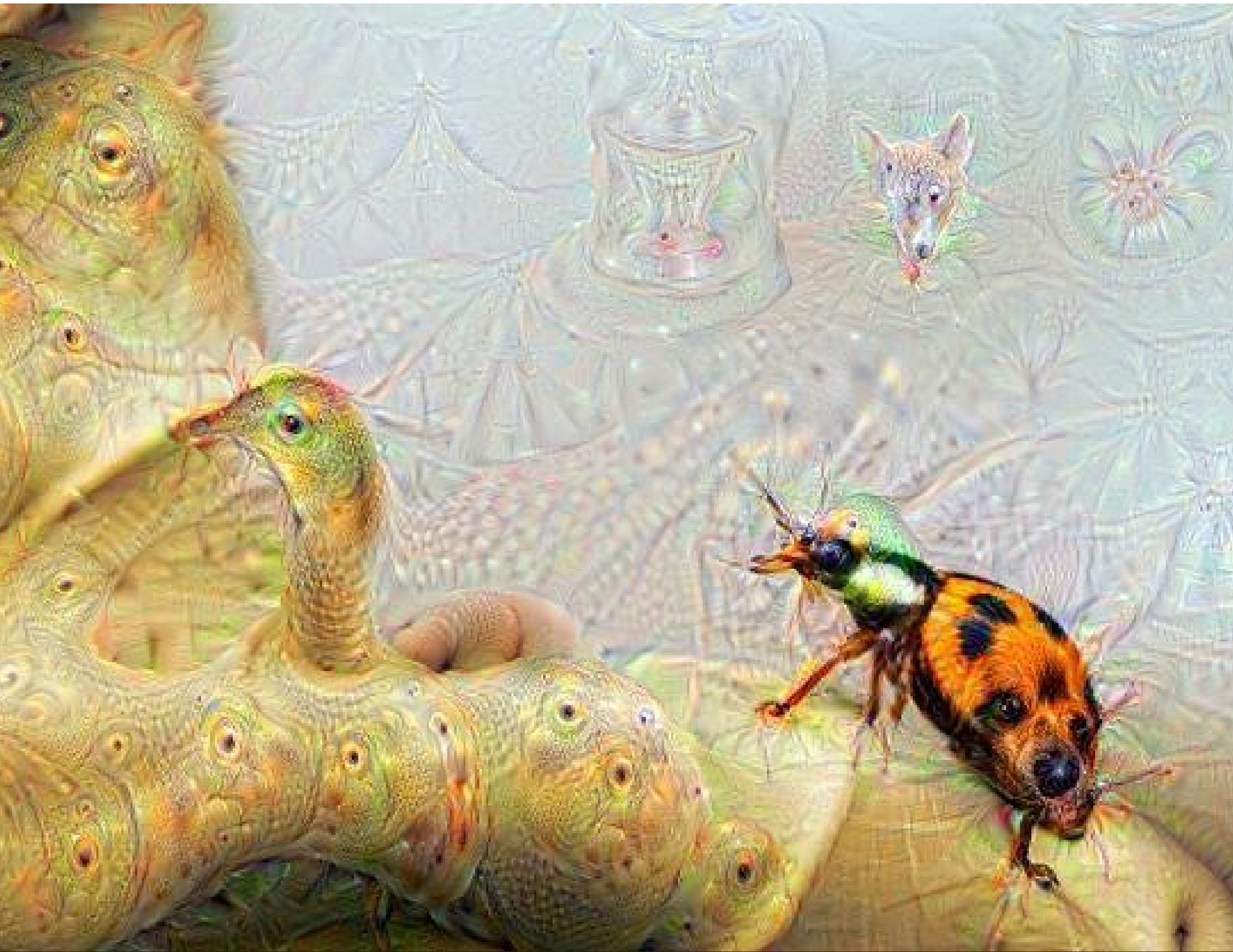
Stockholm
University

Automated image-based taxon identification using deep learning and citizen-science contributions

Miroslav Valan



Doctoral Thesis in Systematic Zoology at Stockholm University, Sweden 2021

Automated image-based taxon identification using deep learning and citizen-science contributions

Miroslav Valan

Academic dissertation for the Degree of Doctor of Philosophy in Systematic Zoology at Stockholm University to be publicly defended on Wednesday 10 March 2021 at 14.00 in Vivi Täckholmsalen (Q-salen), NPQ-huset, Svante Arrhenius väg 20.

Abstract

The sixth mass extinction is well under way, with biodiversity disappearing at unprecedented rates in terms of species richness and biomass. At the same time, given the currentpace, we would need the next two centuries to complete the inventory of life on Earthand this is only one of the necessary steps toward monitoring and conservation of species. Clearly, there is an urgent need to accelerate the inventory and the taxonomic researchrequired to identify and describe the remaining species, a critical bottleneck. Arguably, leveraging recent technological innovations is our best chance to speed up taxonomic research. Given that taxonomy has been and still is notably visual, and the recent break-throughs in computer vision and machine learning, it seems that the time is ripe to exploreto what extent we can accelerate morphology-based taxonomy using these advances inartificial intelligence. Unfortunately, these so-called deep learning systems often requiresubstantial computational resources, large volumes of labeled training data and sophisticated technical support, which are rarely available to taxonomists. This thesis is devoted to addressing these challenges. In **paper I** and **paper II**, we focus on developing an easy-to-use (’off-the-shelf’) solution to automated image-based taxon identification, which is at the same time reliable, inexpensive, and generally applicable. This enables taxonomists to build their own automated identification systems without prohibitive investments in imaging and computation. Our proposed solution utilizes a technique called feature transfer, in which a pretrained convolutional neural network (CNN) is used to obtain image representations (”deep features”) for a taxonomic task of interest. Then, these features are used to train a simpler system, such as a linear support vector machine classifier. In **paper I** we optimized parameters for feature transfer on a range of challenging taxonomic tasks, from the identification of insects to higher groups --- even when they are likely to belong to subgroups that have not been seen previously --- to the identification of visually similar species that are difficult to separate for human experts. In **paper II**, we applied the optimal approach from paper I to a new set of tasks, including a task unsolvable by humans - separating specimens by sex from images of body parts that were not previously known to show any sexual dimorphism. **Papers I** and **II** demonstrate that off-the-shelf solutions often provide impressive identification performance while at the same time requiring minimal technical skills. In **paper III**, we show that phylogenetic information describing evolutionary relationships among organisms can be used to improve the performance of AI systems for taxon identification. Systems trained with phylogenetic information do as well as or better than standard systems in terms of common identification performance metrics. At the same time, the errors they make are less wrong in a biological sense, and thus more acceptable to humans. Finally, in **paper IV** we describe our experience from running a large-scale citizen science project organized in summer 2018, the Swedish Ladybird Project, to collect images for training automated identification systems for ladybird beetles. The project engaged more than 15,000 school children, who contributed over 5,000 images and over 15,000 hours of effort. The project demonstrates the potential of targeted citizen science efforts in collecting the required image sets for training automated taxonomic identification systems for new groups of organisms, while providing many positive educational and societal side effects.

Stockholm 2021
http://urn.kb.se/resolve?urn=urn:nbn:se:su:diva-189460

ISBN 978-91-7911-416-9
ISBN 978-91-7911-417-6

Department of Zoology

Stockholm University, 106 91 Stockholm

AUTOMATED IMAGE-BASED TAXON IDENTIFICATION USING DEEP LEARNING AND CITIZEN-SCIENCE CONTRIBUTIONS

Miroslav Valan

UNIVERSITET STOCKHOLMS

Stockholm
University

Automated image-based taxon identification using deep learning and citizen-science contributions

Miroslav Valan

ISBN print 978-91-7911-416-9
ISBN PDF 978-91-7911-417-6

Printed in Sweden by Universitetsservice US-AB, Stockholm 2021

Dedicated to my daughter Mia and son Teodor without whom this thesis would have been completed two years ago.

Acknowledgment

I would like to express my deepest gratitude to my supervisor **Fredrik Ronquist** for his patience, unconditional support and understanding. I am in his debt for supporting my choices and for finding a way to finish this thesis without too much hassle.

I have received much support from my co-supervisors A**tsuto Maki**, **Karoly Makonyi** and **Nuria Albet-Tores**. Also, I would like to thank my numerous colleagues from the **SU** and **NRM** and especially **Savantic** which I considered my second home during the last five years.

I am very grateful to all my **coauthors** and those who in one way or the other were involved in the work that resulted in this thesis. Thank you all.

I am indebted to my wife **Vlasta**, for her love, patience, encouragement, endless support, sacrifice and forgiveness during the most important journey - life. I want to thank my parents, **Dragica** and **Mirko** and my brother **Dragoslav**.

Lastly and most importantly, I want to thank my children for being an inexhaustible source of joy and happiness. And the best thing is that we are just starting.

To my 3 years old son **Teodor** for making me feel stronger, more organized and better at prioritizing what matters the most in life. You taught me a lesson I once knew: You can succeed only after you try, so never stop trying.

To my giggly and miraculous baby girl **Mia**. I love seeing you happy for no obvious reason. So, just keep giggling.

Abstract

The sixth mass extinction is well under way, with biodiversity disappearing at unprecedented rates in terms of species richness and biomass. At the same time, given the current pace, we would need the next two centuries to complete the inventory of life on Earth and this is only one of the necessary steps toward monitoring and conservation of species. Clearly, there is an urgent need to accelerate the taxonomic research required to identify and describe the remaining species. Arguably, leveraging recent technological innovations is our best chance to speed up taxonomic research. Given that taxonomy has been and still is notably visual, and the recent breakthroughs in computer vision and machine learning, it seems that the time is ripe to explore to what extent we can accelerate morphology-based taxonomy using these advances in artificial intelligence (AI). Unfortunately, these so-called deep learning systems often require substantial computational resources, large volumes of labeled training data and sophisticated technical support, none of which are readily available to taxonomists. This thesis is devoted to addressing these challenges. In *paper I* and *paper II*, we focus on developing an easy-to-use ('off-the-shelf') solution to automated image-based taxon identification, which is at the same time reliable, inexpensive, and generally applicable. Such a system would enable taxonomists to build their own automated identification systems without advanced technical skills or prohibitive investments in imaging or computation. Our proposed solution utilizes a technique called feature transfer, in which a pretrained convolutional neural network is used to obtain image representations ("deep features") for a taxonomic task of interest. Then, these features are used to train a simpler system, such as a linear support vector machine classifier. In *paper I* we optimized parameters for feature transfer on a range of challenging taxonomic tasks, from the identification of insects to higher groups —— even when they are likely to belong to subgroups that have not been seen previously —— to the identification of visually similar species that are difficult to separate for human experts. We find that it is possible to find a solution that performs very well across all of these tasks. In *paper II*, we applied the optimal approach from paper I to a new set of tasks, including a task unsolvable by humans - separating specimens by sex from images of body parts that were not previously known to show any sexual dimorphism. *Papers I* and *II* demonstrate that an off-the-shelf solution can provide impressive identification performance while at the same time requiring minimal technical skills. In *paper III*, we show that information describing evolutionary relationships among organisms can be used to improve the performance of AI systems for taxon identification. Systems trained with taxonomic or phylogenetic information do as well as or better than standard systems in terms of generally accepted identification performance metrics. At the same time, the errors they make are less wrong in a biological sense, and thus more acceptable to humans. Finally, in *paper IV* we describe our experience from running a large-scale citizen science project organized in summer 2018, the Swedish Ladybird Project, to collect images for training automated identification systems for ladybird beetles. The project engaged more than 15,000 school children, who contributed over 5,000 images and over 15,000 hours of effort. The project demonstrates the potential of targeted citizen science efforts in collecting the required image sets for training automated taxonomic identification systems for new groups of organisms, while providing many positive educational and societal side effects.

Abstrakt

Vi är mitt inne i den sjätte massutrotningen, och den biologiska mångfalden försvinner i en rasande fart. Arter förloras för alltid och och den totala biomassan minskar stadigt. Samtidigt skulle vi behöva tvåhundra år för att slutföra inventeringen av livet på jorden i nuvarande takt, och detta är bara ett av de nödvändiga stegen mot övervakning och bevarande av den biologiska mångfalden. Med tanke på detta står det klart att vi måste försöka påskynda den taxonomiska forskning som krävs för att identifiera och beskriva de återstående arterna. Att utnyttja de senaste årens tekniska framsteg är sannolikt vår bästa chans att göra detta. Med tanke på att taxonomi har varit och fortfarande är baserat till tor del på visuella karaktärer, och att det har gjorts stora framsteg de senaste åren inom datorseende och maskininlärning, är det hög tid att utforska i vilken utsträckning vi kan accelerera morfologibaserad taxonomi med hjälp av artificiell intelligens (AI). De senaste framstegen bygger på så kallad djupinlärning ("deep learning"), vilket ofta kräver betydande beräkningsresurser, stora volymer träningsdata och avsevärd teknologisk kompetens. Dessa resurser är sällan tillgängliga för taxonomer. Forskningen som redovisas i denna avhandling syftar till att avhjälpa dessa problem. I *uppsats I* och *uppsats II* fokuserar vi på att utveckla en lättanvänd standardlösning för automatiserad bildbaserad taxonidentifiering, som samtidigt är tillförlitlig, lättillgänglig och allmänt tillämplig. Ett sådant standardsystem skulle göra det möjligt för taxonomer att bygga sina egna automatiserade identifieringssystem utan oöverkomliga investeringar i beräkningsresurser eller i att ackumulera stora digitala bilddatabaser. Vår lösning använder en teknik som bygger på att extrahera de element eller egenskaper som uppfattas av ett avancerat neuralt nätverk (ett

“convolutional neural network”) tränat för en generell bildklassificeringsuppgift i bilder avsedda för en annan uppgift, en taxonomisk identifieringsuppgift. De extraherade bildegenskaperna kan sedan användas för att träna ett enklare klassificeringssystem, till exempel en så kallad stödvektormaskin (“support vector machine”). Vi optimerade parametrarna för den här typen av system på en rad utmanande taxonomiska uppgifter, från identifiering av insekter till högre taxa —– även när de sannolikt tillhör undergrupper som inte har setts tidigare —– till identifiering av visuellt snarlika arter som är svåra att särskilja för mänskliga experter. Vi fann att det var möjligt att utforma ett sådant system så att det hade god prestanda för samtliga dessa uppgifter. I *uppsat II* använde vi det optimala systemet från papper I till en ny uppsättning uppgifter, inklusive en uppgift som inte kan lösas av människor - att separera hanar från honor utifrån bilder av kroppsdelar som inte tidigare var kända att visa någon sexuell dimorfism. *Uppsats I* och emph II visar att det går att utveckla standardlösningar som ger imponerande identifieringsprestanda hos de färdiga identifieringssystemen och samtidigt kräver minimala tekniska färdigheter av användaren. I *uppsats III* visar vi att information som beskriver evolutionära släktskapsförhållanden mellan organismer kan användas för att förbättra prestandan hos AI-system för taxonomisk identifiering. System tränade med taxonomisk eller fylogenetisk information presterar lika bra som eller bättre än standardsystem när de utvärderas med allmänt accepterade prestandamått. Samtidigt är felen de gör mindre felaktiga i biologisk mening och därmed mer acceptabla för människor. Slutligen beskriver vi i *uppsats IV* vår erfarenhet av att genomföra ett storskaligt medborgarvetenskapligt projekt som anordnades sommaren 2018, Nyckelpigeförsöket, för att samla in bilder för att träna AI-system för identifiering av nyckelpigor. Projektet engagerade mer än 15 000 skolbarn, som bidrog med över 5,000 bilder och över 15,000 timmars arbete. Projektet visar vilken enorm potential som finns i att engagera medborgarforskare i att samla in de nödvändiga bilderna för att kunna träna AI-system för automatisk identifiering av nya grupper av djur och växter. Samtidigt kan sådana projekt ge många positiva bieffekter. Inte minst kan de väcka allmänhetens nyfikenhet inför den biologiska mångfalden och intresset för att bevara den för framtiden.

Author's contributions

The thesis is based on the following articles, which are referred to in the text by their Roman numerals:

I. **Valan, M.**, Makonyi, K., Maki, A., Vondráček, D., & Ronquist, F. (2019). Automated taxonomic identification of insects with expert-level accuracy using effective feature transfer from convolutional networks. *Systematic Biology*, Volume 68, Issue 6, November 2019, Pages 876–895, https://doi.org/10.1093/sysbio/syz014.

II. **Valan, M.**, Vondráček, D., & Ronquist, F. Awakening taxonomist's third eye: exploring the utility of computer vision and deep learning in insect systematics. *Submitted.*

III. **Valan, M.**, Nylander A. A. J. & Ronquist F. AI-Phy: improving automated image-based identification of biological organisms using phylogenetic information. *Manuscript.*

IV. **Valan, M.**, Bergman M., Forshage M. & Ronquist F. The Swedish Ladybird Project: Engaging 15,000 school children in improving AI identification of ladybird beetles. *Manuscript.*

Candidate contributions to thesis articles*

Type of contribution	paper I	paper II	paper III	paper IV
Conceived the study	A	A	A	A
Designed the study	A	A	A	A
Collected the data	A	A	A	A
Analyzed the data	A	A	A	A
Manuscript preparation	A	A	A	A

Table 1: **Contribution Explanation:**

A - Substantial: took the lead role and performed the majority of the work.

B - Significant: provided a significant contribution to the work

C - Minor: contribution in some way, but contribution was limited.

Contents

Abbreviations

A list of abbreviations used in the thesis:

AI - Artificial Intelligence

ATI - Automated Taxonomic Identification aystem

CAM - Class Activation Maps

CNN - Convolutional Neural Network

CS - Citizen Science

DL - Deep Learning

GBIF - Global Biodiversity Information Facility

GPU - Graphical Processing Unit

FC - Fully Connected

LR - Logistic Regression

LS - Label Smoothing

PLS - Phylogenetic Label Smoothing

SLP2018 - Swedish Ladybird Project

SVC - Support Vector Classifier

SVM - Support Vector Machin

TLS - Taxonomic Label Smoothing

Chapter 1

Introduction

> *An understanding of the natural world and what's in it is a source of not only a great curiosity but great fulfillment.*
>
> David Attenborough

Biodiversity is under unprecedented pressure due to climate change and the influence of humans. Based on the alarming rates at which species are disappearing it is more than obvious that the sixth mass extinction is under way (Ehrlich, 1995; Laliberte and Ripple, 2004; Dirzo et al., 2014; Ripple et al., 2014; Maxwell et al., 2016; Ceballos et al., 2017). Precious life forms are lost before we became aware of their existence; forms that took evolution millions of years to create. If we would know what we have and what we may lose it would be easier to convince decision-makers to take appropriate action to stop this devastating loss of biodiversity.

The scientific field charged with the task of describing and classifying life on Earth is taxonomy, an endeavor that is as old as humans. Since the very beginnings, we aimed to understand the World around us; we observed, compared, tried to understand and made some conclusions; then we passed the knowledge on to the coming generations in oral and later in written form. It is easy to imagine how food (i.e. living beings - plants, animals and fungi) was on top of our priorities; we had to learn what is edible and tasty and what not, so we probably relied on some information about anatomical features to distinguish one form of life from another. Years later, this became more structured so different forms of life started to be compared based on the same body parts, or the absence or presence of some morphological structures. The first written descriptions of different species were composed by compiling such observations of characters. This is considered as the beginning of *descriptive taxonomy*. During the 18th century, Carl Linnaeus, a Swedish botanist, zoologist and taxonomist, established universally accepted conventions for classifying nature within a nested hierarchy and for the naming of organisms. Today, this system is still in use and it is known as *Linnaean taxonomy* or *modern taxonomy*.

Taxonomy remained predominantly descriptive until the mid-20th century when it became more quantitative thanks to the developments in statistics. Data such as length, width, angles, counts and ratios, combined with multivariate statistical methods, provided a deeper understanding of patterns in the biological world. This marked the beginning of *traditional morphometrics* (Marcus, 1990). In 1980's, taxonomist applied approaches to quantify and analyse variations in shape (known as *geometric morphometrics* Rohlf and Marcus (1993)), which was based on coordinates of outlines or landmarks. These were useful for graphical visualisation and/or statistical analyses, but they were also used in building some of the first systems for automated taxon identification (see below).

Throughout its historical development, it has become increasingly clear that taxonomy is more than just a descriptive scientific discipline; it is a fundamental science on which other sciences—such as ecology, evolution and conservation—rely. In an important sense, taxonomy represents the World's scientific frontier, marking the boundary between the known and the unknown in our discovery of life forms. Unfortunately, taxonomic research is still slow in expanding this frontier. At the current pace, it is expected that it will take many years to describe all species of biological organisms on the planet. The gaps in our taxonomic knowledge and the shortage of taxonomic expertise is known as *the taxonomic impediment* (Agnarsson and Kuntner, 2007; Walter and Winterton, 2007; Rodman and Cody, 2003; Ebach et al., 2011; Coleman, 2015). Clearly, accelerating taxonomic research would bring many positive effects on a wide range of immensely important decisions our civilization needs to make in the very near future.

One possible approach to combating the taxonomic impediment would be to build sophisticated automated taxon identification systems (ATIs). ATIs could help in two ways. First, they could take care of routine identifications, freeing up the time of taxonomic experts so that they could focus on more challenging and critical tasks in expanding our knowledge of biodiversity. Second, sophisticated ATIs could also directly help in the process of identifying and describing new life forms. Until recently, however, ATIs were not particularly effective in solving these tasks. An important reason for this is that they were based on hand-crafted features. For example, if the purpose were to identify insects, relevant features might be the wing venation patterns, the positions of wing vein junctions, or the outlines of the whole body. After human experts identified some potentially informative features, these features would then have to be identified in images manually or through automated procedures that were specifically designed for the task at hand (Arbuckle et al., 2001; Feng et al., 2016; Francoy et al., 2008; Gauld et al., 2000; Lytle et al., 2010,?; O'Neill, 2007; Schröder et al., 1995; Steinhage et al., 2007; Tofilski, 2007, 2004; Watson et al., 2003; Weeks et al., 1999a,b, 1997). Some of the ATIs developed using these techniques have shown great performance (Martineau et al., 2016), but the approach is difficult to generalize because it requires knowledge of programming and image analysis (to formalize manual or code automatic procedures for feature extraction), of machine learning (to build an appropriate classifier) and of the task itself (expertise on the taxa of interest). Clearly, this approach does not generalize well. For every new task we need to consider factors that determine the best target features, and then hand-craft procedures to encode those features. For these reasons, such ATIs have been presented for only a few groups. Note that a considerable amount of human effort must be spent before we can even evaluate whether it is feasible to solve the identification task at hand using this approach.

1.1 Convolutional neural networks and deep learning

In recent years, more general approaches to image classification have developed greatly (LeCun et al., 2015; Schmidhuber, 2015). This is part of a general trend in computer science towards more sophisticated and intelligent systems, that is, towards more sophisticated artificial intelligence (AI). The trend is driven by improved algorithms, rapidly increasing amounts of data, and faster and cheaper computation. In the field of computer vision, the development has been particularly fast in recent years with the introduction of more complex and sophisticated artificial neural networks, known as convolutional neural networks (CNNs), and the training of advanced (deep) versions of these networks with massive amounts of data, also known as deep learning (DL). The dramatic progress in computer vision has been enabled also by the development of graphical processing units (GPUs), adding a considerable amount of cheap processing power to modern computer systems.

The first super-human performance of GPU-powered CNNs in an image classification task (Cireşan et al., 2011) was reported in 2011 in a traffic sign competition (Stallkamp et al., 2011). The breakthrough came in 2012, when a CNN architecture called AlexNet (Krizhevsky et al., 2012) out-competed all other systems in the ImageNet Large Scale Visual Recognition Challenge (Russakovsky et al., 2015), a larger and more popular image classification challenge. The good news about DL performance spread quickly, and we soon witnessed successful applications in other research areas, such as face verification (Taigman et al., 2014), object localisation (Tompson et al., 2015), image and video translation into natural language (Karpathy and Fei-Fei, 2015), language translation (Sutskever et al., 2014; Jean et al., 2015), speech recognition (Sainath et al., 2013; Hinton et al., 2012; Zhang and Zong, 2015) and question-answer problems (Kumar et al., 2016).

The core of every CNN architecture is a set of convolutional (conv) layers, hence the name convolutional neural network (Fukushima, 1979, 1980; Fukushima et al., 1983; LeCun et al., 1989). The convolutional part of a CNN enables automatic feature learning; it works as a "feature extractor". The resulting features are then fed through one or more fully connected (FC) layers, which deal with the classification task. The FC layers in principle correspond to a traditional multi-layer perceptron (Rosenblatt, 1957) which is a simple fully-connected feed-forward artificial neural network. Learning in a CNN is possible thanks to the backpropagation algorithm (Kelley, 1960; Linnainmaa, 1976; Werbos, 1982; Rumelhart et al., 1986; Schmidhuber, 2014) and gradient-based optimization (Robbins and Monro, 1951; Kiefer et al., 1952; Bottou et al., 2018). Most of the CNNs used today also contain other layers, such as pooling (dimensionality reduction) (Fukushima, 1979, 1980), normalization (e.g. BatchNorm (Ioffe and Szegedy, 2015), helps with stabilizing the training) or regularization layers (Hanson, 1990; Srivastava et al., 2014) (helps with addressing the over-fitting), among many others.

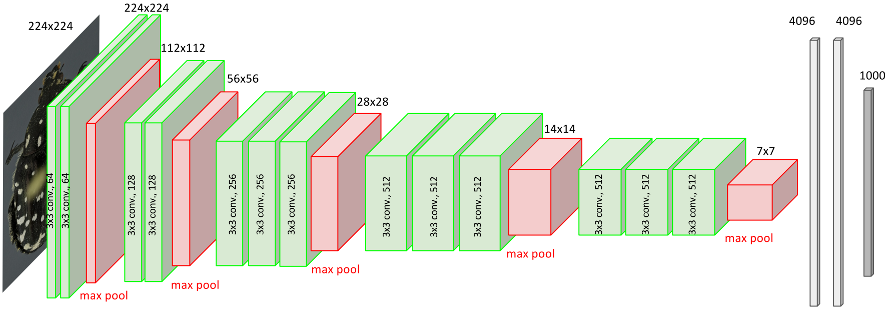


Figure 1.1: Architecture of VGG16 (Simonyan and Zisserman 2014), a simple modern CNN. VGG16 consists of five convolutional blocks, each block consisting of two or three convolutional layers (green) followed by a MaxPooling layer (red). These blocks are followed by three layers of fully connected neurons (gray), the last of which consists of a vector of length 1000. Each element in this vector corresponds to a unique category in the ImageNet Dataset (Russakovsky et al., 2015) for which this architecture was initially built. Adopted from **paper I**

To better understand the basic structure of a CNN, consider Figure 1.1 illustrating a well known deep CNN architecture, VGG16 (Simonyan and Zisserman, 2014). This architecture is simple, yet very powerful and therefore one of the best studied. It has also become one of the most commonly utilized architectures for addressing various research questions. VGG16 consists of five convolutional blocks followed by two FC hidden layers and the output layer (also FC), where the number of nodes corresponds to the number of categories the network is trained for. The convolutional block is made of convolutional (conv) layers followed by a MaxPooling layer. Every conv layer in the VGG family is made of 3x3 filters. The number of layers in each block and the number of filters in each layer vary, so we have 2x64, 2x128, 3x256, 3x512, and 3x512 ”layers x filters” respectively for the five convolutional blocks. Note that some of the recent CNNs have more than a hundred layers including dozens of convolutional layers with much more complex architectures. VGG16 uses max pooling with kernel of size 2x2 and a stride of 2, taking only the maximum value within the kernel (other options would be the average, sum, etc). This results in reduced width and height of the feature matrix by a factor of two, and total amount of data by a factor of four. Unlike a conv layer, where nodes are connected to the input image or previous layer only by the local region of the same size as the corresponding kernel, the nodes in a fully connected layer (FC) are connected to every node in the previous layer (as in a simple multi-layer perceptron).

Modern CNNs often require large sets of labeled images for successful supervised learning. Recently, it has been discovered that features learned by a CNN that has been trained on a generic image classification task (source task) can be beneficial in solving a more specialized problem (target task) using a technique called *transfer learning* (Caruana, 1995; Bengio, 2012; Yosinski et al., 2014; Azizpour et al., 2016)). Transfer learning works primarily because a fair amount of relevant low-level features (edges, corners, etc.) are likely similar between source and target tasks. Intermediate (you can think of eye, nose, mouth, etc.) and high-level (e.g. head, leg) features are more specialized and their usefulness depends on the distance between the source and target tasks.

There are two variants of transfer learning: *feature transfer* and *fine-tuning*. In feature transfer, a pretrained CNN serves as an automated feature extractor (Azizpour et al., 2016; Donahue et al., 2014; Oquab et al., 2014; Sharif Razavian et al., 2014; Zeiler and

Fergus, 2014; Zheng et al., 2016)). Each image is fed through a pretrained CNN, and its representation (feature vector) is extracted from one of the layers of the CNN, capturing low- to high-level image features. Then, these features are used to train a simpler machine learning system, such as a Support Vector Machine (SVM) (Cortes and Vapnik, 1995)), a Logistic Regression (LR) (Cox, 1958), a Random Forest (Breiman, 2001) or a Gradient Boosting (Friedman, 2001). This approach is usually computationally more efficient and it can benefit from properties of the chosen classifier (e.g. SVMs tend to be resistant to overfitting, outliers and class imbalance, while LR is simple, intuitive and efficient to train).

Taking a pretrained CNN (or part of it) as initialization for training a new model is known as fine-tuning. Fine-tuning tends to work well when the specialized task is similar to the original task (Yosinski et al., 2014). Compared to training a CNN from scratch, fine-tuning reduces the hunger for data and improves convergence speed, but it may require a fair amount of computational power. In fine-tuning, the images have to be run through the CNN in a forward pass, and then the computed derivatives from the predictions have to be backpropagated to modify the filters (the latter is the more computationally expensive part). This process of alternating forward and backward passes has to be repeated until our model converges. There is also the problem of defining appropriate learning hyper-parameters in order to enable sufficient flexibility in learning of the new task while avoiding overfitting.

1.2 Aim of the current thesis

With the breakthroughs in deep learning and computer vision outlined above, it is now possible to meet the requirements for highly competent ATIs (Wu et al., 2019; Hansen et al., 2020; Joly et al., 2018; Van Horn et al., 2018; Cui et al., 2018), which can help accelerate taxonomic research. Given a sufficient number of training examples and their labels (e.g. a species name obtained from a taxonomic expert or with DNA sequencing), these new systems learn to identify features important for identification directly from images, without any interference from humans; that is, there is no need for an expert to indicate what is informative, the system finds the relevant image features by itself. However, as indicated above, a limiting factor is access to sufficient amounts of training data, which could be a serious challenge for most species identification tasks. There are various reasons for this. Firstly, the species abundances are usually imbalanced: there are often a few common species, while the majority of species are rarely seen and almost never photographed (or collected). Secondly, the number of images, for those species that are photographed (or collected), is hugely imbalanced towards more attractive groups or subgroups. Among insects, for instance, butterflies are wildly popular targets for nature photographers, while small midges, flies or parasitic wasps are almost never photographed regardless of how common they are. The popularity may also vary among morphs or life stages; for instance, butterfly eggs and pupae are photographed much less than adult butterflies. Thus, collecting enough images of all species and relevant morphs to be able to train a state-of-the-art AI system may be a daunting task. In addition to the challenge of putting together an adequate training set, another serious challenge in training such a state-of-the-art AI system on a dedicated taxonomic task is that it requires advanced technical skills that most taxonomists lack. In this thesis, I address these challenges.

The main focus of the thesis has been on insects because they are diverse, challenging to identify and there are many groups of insects that are poorly studied. In fact, more than half of the known species on Earth are insects (over a million according to Zhang (2011)); and many scientists are suggesting that what we know today is just a fraction of what is left to be discovered (Mora et al., 2011; Stork et al., 2015; Novotny et al., 2002). For illustration, consider estimates of the number of undescribed species of all chordates together (15,000), plants (80,000) and insects (4,000,000) (Chapman et al., 2009). The disparity is even greater if we base the comparison on how much we know about their physiology, behaviour, spatial and temporal distributions. Despite these knowledge gaps, insects play many important roles in our ecosystems, both beneficial ones, for instance as pollinators of crops, but also less favorable ones, for instance as pests, invasive species or even vectors of disease. The enormous diversity of insects, the shortage of taxonomic expertise (Gaston and May, 1992; Gaston and O'Neill, 2004), and the importance of many insect species in our ecosystems combine to emphasize the need for accelerating taxonomic research on insects and the potential use for ATIs in doing so. Nevertheless, the findings presented in the thesis are general and should apply to image-based identification of any group of organisms with AI.

An important goal of the current thesis has been to develop techniques enabling taxonomists to build their own sophisticated ATIs using reliable and computationally inexpensive approaches, and without prohibitive investments in imaging (**paper I** and **II**). In **paper I**, we explored methods that might allow taxonomists to develop ATI systems even when the available image data and machine learning expertise are limited. Specifically, we focused on feature transfer, as previous work has indicated that features obtained from pretrained CNNs is a good starting point for most visual recognition tasks (Sharif Razavian et al., 2014). A CNN pretrained on a general image classification task was used as an automated feature extractor, and the extracted features were then used in training a simpler classification system for the taxonomic task at hand. By optimizing the feature extraction protocol, we were able to develop high-performing ATIs for a range of taxonomic identification tasks using fairly limited image sets as training data. Specifically, we looked at two challenging types of tasks: (1) identification of insects to higher groups, even when they are likely to belong to subgroups that have not been seen previously; and (2) identification of visually similar species that are difficult to separate for human experts. For the first type of task, we looked at the identification of images from the iDigBio repository of Diptera and Coleoptera, respectively, to higher taxonomic groups. For the second type of task, we looked at the identification of beetles of the genus *Oxytherea* to species, using a dataset assembled for the paper, and stonefly larvae to species, using a previously published dataset.

In **paper II**, we aimed to address some questions on automated identification that are frequently asked by insect taxonomists: Which techniques are best suited for a quick start on an ATI project? How much data is needed? What is the needed image resolution? Is it possible to tackle identification tasks that are unsolvable by humans? To answer these questions, we created two novel datasets of 10 visually similar species of the flower chafer beetle genus *Oxythyrea*. The best performing system found in **paper I** was then used as an off-the-shelf solution and applied to these datasets in several experiments designed to answer the questions. In addition, we repeated the same experiments using some state-of-the-art approaches in image recognition. We show that our off-the-shelf system, while offering an "easy-to-use instant-return" approach, is often sufficient for testing interesting hypotheses. In fact, the identification performance of ATIs based on the off-the-shelf system was not too far from that of state-of-the-art approaches in our experiments, and it provided similar insights (feasibility, misidentification patterns, etc.) compared to the more advanced systems. We even demonstrate that our off-the-shelf approach can be successfully used on a challenging task that appears unsolvable to humans.

It is well known that CNNs occasionally make catastrophic errors; e.g., misidentifying one category for a completely unrelated category - a mistake that humans would be very unlikely to make. We address this in a biological setting in **paper III** by leveraging a recently introduced technique called label smoothing (Szegedy et al., 2016). Specifically, we propose label smoothing based on taxonomic information (taxonomic label smoothing) or distances between species in a reference phylogeny (phylogenetic label smoothing). We show that networks trained with taxonomic or phylogenetic information perform at least as well on common performance metrics as standard systems (accuracy, top3 accuracy, f1 score macro), while making errors that are more acceptable to humans and less wrong in an objective biological sense. We validated our proposed techniques on two empirical examples (38,000 outdoor images of 83 species of snakes, and 2,600 habitus images of 153 species of butterflies and moths).

As mentioned above, CNNs typically require large training sets of accurately labeled images. Assembling such training sets for developing ATIs could be addressed by soliciting the help from citizen scientists. We explored this in **paper IV**. In the Swedish Ladybird Project (SLP2018), we engaged more than 15,000 Swedish school children in collecting photos of ladybird beetles (Coccinellidae). The children collected more than 5,000 photos of 30 species of coccinellids. This is almost as many coccinellid images as the rest of the World contributed from around the globe to the Global Biodiversity Information Facility

(GBIF) portal during the same period—the summer of 2018. We found that adding the SLP2018 images to the GBIF data resulted in improvements of ATI model performance across various evaluation metrics for all but the most common ladybird species.

Chapter 2

Summary of papers

Begin at the beginning," the King said, very gravely, "and go on till you come to the end: then stop.

Lewis Carroll

2.1 Paper I

2.1.1 Material and methods

Our experiments in **paper I** were designed to find optimal feature extraction settings for various taxonomic identification tasks and training datasets using a single feed-forward pass through a pretrained CNN. Recent work has indicated that these so-called deep features, although the extraction of them has been learned on a general image classification task, are very robust and, in combination with simple classifiers such as SVMs (Cortes and Vapnik, 1995), can yield results on par with or better than state-of-the-art results obtained with hand-crafted features (Azizpour et al., 2016; Sharif Razavian et al., 2014; Donahue et al., 2014; Oquab et al., 2014; Zeiler and Fergus, 2014; Zheng et al., 2016).

A well known CNN architecture, VGG16 (Simonyan and Zisserman, 2014), and its publicly available checkpoint pretrained on the ImageNet task (Simonyan and Zisserman, 2014), were utilized across all our experiments. Our experiments were based on features extracted after each conv block, and we refer to them as *c1*-*c5*, respectively. The FC layers were excluded because they were dependent on the image input size. In our experiments, we investigated the effects of: input image size, pooling strategy (Max vs Average), features from different layers (feature depth), normalization (*l2* and/or signed square root), feature fusion, non-global pooling and image aspect ratio.

2.1.2 Datasets

To find optimal hyperparameters for feature extraction we created four datasets representing two types of challenging taxonomic tasks; (1) identifying insects to higher groups when they are likely to belong to subgroups that have not been seen previously; and (2) identifying visually similar species that are difficult to separate even for experts.

Three out of four datasets (D1-D4) were assembled specifically for this paper (Table 2.1). The first two datasets (head view of flies and top view of beetles, D1 and D2 respectively) were designed to investigate how far this approach can get us when assigning novel species to known higher taxonomic categories. The remaining two datasets were used to investigate whether the same techniques would be able to discriminate among visually very similar species (top view of sibling beetle species and species of Plecoptera larvae in different life stages, D3 and D4 respectively). Images from all four datasets were taken in lab settings. They all had uniform background (the same uniform background across all images in D3-D4) and with small amounts of image noise (pins, dust, labels, scales, measurements). In all datasets but D4, objects were large, centered and share almost the same object orientation (imaged in a standard taxonomic imaging procedure).

2.1.3 Experiments

Impact of image size. Previous work demonstrated that concatenating features from images of different scales (image sizes) could improve the performance on fine-grained classification tasks (Takeki et al., 2016). However, in order to obtain features from the

Table 2.1: Datasets used in **paper I**. Datasets D1 and D2 are used for a task of assigning novel species to known higher taxonomic categories and the other two datasets for a task of separating specimens of visually similar species. In all datasets, the images were taken in lab settings with uniform background, large centered objects (not in all images in D4); same object orientation (except D4) and small amount of background noise (pins, dust, labels, scales, measurements). Stars (*) indicate datasets composed for *paper I* using images obtained from www.idigbio.org. Adapted from **paper I**.

ID	Insect	Categories	Images per taxa	View	Source
D1	Flies	11 families	24 -159	face	*
D2	Beetles	14 families	18 - 900	top	*
D3	Beetles	3 species	40-205	top	This study
D4	Stoneflies	9 species	107-505	top	Lytle et al. (2010)

same image of different scales one needs to execute multiple feed-forward passes which results in increased computational cost. Unlike this technique, we opted for finding the optimal input size for a single feed-forward pass. In this experiment we restricted our attention to *c5*.

Impact of pooling strategy. Global pooling (Lin et al., 2013) is a common way to reduce dimensionality of deep features. Despite several recently proposed alternatives, the two most common pooling strategies are still global max pooling and global average pooling. We experimented with both pooling strategies and with a simple combination of the two (concatenation). As in the previous experiment, we used *c5* features only.

Impact of feature depth. According to Azizpour et al. (2016), one of the most important factors for the transferability of pretrained features is the distance between the target and the source tasks. If the task is to separate breeds of dogs then we may expect the layers toward the end (FC layers) to perform the best. This is because the source dataset ImageNet has a lot of dog categories so the later layers have probably learned so-called high level features (you can think of body parts and their shapes - legs, head). In contrast, if the task is to separate two visually similar beetle species that differ only in small details, such as the degree of hairiness (corresponding to fine-grained differences in image texture), then we may want to focus on features from earlier layers (conv layers). To investigate how the feature depth affects performance on our taxonomic identification tasks, we compared extracted features from all five convolutional blocks *c1-c5*.

Impact of feature normalization. Reducing the variance of the elements in the feature vectors is known to facilitate classification. We experimented with two common normalization techniques: *l2*-normalization and signed squared root normalization as in Arandjelovic and Zisserman (2013).

Impact of feature fusion. The advantage of combining features from different layers is demonstrated in Zheng et al. (2016). Unlike their work, we only tested fusion of features from conv blocks (*c1-c5*) to avoid dependency on image input size.

Impact of non-global pooling. Feature matrices of the intermediate layers are large. The total size is equal to HxWxF - where H is the height, W is the width and F is the depth or the number of filters of the convolutional block. As the first two dimensions (H,W) of the feature matrices depend on image input size, and the number of filters is large, some dimensionality reduction is necessary in extracting features from intermediate conv layers. Global pooling decreases the feature matrix to a vector of size 1x1xF. This minimizes the computational cost for classifier training, prevents overfitting, but it is also known to result in better performance compared to just flattening raw feature matrices (Zheng et al., 2016).

We investigated the effect of intermediate levels of dimensionality reduction. Specifically, we reduced raw feature matrices to matrices of sizes 2x2xF, 4x4xF, 7x7xF, 14x14xF and 28x28xF, which were then flattened. These intermediate levels of dimensionality reduction increase computational cost but potentially preserve more information.

Impact of image aspect ratio. We maintained the image aspect ratio across all the experiments described above. The images were symmetrically padded with random or uniform pixels, which resulted in preserved object aspect ratio but some loss of information due to the added uninformative pixels. An alternative procedure would be to instead preserve the image information by image resizing, resulting in distorted objects, instead of padding with uninformative pixels. In this experiment, we compared these two approaches to examine whether it was more important to maintain aspect ratio or to preserve image information.

Classifier and evaluation of classification performance

The extracted features were fed into SVM, specifically a one-vs-all support vector classifier (SVC) (Cortes and Vapnik, 1995). This classifier is a common choice for these types of applications because it is memory efficient (uses only support vectors), and because it works well with high dimensional spaces (Vapnik, 2000) and with unbalanced datasets (He and Garcia, 2009). We validated our results using a tenfold stratified random sampling strategy without replacement. In each iteration, one subset was used as the test set, while the classifier was trained on the remaining nine. As the evaluation metric we used accuracy averaged across individual subsets. A similar validation strategy was utilized across other experiments in this thesis unless otherwise noted.

2.1.4 Results

Evaluation of individual steps

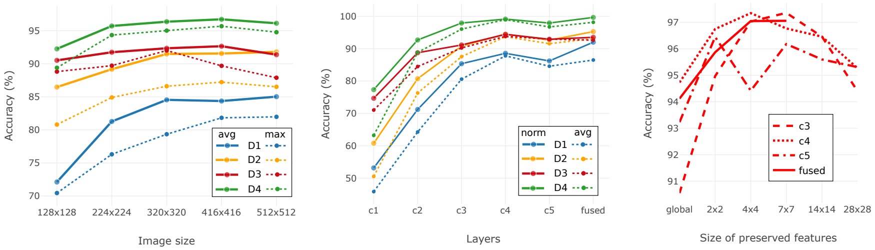


Figure 2.1: We show A) the effect of the image size and pooling strategy (**left**); B) the effect of the feature depth, normalization and feature fusion (**center**); and C) the effect of non-global pooling in one of our datasets - D3 (**right**). Adapted from **paper I**.

Impact of image size. The first step in our experiments was to find an appropriate image size that would perform well across tasks and datasets. We focused on *c5* features and assessed the performance for several input sizes (Fig 2.1 - left). We found that the accuracy increased until the size of *416x416* on most of the datasets, and that in some cases using even larger images resulted in worse performance. Thus, we decided to proceed with *416x416* input image size.

Impact of pooling strategy. In our experiments, global average pooling yielded better results than global max pooling. This could be explained by the fact that, in our datasets, objects occupy a large portion of the image. Thus, by averaging, we allow more object information to affect the final feature state than if we simply take a single value (the max), as in max pooling. Concatenating the two feature vectors obtained from the two different pooling strategies yielded results that were intermediate between the two separate results.

Impact of feature depth. Our results show that c4 features perform the best, while *c3* yields results that are comparable with those of *c5*.

Impact of feature normalization. Signed square root normalization increased the performance but, somewhat surprisingly, we found that *l2*-normalization had a negative effect on accuracy, regardless of whether it was performed with or without signed square root normalization.

Impact of feature fusion. Feature fusion further improved the accuracy. The only exception was on D3, where fusion marginally fell behind the single best layer (only one image difference in accuracy).

Impact of non-global pooling. In this experiment we obtained further improvements on three out of four datasets. For D1-D3, we found that pooling down to 4x4xF yielded the best results, surpassing the globally pooled features. The most evident advantage of an intermediate level of pooling was on D3. We did not see any improvements at all on D4. This was the only dataset with non-centered and occasionally small objects, which could potentially result in having many receptive fields with no information about the object. In the three remaining datasets (D1-D3), the objects were always large, presumably generating a significant number of receptive fields in later CNN layers.

Impact of image aspect ratio. In all experiments reported above, the image aspect ratio was maintained. However, we were able to slightly improve the best performing model from the previous experiment by resizing images without maintaining the aspect ratio. Obviously, such resizing distorts objects but it also provides more information because there are no added uninformative pixels.

Performance on related tasks.

To validate our conclusions, we evaluated our optimal solution on several recently published biological image classification tasks (see Table 2.2). Our approach performed about as well

Table 2.2: Comparison of performance of our method on some recently published biological image classification tasks. We used input images of size *416x416* (*160x160* for Pollen23), global average pooling, fusion of *c1–c5* features, and signed square root normalization. Accuracy is reported using tenfold cross validation except for EcuadorMoths, where we used the same partitioning as in the original study. **Bold** font indicates the best identification performance. The result on EcuadorMoths in parenthesis is from a single c4 block. (fm = families; sp = species). Adapted from **paper I**.

Datasets	Classes	Our	Others	Reference	Method
ClearedLeaf	19 fm	**88.7**	71	Wilf et al. (2016)	SIFT + SVM
CRLeaves	255 sp	**94.67**	51	Carranza-Rojas et al. (2017)	finetune InceptionV3
EcuadorMoths	675 sp	55.4	55.7	Rodner et al. (2015)	AlexNet+SVM
EcuadorMoths	675 sp	(**58.2**)	57.2	Wei et al. (2017)	VGG16+SCDA+SVM
Flavia	32 sp	**99.95**	99.6	Sun et al. (2017)	ResNet26
Pollen23	23 sp	**94.8**	64	Gonçalves et al. (2016)	CST + BOW

as or better than previously published methods.

2.2 Paper II

2.2.1 Material and methods

For the experiments in **paper II**, we created two image datasets on 10 visually similar species of flower chafer beetles of the genus *Oxythyrea*. For the first dataset, we collected images using a standardized taxonomic imaging setup, in which images with different depth of field were stacked together in a single high resolution image. For the second dataset, the same specimens were photographed in a much simpler and faster way using only a smartphone and a cheap 2$ attachable lens (see Figure 2.2).

We first experimented with images of different habitus views (dorsal and ventral). Hu-

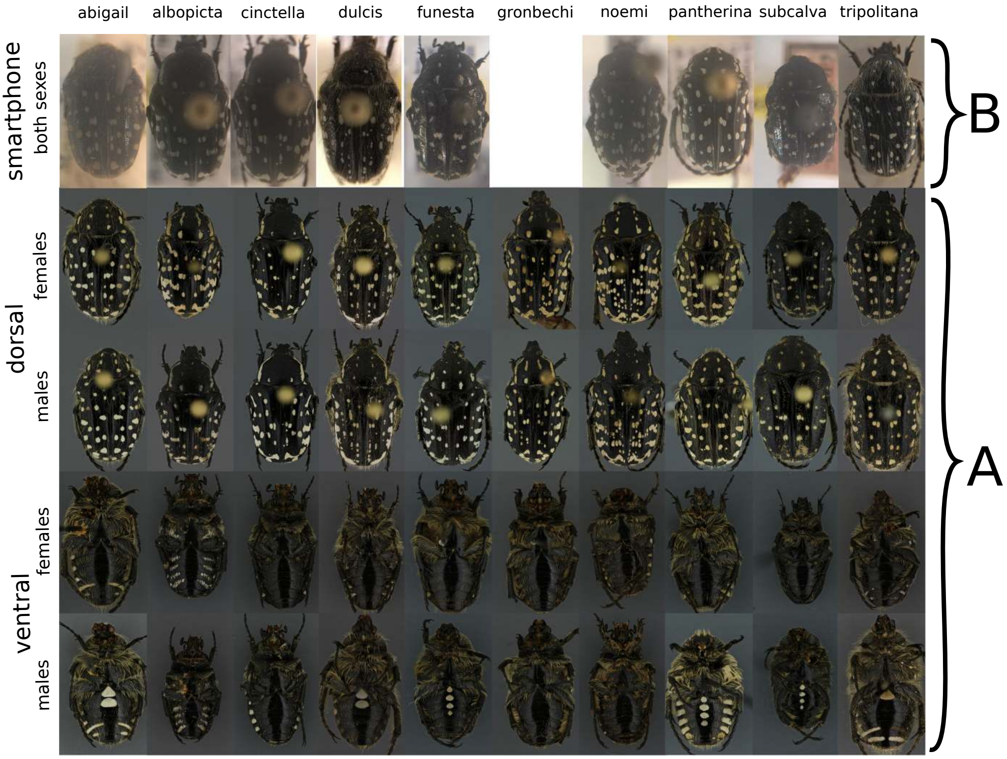


Figure 2.2: **Fig. 2. Datasets of ten *Oxythyrea* species used in *paper II*.** We show example images of **dataset *B*** (first row) collected with a smartphone and a cheap 2$ attachable lens and example images from **dataset *A*** (remaining rows) collected in a standardized taxonomic imaging setting. Dataset *A* contains images of dorsal and ventral habitus including images of both sexes. Note that *Oxythyrea* beetles show sexual dimorphism only on their abdomen (ventral view). Adapted from **paper II**.

mans often find one habitus view more useful for identification than the other and we investigated whether this was also true for the ATIs we developed. Then, we investigated how "quick and dirty" image collection using a smartphone and a cheap attachable lens compares with the time-consuming standard taxonomic imaging setting. Lastly, we explored whether the same approach was applicable on tasks unsolvable by humans. Here, we experimented with separating *Oxythyrea* specimens by sex using images of the dorsal view. According to previous work on *Oxythyrea*, these species do not show any sexual dimorphism in this view.

The experiments in **paper II** were based on optimal parameters from **paper I**. Specifically, we used VGG16 (Simonyan and Zisserman, 2014) pretrained on ImageNet (Russakovsky et al., 2015) for feature extraction and SVC (Cortes and Vapnik, 1995) for classification. Features from all five convolutional blocks were reduced using global average pooling, then concatenated and normalized using the signed square-root method. The only difference was the input image size, which was set to smaller size (*224x224*) in order to speed up the experiments. The validation approach was the same as in **paper I**.

In addition to the off-the-shelf approach developed in **paper I**, we repeated the same experiments using some current state-of-art techniques in image recognition. Specifically, we utilized a well known CNN *SE-ResNext101-32x4* (Hu et al., 2018) and fine-tuned a publicly available checkpoint pre-trained on the ImageNet dataset (Russakovsky et al., 2015). This architecture is a variant of ResNets (He et al., 2016) (networks with residuals modules) with many improvements added subsequently, such as "cardinality" (Xie et al., 2017) and a squeeze and excite block (Hu et al., 2018). The learning rate (*lr*) was adjusted using the one cycle policy ((Smith, 2018, 2015; Smith and Topin, 2017)), with a maximum *lr* set to 0.0006. We set the batch size to 12 and back-propagated accumulated gradients on every two iterations for a total of 12 epochs. As our optimization strategy, we used an adaptive learning rate optimization algorithm called Adam (Kingma and Ba, 2014) with a momentum of 0.9. Regularization was done using *i)* a dropout (Srivastava et al., 2014) layer (0.5) inserted before the last layer; *ii)* label smoothing as our classification loss function (Szegedy et al., 2016); and *iii)* augmentation (zooming, flipping, shifting, brightness, lighting, contrast and mixup (Zhang et al., 2017)). Lastly, with Class Activation Maps (CAM) (Zhou et al., 2016), we visualized so-called heat maps for relevant category-specific regions of images. These heat maps light up the regions that are important for the AI system in identifying an image as belonging to a particular category.

2.2.2 Results

Off-the-shelf approach

Our results using the off-the-shelf system suggest that either one of the habitus views (dorsal or ventral) can be successfully utilized in identifying the ten species of *Oxythyrea*. However, better accuracy is achieved with images of dorsal habitus compared to ventral (3x smaller error rate). This finding corresponds to human perception of the difficulty of the task. Combining information from both views, we observed only slight improvements. This is likely caused by the above-mentioned difference in error rate between the views. If we had combined information from similarly performing views, one would have expected a greater positive impact of the combination.

The images collected by a smartphone and a cheap attachable lens performed almost as well as the high-resolution images. Although humans find such images difficult to use for identification, clearly inferior to high-resolution images, they are apparently often sufficient for machine identification. A possible reason for this is that the images fed to current AI systems for image classification are reduced in size (often to around 224x224 pixels). After reduction in size, the high-resolution taxonomic images are probably quite similar to the smartphone ones, possibly explaining why they do not result in significantly better identification performance.

The off-the-shelf feature transfer solution found sexual dimorphism of the dorsal habitus in at least some *Oxythyrea* species. In two species, *O. albopicta* and *O. noemi*, the model seemed to be able to identify most of the males correctly, while the sex identifications for female specimens were comparable to guessing randomly.

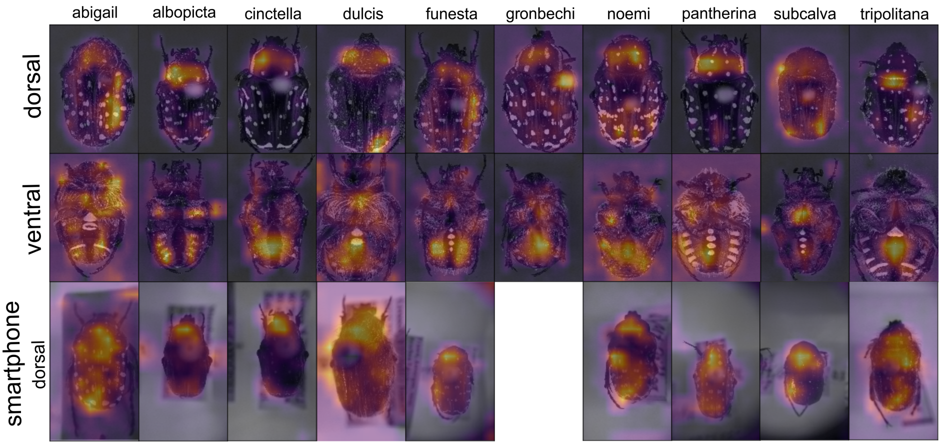


Figure 2.3: Class Activation Maps for the specimens from the species/sex task (only select specimens are depicted). Adapted from **paper II**.

Beyond off-the-shelf solution

As expected, fine-tuning yielded better results than the off-the-shelf solution, drastically reducing the error rates (2-5x). This approach also allowed us to compute heat maps, which made it possible to compare machine reasoning about the identification task to the reasoning of taxonomists. The heat maps showed that the model was often focusing on the pronotum. In species considered easier to identify, this was the only region that was highlighted (*O. albopicta*, *O. cinctella*), while on more difficult species (e.g. *O. dulcis*, *O. noemi*) the model used information from a wider region including the whole elytra. According to the heat maps, *O. tripolitana* was easily identified using information from the small part between the pronotum and the scutellum. This species has an accumulation of setae in this region, which until now has not been seen as a reliable morphological character among taxonomists.

When the task was sorting to sex alongside species based on the dorsal habitus, the model again focused on the pronotum and sometimes on the elytra. It is not clear what exactly is the discriminating feature. Looking at the heat maps from the ventral side, for six species the model easily recognized the median region, where the pattern of white dots was present only in males (Fig. 2.3). In the remaining species, this pattern was not present in either sex. However, the same region was still highlighted. The reason was likely a sex-specific grove present in the median part of the abdomen, which is clearly visible if the abdomen is examined from the side and with appropriate illumination. However, this groove was impossible for humans to see in most of the images of the ventral habitus used in the experiment.

2.3 Paper III

2.3.1 Material and methods

In **paper III**, we explored the utility of taxonomic or phylogenetic information in training and evaluating CNNs for identification of biological species, with the aim of improving these systems so that they make less catastrophic errors. Specifically, we included the biological information during the training by adjusting targets with label smoothing (LS) based on taxonomic information (taxonomic label smoothing - TLS) or distances between species in a reference phylogeny (phylogenetic label smoothing - PLS). Similar to *paper II* we used finetunning from a pretrained checkpoint and we compared our approach against two well established baselines: one-hot encoding and standard LS (Szegedy et al., 2016).

In one-hot encoding, categories are represented as binary vectors of length equal to the number of unique categories, with 1 for the correct target and 0 for the other categories. LS is a weighted average of the one-hot encoded labels and the uniform probability distribution over labels. In both scenarios, networks are optimized toward targets using a distribution, in which categories are equally distant from each other. In such a setting, a network is equally penalized for every misidentification it makes. With our approach, hierarchical information based on biological relatedness is incorporated in the target encoding. This results in the network being less penalized for misidentifying categories that are closely related biologically, and more penalized for mixing up distant categories.

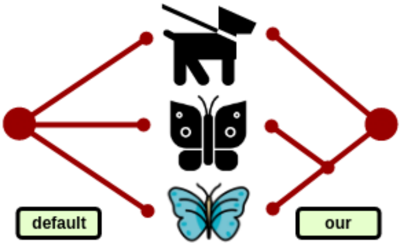


Figure 2.4: Systems optimized toward one-hot or LS (Szegedy et al., 2016) targets assume all categories are equally distant from each other and hence all errors are penalized equally. We propose to smooth the targets using hierarchical biological relatedness information (taxonomy or phylogeny) so that systems are penalized more for erroneous identifications that are farther away from the correct category.

Our approach differs from one-hot encoding and standard label smoothing only in the target encoding. We use a weighted average of one-hot encoded labels and a non-uniform distribution over labels representing hierarchical biological information about the categories based on taxonomic ranks, anagenetic distance, or cladogenetic distance. For taxonomic rank, we acounted the number of shared taxonomic levels (genus and family) between the correct target and each of the other categories. For anagenetic-distance smoothing, we used the branch lengths separating the correct target and each of the other categories on the reference tree as the distance measure. For cladogenetic-distance smoothing, we instead used the number of edges separating the categories on the reference tree as the distance measure. Lastly, we normalized the resulting values by subtracting them from the maximum value (so that closely related categories had the highest values); and then we normalized the values, that is, we divided the values with the sum over all categories so that the sum of the distribution was equal to 1 as in one-hot labels. For all hierarchical smoothing methods, we explored several mix-in proportions (smoothing values), β, of hierarchical information to binary information ($\beta \in \{0.025, 0.05, 0.1, 0.2, 0.4\}$) on two image data sets: 38,000 outdoor images of 93 species of snakes and 2,600 habitus images of 153 species of Lepidoptera (butterflies and moths) images obtained from GBIF (2020).

In addition to accuracy used in papers I-II, here we used two more common evaluation metrics: Top-N-accuracy or topN - if the correct category is among the N highest probabilities, we count the answer as correct (in this study N=3); and f1 score macro - a weighted average of recall and precision, where recall is the ratio of correctly predicted positive observations to all observations in the actual category, and precision is the ratio of correctly predicted positive observations to the total predicted positive observations. All three of these evaluation metrics assume that all errors are equally bad. For that reason we also measured the accuracy at the genus and family levels, and for snake dataset we report the accuracy of predicting a relevant biological trait, namely whether a snake species is venomous or not.

2.3.2 Results

Firstly, we evaluated the experiments with standard evaluation metrics. On the first dataset, one-hot encoding gave better results than LS on accuracy and top3-accuracy, but slightly worse on f1 scores with macro averaging. Systems trained using phylogenetically informed targets (TLS or PLS) with small smoothing values ($0.025 \leq \beta \leq 0.1$) gave slightly better results than both benchmarks, with exception of anagenetic-distance smoothing, which performed on par with the benchmarks. On the second dataset, the LS benchmark consistently gave better results than the one-hot benchmark. Results from experiments with TLS and PLS were consistently better than the one-hot benchmark. When compared to the second benchmark, LS, the inclusion of hierarchical information (TLS or PLS) with intermediate smoothing values ($0.05 \leq \beta \leq 0.4$) gave similar accuracy and f1 scores, while it often gave better top3.

Our phylogenetically or taxonomically informed approach performed better than both benchmarks on evaluation metrics that take into account the hierarchical information. Specifically, we found that TLS or PLS (based on anagenetic or cladogenetic distances), across a range of different smoothing values, yielded better results on the accuracy of identifications at the genus or family level, or the accuracy of predicting an important biological trait (a snake being venomous). Thus, even if the systems trained using TLS or PLS made the same number of errors as the benchmark reference systems, the errors tended to be less serious in that the misidentified categories involved organisms that were more closely related to each other.

2.4 Paper IV

2.4.1 Design of the study. Material and methods

The aim of **paper IV** was to describe lessons learned during Swedish Ladybird Project 2018 (SLP2018), a citizen-science project focused on collecting smartphone images in order to develop an ATI tool for Swedish species of ladybird beetles. The first phase, the citizen-science (CS) part, was organized in the summer of 2018. Initially, we aimed at schoolchildren (ages 6-16), but after the initial press release the project caught the attention of many local and national media and attracted a lot of interest from potential participants. Therefore, we decided to extend the project to include the interested public in general, and preschool kids (up to age 6) in particular. We offered teachers to register in advance to allow direct communication with the project team and to receive additional support. The preregistered classes were also provided with an "experiment kit", which included a guide for teachers, a macro-lens for mobile devices. and a recently published comprehensive field guide to Swedish ladybird species. In total, 700 experiment kits were dispensed among participants. The aim was to provide one kit per 15 participating kids, so larger classes or sets of classes received more than one kit. Contributions were submitted through an app specifically made for this project. For the identification of the ladybird species in the collected images, we relied on experts. After completion of the project an evaluation survey was sent out the registered participants.

For comparison purposes, we downloaded all the images of Swedish ladybirds available through the Global Biodiversity Information Facility (GBIF) GBIF (2020). The taxonomic identifications of the images provided by GBIF were taken at face value. GBIF images contributed in 2018 were used for the evaluation of ATIs ("GBIF2018"), while the remaining images (collected before 2018; "GBIF_training") were used as an additional training set that we could compare to the image set collected through the SLP2018 project. Specifically, to evaluate the SLP2018 contribution, we trained networks on SLP2018, GBIF_training and SLP2018+GBIF_training, and compared the identification performance of the resulting ATIs. We fine-tuned a well known and light architecture ResNet50 (He et al., 2015) from a publicly available checkpoint pretrained on ImageNet (Russakovsky et al., 2015). The finetunning procedure resembled the one used in **paper II** but with different hyperparameters.

As our evaluation metrics we used accuracy as described in **paper I** and f1 score macro as described in **paper III**. In addition to these common metrics, we used f1 score macro on subsets of species created based on the number of images per species on GBIF: with the two dominant species, *Harmonia axyridis* and *Coccinella septempunctata*; other common species (ranked 3-10 in abundance); and the remaining species.

2.4.2 Results

Almost 400 teachers registered to participate in the project. According to the 24 replies we received in the evaluation survey, many of them supervised more than one class. In average, the number of students per registered teacher was 31.5 (range 16-67), or in total around 12,000 children. If we consider unregistered participants, which also include schools and preschools, and expeditions for kids organized by amateur entomologists and natural history museums, we estimate that the total number of children participating in the project was around 15,000.

The registered teachers were instructed to spend 1-2 hours of effort on field work. The survey revealed that most of the groups had searched ladybirds multiple times (up to 28 times) and they mostly invested up to 2h (80%) on the task on each field trip. Perhaps even more astonishing was that some groups, presumably preschool children, had spent orders of magnitude more time than expected (replies included “every day”, “all the time throughout the project period” and “many times”).

In total, we received over 5,000 images for almost half of the Swedish ladybird species

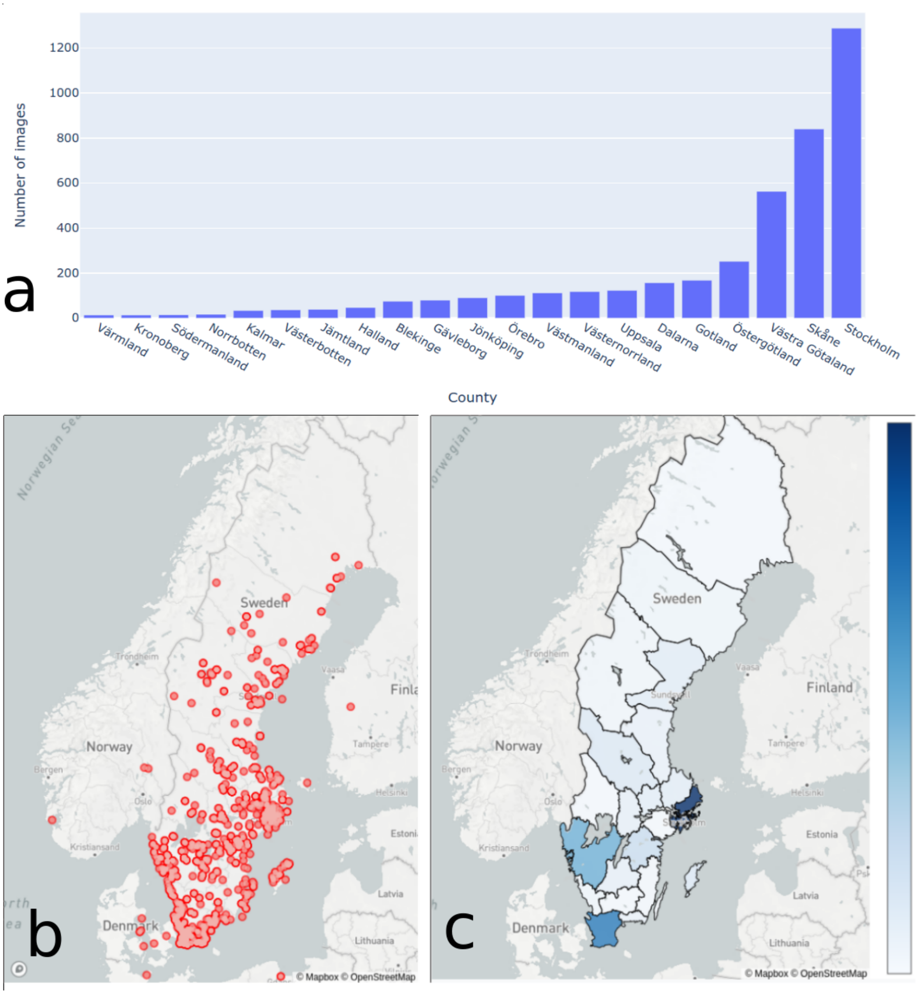


Figure 2.5: Geographical distribution of the contributed images. Most of the images came from the most populated counties Stockholm, Skåne and Västra Götaland (a). In (b) we show locations where images are taken (each dot represents a single image) and in (c) we show the normalized contribution per capita for each of the 21 counties (darker is more). Adapted from **paper IV**.

(30/71). The first images were submitted in early June, followed by the first peak during the first three weeks of summer (weeks 26-28). The summer was calm until weeks 35-38, which brought the biggest contributions.

The contributions were received from all of the 21 counties of Sweden. The most populated counties—Stockholm, Västra Götaland and Skåne—contributed most of the images. However, when accounting for population size, Gotland stood out with 7 times as many contributions per capita as the country's average.

The majority of the insects from contributed photos (68%) were identified as *Coccinella septempunctata*. Similar patterns were found in the individual months, with *C. septumpunctata* comprising between 60 and 80% of the monthly contributions. In addition to *C. septempunctata*, only three more species were represented by more than 3% of the total number of images: *Psyllobora vigintiduopunctata* (7.7%), *Harmonia axyridis* (5.8%) and *Adalia bipunctata* (5.4%).

The experiments designed to analyze the contribution of SLP2018 in the context of developing ATIs showed that the estimated value of the SLP2018 data heavily depended on the choice of evaluation metric. The SLP2018 data, when compared to the GBIF data, did not contribute significantly to improving the accuracy or top3 scores. These metrics favor majority categories (such as *C. septempunctata* and *H. axyridis* in our study) and these were already well represented in the GBIF dataset with 80% of all the ladybirds images on GBIF belonging to one of these two species.

However, adding SLP2018 data to GBIF_training notably improved our second metric, F1 score with macro averaging. The improvement came from minority categories, and we could see that by comparing F1 scores over three subsets of species. The scores for the first subset, comprising the two dominant species (*C. septempunctata* and *H. axyridis*), remained largely unchanged. The scores for the second subset, containing the remainder of the common species, were slightly but clearly improved, while the scores for the last subset, containing the rare (minority) species, benefited the most from the addition of the SLP2018 images.

Chapter 3

Discussion

The scientists of today think deeply instead of clearly. One must be sane to think clearly, but one can think deeply and be quite insane.

Nikola Tesla

Our results from **paper I** show that there is considerable room for optimization of current DL techniques so that they generate better ATIs in settings that are typically encountered by taxonomists. Surprisingly, our experiments in **paper I** show that it is possible to find strategies that significantly boost the performance of ATI systems across a range of taxonomic tasks and datasets, and these strategies seem to be successful also on related tasks. Specifically, the results obtained in **paper I** are based on a computationally efficient approach, namely feature transfer from a single feed-forward pass through a publicly available checkpoint of the VGG16 (Simonyan and Zisserman, 2014) system pretrained on the ImageNet classification task (Russakovsky et al., 2015). When compared to a baseline application of this approach, we were able to significantly improve the performance of the resulting ATI by introducing each of the following modifications or additions: larger image input size *416x416*, global average pooling, fusion of features from all five convolutional blocks and signed square root normalization.

Our results also indicate that it may be possible to further improve identification performance by optimizing the dimensionality reduction. In our experiments, global pooling (1x1xF) was not always optimal; pooling to small but not minimal feature matrices (2x2xF or 4x4xF) often resulted in better performance. Specifically, intermediate pooling resulted in significant improvements on three out of four datasets.

It turned out that intermediate-level features extracted from the *c4* layer were the most useful ones for the taxonomic tasks we tackled in **paper I**. According to Zheng et al. (2016), this is not the case for generic image classification tasks (e.g. Caltech-101 (Li Fei-Fei et al., 2006), Caltech-256 (Griffin et al., 2007), and PASCAL VOC07 (Everingham et al., 2015)), nor for popular fine-grained identification tasks (Bird-200-2011 (Wah et al., 2011) and Flower-102 (Nilsback and Zisserman, 2008)), where the most discriminating layers are the first fully connected layer and *c5*, respectively. A possible reason for this discrepancy could be that our images are more distant from the ImageNet dataset than any of the above-mentioned image datasets, or that differences between categories in our datasets are even more subtle than those in the previous fine-grained identification tasks.

From the perspective of insect taxonomy, perhaps the most interesting insight from **paper I** is that, with optimized feature transfer, it is possible to develop high-performance ATIs for a wide range of tasks without expertise in image analysis or machine learning. Developing ATIs with identification accuracy on par with or exceeding that of human experts clearly seems to be within reach for non-experts. Thus, many taxonomists should be able themselves to leverage the latest advances in CNNs and deep learning to develop high-performance ATIs for just about any classification task of interest in insect systematics.

In **paper II**, we examined the potential of the system with the optimal feature transfer protocol identified in **paper I** in testing a set of new taxonomic hypotheses. With no further optimization, our off-the-shelf solution from **paper I** proved to be very useful in tackling several challenging identification tasks, some of which were previously thought to be unsolvable by humans. A particularly striking example is the identification of sex from images of the dorsal habitus of *Oxythyrea* beetles; sexual dimorphism in dorsal habitus has not been noted previously in these beetles.

It is perhaps not surprising that the amount of data is important for good performance of machine identification (the more examples, the better; **paper I**) or that the difficulty increases with the amount of morphological variation within categories (sexual dimorphism, life stages, age, etc.; **paper I**, **II**). Thus, collecting a sufficient number of accurately labeled images for developing a sufficiently competent ATI is important, but may appear to be a difficult task, particularly since these image sets often have to be assembled essentially from scratch for many taxonomic tasks. Fortunately, our results (**paper II**) indicate that, at least for some tasks, images collected with a smartphone and a cheap attachable lens are quite sufficient. Thus, photographing specimens with a smartphone sometimes offers an interesting alternative to the time-consuming procedure of collecting high resolution images in a standardized taxonomic imaging setup. We estimated that by using a smartphone, one may collect 40x more images in the same amount of time as that required for the traditional setup. It is worth noting that an ATI based on a larger set of smartphone images may actually perform better than a system trained on a smaller number of high-resolution images. The smartphone approach is also much easier to scale as everyone has a smartphone today, while the setup used for high-resolution imaging of insects is mostly restricted to entomologists at larger institutions and requires a substantial investment in equipment.

In both **paper I** and **II**, we observed that machines and humans tend to find the same or similar tasks challenging. Good examples are separating visually similar species (sibling species) (**paper I, II**) or placing a species to the correct higher taxonomic group (family or tribe in our case) when the subgroup it belongs to (species and genus in our case) has not been seen before (**paper I**). However, we found occasional cases where machines made catastrophic errors that no human would make, something we tried to address in **paper III** (see below).

An off-the-shelf solution can be sufficient for developing an ATI with adequate performance for many tasks, as shown in **papers I-II**, but even when this is not the case, such a solution is useful in testing whether an identification task is feasible at all, and whether it is worth spending more resources on developing a more sophisticated system. Going beyond off-the-shelf and utilizing state-of-the-art techniques can improve AI performance, but also give additional insights, as illustrated by our heat maps (class activation maps). This technique highlights the regions of images containing the features that a model finds informative for a given task (Fig. 2.3). Our experiments (**paper II**) indicate that human experts often find features occurring in the highlighted regions useful for species discrimination. However, these heat maps are not always easy to interpret. They do not tell us exactly what the discriminating features are but rather indicate where they are located. For example, consider an image of a beetle specimen, and assume that a region of the specimen with a lot of hairs is highlighted by the heat map (as in our experiments). This does not tell us what feature of the hairs is important for discrimination (their number, distribution, thickness, length or something else). In fact, we cannot even be sure that the discriminating feature has anything to do with the hairs, it could be some underlying feature, such as the shape or size of the hairy body part. Therefore, we should not expect the heat maps to point exactly to an interesting morphological feature but rather to give us hints of where to look. Then, it is up to the expert to hypothesize what the actual discriminating feature is.

Another interesting phenomenon that we were able to demonstrate with the heat maps was the fact that CNNs learn from whole images, including information that may be irrelevant for the task. Specifically, the heat maps from one of the species studied in **paper** highlighted the region where the pin is located. It turned out that all the imaged specimens of this species were collected by a single entomologist, who was pinning them from a non-standard angle, and this entomologist did not contribute any specimens of the other species. This is an excellent example of how bias can be introduced in image training sets.

Intuitively, one might expect that more sophisticated training of AI systems could reduce the risk of such irrelevant biases, or the risk of making catastrophic errors, that is, misidentifying categories that are completely unrelated. In **paper III** we explore this idea by proposing to utilize phylogenetic relationships among biological organisms (TLS and PLS) in the training of ATIs. Our solution is inspired by the recently introduced LS (Szegedy et al., 2016) technique, which uses a weighted average of one-hot targets and uniform distribution over labels during training. However, unlike both LS and one-hot encoding, TLS and PLS use target encodings based on non-uniform distributions. Specifically, they mix one-hot encoding with a small proportion of a distribution based on a hierarchical scoring scheme reflecting either taxonomic (TLS) or phylogenetic (PLS) information. This rewards the system being trained for predicting the correct category, while punishing it for errors in proportion to how much these errors violate what we know about the biological relationships among the categories. The technique we propose could easily be applied also in other cases where there is a natural hierarchy describing the similarity relations among the categories. As demonstrated by our results, systems trained in this way tend to make less serious errors. For this reason, hierarchically aware systems might be preferred in many practical applications over standard systems, even when they make more errors in total. Examples may include cases where the cost of making an error is very high, such as misidentifying a venomous snake for a non-venomous one.

The results presented in **paper III** indicate that systems trained with phylogenetic information often perform on par with or better than baseline systems in terms of common evaluation metrics, such as accuracy, topK-accuracy and f1 score macro. When evaluated with custom metrics that take the biological context into account (accuracy on the genus and family levels, or accuracy in predicting a biological trait) we observed that the systems trained using TLS or PLS often outperformed both of the baseline systems. In our experiments, a range of smoothing ratios (β values) was tested, and although 0.05 and 0.1 seemed to perform well on both datasets, we noticed that applying the same smoothing ratio across tasks might be suboptimal given the current methodological setup. The reason for this is possibly the difference in the number of categories across tasks and the fact that we distribute total smoothing value β across all categories. As a result, tasks with a larger number of categories have smaller average smoothing value per category. A possible solution could be to distribute the smoothing only across a fixed number of the most related categories and keep the remaining categories at zero in the smoothing distribution.

Assembling sufficiently large training sets of accurately labeled images will remain a challenge in the development of ATIs for many organism groups. Using citizen-science contributions is an obvious possibility, and our attempt in this direction (**paper IV**) exceeded all expectations. We estimated that approximately 15,000 children participated in the project. Given our conservative estimate that each child devoted 1 hour to the task, this translates to a contribution of approximately 15,000 hours (375 weeks or 83 person months) of effort to the project. If a single person would be required to perform this task by himself/herself, it would take that person more than 10 years to complete it (1,600 annual working hours in Sweden (Charlie Giattino and Roser, 2013)).

Over 5,000 images were received in two major waves, one in the early summer and one in the late summer. These waves were anticipated as a result of the design of the project, which closely followed the biology of ladybirds in Sweden. The adult ladybirds become active in May-June searching for food and to mate, then they slow down until the end of the summer when the new generation emerges, and joins the adult beetles who overwintered in the search for food to stack reserves for the upcoming winter. The natural life-history fluctuation in adult beetle activity were probably further accentuated by the timing of project activities. The first wave coincided with the project kick-off and the launch of an extensive project marketing campaign. We did have participants that contributed images during the whole summer (preschools, field stations, museums and groups of amateur entomologists), but the image contributions were fewer during this period. Finally, the majority of the contributed images were received in the second wave in the late summer, when schoolchildren returned to school and teachers had scheduled project activities in the field.

The results were impressive in terms of both species coverage and taxonomic composition. In total, 30 of the 71 naturally occurring coccinellid species were photographed. The most common species were photographed the most, while some rare species were completely absent from the contributed images. An interesting finding was that *Harmonia axyridis*, an invasive species that was first recorded from Sweden a decade ago (Brown et al., 2007), was the third most commonly photographed species. As one might expect, we noted some biases toward species that thrive in urban environments and against those that occur in habitats less suitable for field excursions with children or in less populated parts of the country (northern species).

The 5,119 images collected by the SLP2018 project represent a significant expansion of what was at the time available through GBIF. SLP2018 images increased the number of Swedish images of ladybird beetles by a factor of 50. The number of SLP2018 images approached the number of images submitted that year to GBIF from around the world (7,264). and represent 21% of all GBIF images collected over the last 20 years from all over the globe of species of ladybird beetles occurring in Sweden. These numbers clearly show the potential of citizen-science projects in quickly assembling sizeable sets of images for suitable organism groups.

Our experiments also confirm that the SLP2018 data provided valuable information for training ATIs. The increase in identification accuracy was particularly obvious for the less common species, the ones for which the number of images available previously was insufficient for adequate training of ATIs. It is interesting to speculate on how useful citizen-science projects might be in general for assembling suitable image sets for training ATIs. Clearly, the most significant boosts will be expected for organism groups and species for which there are few existing images. The group also needs to be fairly easy to identify and to photograph for the participating citizen scientists. We think that a fair number of groups of insects, other invertebrates, fungi and plants may fit these criteria. For some of the more obscure groups, it may sometimes be possible to find more specialized amateur naturalists that are willing to contribute, while other groups are suitable for larger-scale projects targeting non-specialists as in the SLP2018 project. An interesting lesson is the outsize contribution of the images of the rare species and of the rarely photographed subgroups, such as some of the subfamilies of Coccinellidae that were poorly represented in both the SLP2018 and GBIF image sets. It may well be possible to increase the effectiveness of citizen-science projects by directing the attention of participants towards these subgroups or species, perhaps by considering various point reward systems. Clearly, citizen science provides many valuable opportunities for advancing the development of ATIs. At the same time, these projects can also help raise the awareness of the value of biodiversity and have many other positive societal side effects.

Chapter 4

Concluding remarks and thoughts on not so far future

The present is theirs; the future, for which I really worked, is mine.

Nikola Tesla

The research presented in this thesis demonstrates how current AI technologies — specifically CNNs and DL — can solve several types of challenging taxonomic identification tasks (**papers I-II**). We managed to develop an optimized feature extraction protocol that was successful in tackling a broad range of taxonomic identification tasks, making it possible to offer an easy-to-use instant-return approach for taxonomists interested in investigating the feasibility of using AI in addressing particular identification problems. Specifically, we demonstrated the usefulness of our optimized solution in: (1) identifying insects to higher groups when they are likely to belong to subgroups that have not been seen previously; (2) identifying visually similar species that are difficult to separate even for experts; and (3) solving some identification tasks that appear unsolvable to humans, such as detecting the sex of a specimen when there is no prior evidence in the literature of sexual dimorphism.

We also showed that going beyond such an off-the-shelf solution by utilizing state-of-the-art AI techniques can yield further improvements in identification accuracy while providing additional insights. A particularly striking example of the latter is provided by the heat maps presented in **paper II**, highlighting regions that are particularly important for the ATI in discriminating among categories of imaged objects. This type of approach may well turn out to be a powerful tool that can guide experts in identifying the informative morphological features that are critical for both taxonomic and phylogenetic research.

In **paper III**, we demonstrated that by leveraging taxonomic or phylogenetic information we can train systems that are as good as standard systems in terms of common evaluation metrics (accuracy, topK accuracy, f1 score macro), but make errors that are less wrong in an objective biological sense (**paper III**). An obvious research direction inspired by this paper would be to explore the use of CNNs for phylogenetic inference. My initial experiments show that models optimized only toward targets based on phylogenetic information are quite good at approximating the position of species in the tree of life. Of course, this requires that a reference phylogeny is available for training. However, there are numerous ways in which such phylogeny-trained AI systems could be useful. For instance, a system trained using an approximate or incomplete reference phylogeny could learn to distinguish phylogenetically important features, and thus refine the phylogeny in the way it perceives similarities among images. It should also be able to place unseen species in a phylogenetic context more accurately than other systems. It may also turn out that AI "knowledge" about what image features structure the phylogeny of one group of organisms can transfer to related groups. If so, then an AI system trained to reconstruct phylogeny may well be able to infer phylogeny of a group of organisms it has not seen previously using only images of those organisms.

We already mentioned that the credit for recent developments in computer vision goes to the faster and cheaper computational power, improved algorithms, and increasing amounts of image training data available on the internet. We are witnessing constant, fast-paced improvements in the first two, regardless of what happens in the taxonomic community. However, the taxonomic community could contribute substantially to the field making even bigger strides forward by generating sufficient amounts of accurately labeled image data for training of AI systems. An obvious way to generate more data is to use CS efforts similar to our project described in **paper IV**. However, the CS path does not seem a viable solution to the problem of accumulating adequate training sets for a signification number of insect species. Some organism groups are simply too difficult for citizen scientists to find in the field, or to photograph in such a way that it will be possible to identify the imaged species. In some cases one may be able to extend the limits of what can be done by focusing on specialized amateur naturalists, to train amateur naturalists to work with new groups, or to elicit the help of CS in imaging existing collections of specimens. However, in the end, accumulating sufficient training data has to be achieved using other approaches for many organism groups.

An interesting alternative approach to CS would be to use camera traps for accumulating images for training. Camera traps have the additional advantage that they can also be used for continuous monitoring. There is already considerable experience with camera traps for mammals, and with developing AI techniques for these data. For example, in the Snapshot Serengeti project (Swanson et al., 2015), CNN-based models have now been developed that are better than humans in terms of accuracy at recognizing common species (zebra, lion, elephant)(Norouzzadeh et al., 2018; Willi et al., 2019). There are also several successful examples targeting insects, mainly driven by industrial applications (e.g. see the recent review paper on automated monitoring of pests (Cardim Ferreira Lima et al., 2020) and references therein). However, as in the CS approach, we need to find ways how to generate labels for the images we collect. For instance, by the year 2015, Snapshot Serengeti had accumulated over 1,500,000 labeled images by using an army of 100,000 volunteers. In this project, it was found that having an image labeled by one volunteer gave only 85% accuracy in the identification. This improved further to 95% and 98% accuracy when using the voting consensus of 10 and 20 volunteers, respectively (see Swanson et al. (2015) for details of the setting). A single expert could label images with 96.6% accuracy, CNNs alone reached 97% accuracy, and CNN+expert in a human-in-the-loop approach reached 99.8% accuracy. Unfortunately, insect identification is considerably more challenging for humans than mammal identification for several reasons, including the enormous diversity of insects, and becoming an expert on even a single group can require a lifetime of training.

Thus, entomologists need to be more creative than mammalogists in order to get over this hump caused by the need for sizeable collections of accurately labeled training data.

Most of the challenges we find in building AI identification systems for taxonomy are not unique to our field. In fact, we can count on constant advancements in DL and computer vision to also benefit the construction of ATIs. For instance, there has recently been significant progress in utilizing unlabeled images in training (Chen et al., 2020; He et al., 2020), which considerably reduces the need for labeled data at the expense of computational resources. Because computational resources are getting cheaper and faster, while human labor costs for labeling are increasing, this seems a particularly promising future direction for the development of ATIs and many other AI applications. Another research area of potential interest to taxonomists, where the DL community is very active, is so-called automated machine learning (AutoML) (see (He et al., 2021) and references therein). The idea is to build solutions based on cutting-edge techniques but without any human involvement, thus completely removing the need for machine learning expertise. In other words, the idea is to completely automate data preparation, feature engineering, hyper-parameter optimization, and neural architecture search. There are many other promising ideas currently pursued by the DL community, which will, sooner or later, affect many aspect of taxonomy and transform the way we work with species discovery, description and identification.

My impression is that the time is ripe for biologists experimenting with AI techniques to set bigger goals that could provide more substantial benefits and bring more transformative changes. For instance, imagine a robot with a camera, which could capture specimens, automatically detect them and recognize most of the species, collect behavioural data, and then, depending on a predefined set of rules, select specimens to be collected and preserved for future study, eliminating specimens of invasive or harmful species, and releasing the remainder. How about nano-robots small enough so that millions could fit on a single head of an insect pin; if they can flow through our bloodstreams, detect and cure diseases then it might be possible to have similar nano-robots fly around in the wild, detect an interesting specimen, collect behavioural data and sample tissue for genetic analysis in a non-destructive way. The recent progress in technology gives me confidence that these types of systems might be feasible in the near future, and I hope I will be able to contribute to the journey towards building them.

In conclusion, today's ”machines” can see, sometimes better than us, they learn much faster, they never forget and their knowledge is transferable to new tasks in similar domains. Technology enables us already today to automatize many aspects of taxonomic work, and it is only a matter of time until we will be able to put together some pieces of technology already perfected in various other industries to achieve even more impressive tasks. Taking advantage of these opportunities will enable us to progress more forcefully than ever toward the goal of saving the Earth. Despite all the challenges, I think the future is bright, as those challenges are nothing but new opportunities.

Bibliography

Agnarsson,I. and Kuntner,M. Taxonomy in a changing world: seeking solutions for a science in crisis. *Systematic Biology*, 56(3):531–539, 2007.

Arandjelovic,R. and Zisserman,A. All about VLAD. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pages 1578–1585. cv-foundation.org, 2013.

Arbuckle,T., Schröder,S., Steinhage,V., and Wittmann,D. Biodiversity informatics in action: identification and monitoring of bee species using ABIS. In *Proc. 15th Int. Symp. Informatics for Environmental Protection*, volume 1, pages 425–430. enviroinfo.eu, 2001.

Azizpour,H., Razavian,A. S., Sullivan,J., Maki,A., and Carlsson,S. Factors of transferability for a generic ConvNet representation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 38 (9):1790–1802, September 2016.

Bengio,Y. Deep learning of representations for unsupervised and transfer learning. In *Proceedings of ICML workshop on unsupervised and transfer learning*, pages 17–36, 2012.

Bottou,L., Curtis,F. E., and Nocedal,J. Optimization methods for large-scale machine learning. *Siam Review*, 60(2):223–311, 2018.

Breiman,L. Random forests. *Machine learning*, 45(1):5–32, 2001.

Brown,P., Adriaens,T., Bathon,H., Cuppen,J., Goldarazena,A., Hägg,T., Kenis,M., Klausnitzer,B., Kovář,I., Loomans,A., et al. Harmonia axyridis in europe: spread and distribution of a non-native coccinellid. In *From Biological Control to Invasion: the Ladybird Harmonia axyridis as a Model Species*, pages 5–21. Springer, 2007.

Cardim Ferreira Lima,M., Damascena de Almeida Leandro,M. E., Valero,C., Pereira Coronel,L. C., and Gonçalves Bazzo,C. O. Automatic detection and monitoring of insect pests—a review. *Agriculture*, 10(5):161, 2020.

Carranza-Rojas,J., Goeau,H., Bonnet,P., Mata-Montero,E., and Joly,A. Going deeper in the automated identification of herbarium specimens. *BMC Evol. Biol.*, 17(1):181, August 2017.

Caruana,R. Learning many related tasks at the same time with backpropagation. In *Advances in neural information processing systems*, pages 657–664, 1995.

Ceballos,G., Ehrlich,P. R., and Dirzo,R. Biological annihilation via the ongoing sixth mass extinction signaled by vertebrate population losses and declines. *Proceedings of the national academy of sciences*, 114(30):E6089–E6096, 2017.

Chapman,A. D. et al. Numbers of living species in australia and the world. 2009.

Charlie Giattino,E. O.-O. and Roser,M. Working hours. *Our World in Data*, 2013. https://ourworldindata.org/working-hours.

Chen,T., Kornblith,S., Norouzi,M., and Hinton,G. A simple framework for contrastive learning of visual representations. *arXiv preprint arXiv:2002.05709*, 2020.

Cireşan,D., Meier,U., Masci,J., and Schmidhuber,J. A committee of neural networks for traffic sign classification. In *The 2011 international joint conference on neural networks*, pages 1918–1921. IEEE, 2011.

Coleman,C. O. Taxonomy in times of the taxonomic impediment–examples from the community of experts on amphipod crustaceans. *Journal of Crustacean Biology*, 35(6): 729–740, 2015.

Cortes,C. and Vapnik,V. Support-vector networks. *Mach. Learn.*, 20(3):273–297, September 1995.

Cox,D. R. The regression analysis of binary sequences. *Journal of the Royal Statistical Society: Series B (Methodological)*, 20(2):215–232, 1958.

Cui,Y., Song,Y., Sun,C., Howard,A., and Belongie,S. Large scale fine-grained categorization and domain-specific transfer learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

Dirzo,R., Young,H. S., Galetti,M., Ceballos,G., Isaac,N. J., and Collen,B. Defaunation in the anthropocene. *science*, 345(6195):401–406, 2014.

Donahue,J., Jia,Y., Vinyals,O., Hoffman,J., Zhang,N., Tzeng,E., and Darrell,T. DeCAF: A deep convolutional activation feature for generic visual recognition. In *International Conference on Machine Learning*, pages 647–655. jmlr.org, January 2014.

Ebach,M. C., Valdecasas,A. G., and Wheeler,Q. D. Impediments to taxonomy and users of taxonomy: accessibility and impact evaluation. *Cladistics*, 27(5):550–557, 2011.

Ehrlich,P. R. The scale of human enterprise and biodiversity loss. *Extinction rates*, 1995.

Everingham,M., Eslami,S. M. A., Van-Gool,L., Williams,C. K. I., Winn,J., and Zisserman,A. The Pascal Visual Object Classes Challenge: A Retrospective. *International Journal of Computer Vision*, 111(1):98–136, 2015.

Feng,L., Bhanu,B., and Heraty,J. A software system for automated identification and retrieval of moth images based on wing attributes. *Pattern Recognit.*, 51:225–241, March 2016.

Francoy,T. M., Wittmann,D., Drauschke,M., Müller,S., Steinhage,V., Bezerra-Laure,M. A. F., De Jong,D., and Gonçalves,L. S. Identification of africanized honey bees through wing morphometrics: two fast and efficient procedures. *Apidologie*, 39(5):488–494, September 2008.

Friedman,J. H. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232, 2001.

Fukushima,K. Neocognitron: a self organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological cybernetics*, 36(4):193–202, 1980. URL `http://www.ncbi.nlm.nih.gov/pubmed/7370364`.

Fukushima,K. Neural network model for a mechanism of pattern recognition unaffected by shift in position-neocognitron. *IEICE Technical Report, A*, 62(10):658–665, 1979.

Fukushima,K., Miyake,S., and Ito,T. Neocognitron: A neural network model for a mechanism of visual pattern recognition. *IEEE Transactions on Systems, Man, and Cybernetics*, SMC-13(5):826–834, sep 1983. ISSN 0018-9472. doi: 10.1109/TSMC.1983.6313076. URL `http://ieeexplore.ieee.org/document/6313076/`.

Gaston,K. J. and O'Neill,M. A. Automated species identification: why not? *Philosophical Transactions of the Royal Society B: Biological Sciences*, 359(1444):655–667, apr 2004. ISSN 0962-8436. doi: 10.1098/rstb.2003.1442. URL `http://rstb.royalsocietypublishing.org/content/359/1444/655`.

Gaston,K. J. and May,R. M. Taxonomy of taxonomists. *Nature*, 356(6367):281–282, mar 1992. ISSN 0028-0836. doi: 10.1038/356281a0. URL `http://dx.doi.org/10.1038/356281a0`.

Gauld,I. D., O'Neill,M. A., Gaston,K. J., and Others. Driving miss daisy: the performance of an automated insect identification system. *Hymenoptera: evolution, biodiversity and biological control*, pages 303–312, 2000.

GBIF. Home page. Available from: https://www.gbif.org, 2020. Accessed: November 2020.

Gonçalves,A. B., Souza,J. S., Silva,G. G. d., Cereda,M. P., Pott,A., Naka,M. H., and Pistori,H. Feature extraction and machine learning for the classification of brazilian savannah pollen grains. *PLoS One*, 11(6):e0157044, June 2016.

Griffin,G., Holub,A., and Perona,P. Caltech-256 Object Category Dataset. Technical report, 2007.

Hansen,O. L., Svenning,J.-C., Olsen,K., Dupont,S., Garner,B. H., Iosifidis,A., Price,B. W., and Høye,T. T. Species-level image classification with convolutional neural network enables insect identification from habitus images. *Ecology and evolution*, 10(2):737–747, 2020.

Hanson,S. J. A stochastic version of the delta rule. *Physica D: Nonlinear Phenomena*, 42 (1-3):265–272, 1990.

He,H. and Garcia,E. A. Learning from imbalanced data. *IEEE Transactions on knowledge and data engineering*, 21(9):1263–1284, 2009.

He,K., Zhang,X., Ren,S., and Sun,J. Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, 37(9):1904–1916, September 2015.

He,K., Zhang,X., Ren,S., and Sun,J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

He,K., Fan,H., Wu,Y., Xie,S., and Girshick,R. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

He,X., Zhao,K., and Chu,X. Automl: A survey of the state-of-the-art. *Knowledge-Based Systems*, 212:106622, 2021. ISSN 0950-7051. doi: https://doi.org/10.1016/j.knosys.2020.106622. URL `http://www.sciencedirect.com/science/article/pii/S0950705120307516`.

Hinton,G., Deng,L., Yu,D., Dahl,G., Mohamed,A.-r., Jaitly,N., Senior,A., Vanhoucke,V., Nguyen,P., Sainath,T., and Kingsbury,B. Deep Neural Networks for Acoustic Modeling in Speech Recognition: The Shared Views of Four Research Groups. *IEEE Signal*

Processing Magazine, 29(6):82–97, nov 2012. ISSN 1053-5888. doi: 10.1109/MSP.2012.2205597. URL http://ieeexplore.ieee.org/document/6296526/.

Hu,J., Shen,L., and Sun,G. Squeeze-and-excitation networks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

Ioffe,S. and Szegedy,C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*, 2015.

Jean,S., Cho,K., and Memisevic,R. On Using Very Large Target Vocabulary for Neural Machine Translation. In *ACL-IJCNLP*, pages 1–10, 2015.

Joly,A., Goëau,H., Botella,C., Glotin,H., Bonnet,P., Vellinga,W.-P., Planqué,R., and Müller,H. Overview of lifeclef 2018: a large-scale evaluation of species identification and recommendation algorithms in the era of ai. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, pages 247–266. Springer, 2018.

Karpathy,A. and Fei-Fei,L. Deep Visual-Semantic Alignments for Generating Image Descriptions. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3128–3137, 2015.

Kelley,H. J. Gradient theory of optimal flight paths. *Ars Journal*, 30(10):947–954, 1960.

Kiefer,J., Wolfowitz,J., et al. Stochastic estimation of the maximum of a regression function. *The Annals of Mathematical Statistics*, 23(3):462–466, 1952.

Kingma,D. P. and Ba,J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Krizhevsky,A., Sutskever,I., and Hinton,G. E. ImageNet Classification with Deep Convolutional Neural Networks. *Advances In Neural Information Processing Systems*, pages 1–9, 2012. ISSN 10495258.

Kumar,A., Irsoy,O., Ondruska,P., Iyyer,M., Bradbury,J., Gulrajani,I., Zhong,V., Paulus,R., and Socher,R. Ask Me Anything: Dynamic Memory Networks for Natural Language Processing. jun 2016. URL `http://arxiv.org/abs/1506.07285`.

Laliberte,A. S. and Ripple,W. J. Range contractions of north american carnivores and ungulates. *BioScience*, 54(2):123–138, 2004.

LeCun,Y., Boser,B., Denker,J. S., Henderson,D., Howard,R. E., Hubbard,W., and Jackel,L. D. Backpropagation Applied to Handwritten Zip Code Recognition. *Neural Computation*, 1(4):541–551, dec 1989. ISSN 0899-7667. doi: 10.1162/neco.1989.1.4.541. URL `http://www.mitpressjournals.org/doi/abs/10.1162/neco.1989.1.4.541`.

LeCun,Y., Bengio,Y., and Hinton,G. Deep learning. *nature*, 521(7553):436–444, 2015.

Li Fei-Fei, Fergus,R., and Perona,P. One-shot learning of object categories. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, volume 28, pages 594–611, apr 2006. doi: 10.1109/TPAMI.2006.79. URL `http://ieeexplore.ieee.org/document/1597116/`.

Lin,M., Chen,Q., and Yan,S. Network in network. December 2013.

Linnainmaa,S. Taylor expansion of the accumulated rounding error. *BIT Numerical Mathematics*, 16(2):146–160, 1976.

Lytle,D. A., Martínez-Muñoz,G., Zhang,W., Larios,N., Shapiro,L., Paasch,R., Moldenke,A., Mortensen,E. N., Todorovic,S., and Dietterich,T. G. Automated processing and identification of benthic invertebrate samples. *J. North Am. Benthol. Soc.*, 29(3): 867–874, June 2010.

Marcus,L. F. Traditional morphometrics. *Proceedings of the Michigan morphometrics*, 2:77–122, 1990. URL `https://www.researchgate.net/profile/F{%}7B{_}{%}7DRohlf/publication/30850286{%}7B{_}{%}7DProceedings{%}7B{_}{%}7Dof{%}7B{_}{%}7Dthe{%}7B{_}{%}7DMichigan{%`
`links/5566227508aefcb861d1971b.pdf{%}7B{#}{%}7Dpage=87`.

Martineau,M., Conte,D., Raveaux,R., Arnault,I., Munier,D., and Venturini,G. A survey on image-based insects classification. *Pattern Recognit.*, 2016.

Maxwell,S. L., Fuller,R. A., Brooks,T. M., and Watson,J. E. Biodiversity: The ravages of guns, nets and bulldozers. *Nature News*, 536(7615):143, 2016.

Mora,C., Tittensor,D. P., Adl,S., Simpson,A. G., and Worm,B. How many species are there on earth and in the ocean? *PLoS biology*, 9(8):e1001127, 2011.

Nilsback,M.-E. and Zisserman,A. Automated flower classification over a large number of classes. In *Indian Conference on Computer Vision, Graphics and Image Processing*, Dec 2008.

Norouzzadeh,M. S., Nguyen,A., Kosmala,M., Swanson,A., Palmer,M. S., Packer,C., and Clune,J. Automatically identifying, counting, and describing wild animals in camera-trap images with deep learning. *Proceedings of the National Academy of Sciences*, 115 (25):E5716–E5725, 2018.

Novotny,V., Basset,Y., Miller,S. E., Weiblen,G. D., Bremer,B., Cizek,L., and Drozd,P. Low host specificity of herbivorous insects in a tropical forest. *Nature*, 416(6883):841, 2002.

O'Neill,M. A. DAISY: a practical tool for semi-automated species identification. *Automated taxon identification in systematics: theory, approaches, and applications. CRC Press/Taylor & Francis Group, Boca Raton/Florida*, pages 101–114, 2007.

Oquab,M., Bottou,L., Laptev,I., and Sivic,J. Learning and transferring mid-level image representations using convolutional neural networks. *Proc. IEEE*, 2014.

Ripple,W. J., Estes,J. A., Beschta,R. L., Wilmers,C. C., Ritchie,E. G., Hebblewhite,M., Berger,J., Elmhagen,B., Letnic,M., Nelson,M. P., et al. Status and ecological effects of the world's largest carnivores. *Science*, 343(6167), 2014.

Robbins,H. and Monro,S. A stochastic approximation method. *The annals of mathematical statistics*, pages 400–407, 1951.

Rodman,J. E. and Cody,J. H. The taxonomic impediment overcome: Nsf's partnerships for enhancing expertise in taxonomy (peet) as a model. *Systematic biology*, 52(3):428–435, 2003.

Rodner,E., Simon,M., Brehm,G., Pietsch,S., Wolfgang Wägele,J., and Denzler,J. Fine-grained recognition datasets for biodiversity analysis. July 2015.

Rohlf,J. F. and Marcus,L. F. A revolution morphometrics. *Trends in ecology & evolution*, 8(4):129–32, apr 1993. ISSN 0169-5347. doi: 10.1016/0169-5347(93)90024-J. URL `http://www.sciencedirect.com/science/article/pii/016953479390024Jhttps://www.researchgate.net/publication/49756524{%}7B{_}{%}7DA{%}7B{_}{%}7DRevolution{%}7B{_}{%}7Din{%}7B{_}{%}7DMorphometrics`

Rosenblatt,F. *The perceptron, a perceiving and recognizing automaton Project Para*. Cornell Aeronautical Laboratory, 1957.

Rumelhart,D. E., Hinton,G. E., and Williams,R. J. Learning representations by back-propagating errors. *nature*, 323(6088):533–536, 1986.

Russakovsky,O., Deng,J., Su,H., Krause,J., Satheesh,S., Ma,S., Huang,Z., Karpathy,A., Khosla,A., Bernstein,M., Berg,A. C., and Fei-Fei,L. ImageNet large scale visual recognition challenge. *Int. J. Comput. Vis.*, 115(3):211–252, December 2015.

Sainath,T. N., Mohamed,A.-r., Kingsbury,B., and Ramabhadran,B. Deep convolutional neural networks for LVCSR. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 8614–8618. IEEE, may 2013. ISBN 978-1-4799-0356-6. doi: 10.1109/ICASSP.2013.6639347. URL `http://ieeexplore.ieee.org/document/6639347/`.

Schmidhuber,J. Who invented backpropagation? *More in [DL2]*, 2014.

Schmidhuber,J. Deep learning in neural networks: An overview. *Neural networks*, 61: 85–117, 2015.

Schröder,S., Drescher,W., Steinhage,V., and Kastenholz,B. An automated method for the identification of bee species (hymenoptera: Apoidea). In *Proc. Intern. Symp. on Conserving Europe's Bees. Int. Bee Research Ass. & Linnean Society, London*, pages 6–7, 1995.

Sharif Razavian,A., Azizpour,H., Sullivan,J., and Carlsson,S. CNN features off-the-shelf: an astounding baseline for recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 806–813. cv-foundation.org, 2014.

Simonyan,K. and Zisserman,A. Very deep convolutional networks for Large-Scale image recognition. September 2014.

Smith,L. N. No more pesky learning rate guessing games. *CoRR*, abs/1506.01186, 2015. URL `http://arxiv.org/abs/1506.01186`.

Smith,L. N. A disciplined approach to neural network hyper-parameters: Part 1–learning rate, batch size, momentum, and weight decay. *arXiv preprint arXiv:1803.09820*, 2018.

Smith,L. N. and Topin,N. Super-convergence: Very fast training of residual networks using large learning rates. *CoRR*, abs/1708.07120, 2017. URL `http://arxiv.org/abs/1708.07120`.

Srivastava,N., Hinton,G., Krizhevsky,A., Sutskever,I., and Salakhutdinov,R. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958, 2014.

Stallkamp,J., Schlipsing,M., Salmen,J., and Igel,C. The german traffic sign recognition benchmark: a multi-class classification competition. In *The 2011 international joint conference on neural networks*, pages 1453–1460. IEEE, 2011.

Steinhage,V., Schröder,S., Cremers,A. B., and Lampe,K.-H. Automated extraction and analysis of morphological features for species identification. In *Automated taxon identification in systematics: Theory, approaches and applications*, pages 115–129. CRC Press, 2007.

Stork,N. E., McBroom,J., Gely,C., and Hamilton,A. J. New approaches narrow global species estimates for beetles, insects, and terrestrial arthropods. *Proceedings of the National Academy of Sciences*, 112(24):7519–7523, 2015.

Sun,Y., Liu,Y., Wang,G., and Zhang,H. Deep learning for plant identification in natural environment. *Comput. Intell. Neurosci.*, 2017:7361042, May 2017.

Sutskever,I., Vinyals,O., and Le,Q. V. Sequence to Sequence Learning with Neural Networks. In *Advances in Neural Information Processing Systems*, pages 3104–3112, 2014.

Swanson,A., Kosmala,M., Lintott,C., Simpson,R., Smith,A., and Packer,C. Snapshot serengeti, high-frequency annotated camera trap images of 40 mammalian species in an african savanna. *Scientific data*, 2(1):1–14, 2015.

Szegedy,C., Vanhoucke,V., Ioffe,S., Shlens,J., and Wojna,Z. Rethinking the inception architecture for computer vision. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

Taigman,Y., Yang,M., Ranzato,M., and Wolf,L. DeepFace: Closing the Gap to Human-Level Performance in Face Verification. In *Taigman, Y., Yang, M., Ranzato, M. & Wolf, L. Deepface: closing the gap to human-level performance in face verification. In Proc. Conference on Computer Vision and Pattern Recognition 1701–1708*, pages 1701–1708, 2014.

Takeki,A., Trinh,T. T., Yoshihashi,R., Kawakami,R., Iida,M., and Naemura,T. Combining deep features for object detection at various scales: finding small birds in landscape images. *IPSJ transactions on computer vision and applications*, 8(1):1–7, 2016.

Tofilski,A. Drawwing, a program for numerical description of insect wings. *J. Insect Sci.*, 4:17, May 2004.

Tofilski,A. Automatic measurement of honeybee wings. In *Automated Taxon Identification in Systematics: Theory, Approaches and Applications*, pages 277–288. CRC Press, 2007.

Tompson,J., Goroshin,R., Jain,A., LeCun,Y., and Bregler,C. Efficient Object Localization Using Convolutional Networks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 648–656, 2015.

Van Horn,G., Mac Aodha,O., Song,Y., Cui,Y., Sun,C., Shepard,A., Adam,H., Perona,P., and Belongie,S. The inaturalist species classification and detection dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

Vapnik,V. N. The nature of statistical learning theory. Statistics for Engineering and Information Science. *Springer-Verlag, New York*, 2000.

Wah,C., Branson,S., Welinder,P., Perona,P., and Belongie,S. The caltech-ucsd birds-200-2011 dataset. 2011.

Walter,D. E. and Winterton,S. Keys and the crisis in taxonomy: extinction or reinvention? *Annu. Rev. Entomol.*, 52:193–208, 2007.

Watson,A. T., O'Neill,M. A., and Kitching,I. J. Automated identification of live moths (macrolepidoptera) using digital automated identification system (daisy). *System. Biodivers.*, 1(3):287–300, 2003.

Weeks,P. J. D., Gauld,I. D., Gaston,K. J., and O'Neill,M. A. Automating the identification of insects: a new solution to an old problem. *Bull. Entomol. Res.*, 87(02):203–211, 1997.

Weeks,P. J. D., O'Neill,M. A., Gaston,K. J., and Gauld,I. D. Species–identification of wasps using principal component associative memories. *Image Vis. Comput.*, 17(12): 861–866, October 1999a.

Weeks,P. J. D., O'Neill,M. A., Gaston,K. J., and Gauld,I. D. Automating insect identification: exploring the limitations of a prototype system. *J. Appl. Entomol.*, 123(1):1–8, January 1999b.

Wei,X.-S., Luo,J.-H., Wu,J., and Zhou,Z.-H. Selective convolutional descriptor aggregation for Fine-Grained image retrieval. *IEEE Trans. Image Process.*, 26(6):2868–2881, June 2017.

Werbos,P. J. Applications of advances in nonlinear sensitivity analysis. In *System modeling and optimization*, pages 762–770. Springer, 1982.

Wilf,P., Zhang,S., Chikkerur,S., Little,S. A., Wing,S. L., and Serre,T. Computer vision cracks the leaf code. *Proc. Natl. Acad. Sci. U. S. A.*, 113(12):3305–3310, March 2016.

Willi,M., Pitman,R. T., Cardoso,A. W., Locke,C., Swanson,A., Boyer,A., Veldthuis,M., and Fortson,L. Identifying animal species in camera trap images using deep learning and citizen science. *Methods in Ecology and Evolution*, 10(1):80–91, 2019.

Wu,X., Zhan,C., Lai,Y., Cheng,M., and Yang,J. Ip102: A large-scale benchmark dataset for insect pest recognition. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8779–8788, 2019. doi: 10.1109/CVPR.2019.00899.

Xie,S., Girshick,R., Dollár,P., Tu,Z., and He,K. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1492–1500, 2017.

Yosinski,J., Clune,J., Bengio,Y., and Lipson,H. How transferable are features in deep neural networks? In *Advances in neural information processing systems*, pages 3320–3328, 2014.

Zeiler,M. D. and Fergus,R. Visualizing and Understanding Convolutional Networks. *Computer Vision – ECCV 2014*, 8689, 2014. doi: 10.1007/978-3-319-10590-1. URL `http://link.springer.com/10.1007/978-3-319-10590-1`.

Zhang,H., Cisse,M., Dauphin,Y. N., and Lopez-Paz,D. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017.

Zhang,J. and Zong,C. Deep Neural Networks in Machine Translation: An Overview. *IEEE Intelligent Systems*, 30(5):16–25, sep 2015. doi: 10.1109/MIS.2015.69. URL `http://ieeexplore.ieee.org/document/7243232/`.

Zhang,Z.-Q. *Animal biodiversity: An outline of higher-level classification and survey of taxonomic richness*, volume 3148. Magnolia Press, 2011.

Zheng,L., Zhao,Y., Wang,S., Wang,J., and Tian,Q. Good practice in CNN feature transfer. April 2016.

Zhou,B., Khosla,A., Lapedriza,A., Oliva,A., and Torralba,A. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2921–2929, 2016.