U.PORTO
FACULDADE DE CIÊNCIAS
UNIVERSIDADE DO PORTO

MSc

2.º
CICLO

FCUP
2020

U.PORTO

Deep learning to automate the assessment of cultural ecosystem services from social media data

Ana Sofia Cabral Cardoso

FC

U.PORTO
FACULDADE DE CIÊNCIAS
UNIVERSIDADE DO PORTO

Deep learning to automate the assessment of cultural ecosystem services from social media data

Ana Sofia Cabral Cardoso

Dissertação de Mestrado apresentada à
Faculdade de Ciências da Universidade do Porto em
Bioinformática e Biologia Computacional
2020

Deep learning to automate the assessment of cultural ecosystem services from social media data

Ana Sofia Cabral Cardoso

Mestrado em Bioinformática e Biologia Computacional

Departamento de Biologia | Departamento de Ciência de Computadores

2020

Orientador

Ana Sofia Vaz, Postdoctoral Researcher, Andalusian Inter-university Institute for Earth System Research (iEcolab), University of Granada, Spain & Research Centre in Biodiversity and Genetic Resources, Associate Laboratory, University of Porto, Portugal

Coorientador

Francesco Renna, Postdoctoral Researcher, Instituto de Telecomunicações, Departamento de Ciência de Computadores da Universidade do Porto

U. PORTO

FACULDADE DE CIÊNCIAS
UNIVERSIDADE DO PORTO

Todas as correções determinadas pelo júri, e só essas, foram efetuadas.
O Presidente do Júri,

Porto, ______/______/_________

Abstract

Ecosystem services (ES) are the contributions that nature provides to human well-being, either through the provision of material outputs (such as food and timber) or the regulation and maintenance of ecological processes and functions (e.g. disease protection or climate regulation). Besides provisioning and regulation ES, nature also provides cultural ecosystem services (CES), frequently known as nature-based experiences and preferences, that result from the interactions between humans and nature, contributing to people's physical and mental well-being (e.g., through inspiration or recreation). CES have been mostly evaluated through revealed and stated preference methods (e.g. social surveys). Nevertheless, with the rise of digital data and technological advances during the last years, CES analysis has become popular in social media analytics.

Most of the contemporary social media content analyses considered in the context of CES are based on the manual classification of photographs or texts shared by social media users. Inevitably, the manual classification of big photographic data is too time consuming and costly, particularly when it comes to study large geographic areas, time periods and audiences. In this context, advances in automated techniques for image classification have been showing great relevance to address CES. Among these are the convolutional neural networks (CNNs), which constitute the current deep learning state-of-the-art method for visual imagery analyses, due to their ability to capture nonlinear patterns, avoiding the necessity of manually extracting features from the images, as they are automatically learned by the algorithm. Despite deep learning advances and opportunities, the application of CNNs to advance CES assessments has been underexplored.

This thesis aims to advance an automated classification of social media photographs (more precisely, from Flickr and Wikiloc platforms) of the "Peneda-Gerês" protected area (Northern Portugal), that can be useful for CES evaluation, as well as for offering innovative solutions to the scientific community. To achieve that, two CNNs architectures where implemented – VGG16 and ResNet152 –, in conjunction with three approaches: two based in two transfer learning scenarios, one using the Places365 weights and other using the ImageNet weights, and one based in the weights obtained by training only over our dataset. The transferability and generalization of the models was also tested using Flickr's photographs from the protected area of "Sierra Nevada" (Southern Spain).

The models implemented with both of the transfer learning scenarios were more accurate than the ones with the weights obtained by training only over our dataset, suggesting that transfer learning constitutes a reliable solution for training when a small dataset is under analysis. Also, out of the two network architectures, ResNet152 achieved a slightly better performance than VGG16. The same was verified for the ImageNet and Places365, where ImageNet led to a finer model's performance, probably since the training was conducted with a greater number of images. The transferability and generalization capacity of the models when in contact with new and unseen data was not as accurate as expected, which can be related to the features/elements of the photographs which are very distinct among the datasets. Specifically, in "Sierra Nevada", cold and neutral colours, such as white, grey and blue, predominate, while in "Peneda-Gerês", warm and cold colours, like green, blue and brown, are the most common. Overall, the results revealed that deep learning methods can offer significant contributions to assist in CES evaluation.

Keywords: bioinformatics, convolutional neural networks, digital conservation, iEcology, nature-based experiences, transfer learning.

Resumo

Os serviços dos ecossistemas (SE) são as contribuições que a natureza proporciona ao bem-estar humano, seja por meio do fornecimento de produtos materiais (tais como alimentos e madeira) ou pela regulação e manutenção de processos e funções ecológicas (por exemplo, proteção contra doenças ou regulação do clima). Além dos SE de provisão e regulação, a natureza também proporciona serviços dos ecossistemas culturais (SEC), frequentemente reconhecidos como as experiências e preferências baseadas na natureza, que resultam das interações entre o homem e os ecossistemas, contribuindo para o bem-estar físico e mental das pessoas (e.g. através de inspiração ou atividades de recreio). Os SEC têm sido maioritariamente avaliados através de métodos de preferência revelada e declarada (tais como pesquisas sociais). Todavia, com a ascensão dos dados digitais e avanços tecnológicos que se verificou nos últimos anos, o estudo dos SEC através de análises de dados de redes sociais tem vindo a tornar-se popular.

A maioria das análises contemporâneas de conteúdo de redes sociais considerada no contexto dos SEC baseia-se na classificação manual de fotografias ou textos partilhados pelos utilizadores das redes sociais. Inevitavelmente, a classificação manual de grande conteúdo fotográfico constitui um processo demorado e dispendioso, especialmente quando se trata do estudo de grandes áreas, longos períodos temporais e públicos de larga escala. Neste contexto, os avanços em técnicas automatizadas de classificação de imagens têm demonstrado ser de extrema relevância na abordagem aos SEC. Entre estas encontram-se as redes neuronais convolucionais (RNCs), as quais constituem o atual estado da arte, em aprendizagem profunda, para análises de imagens, devido à sua capacidade de captar padrões visuais não lineares. Deste modo, evita-se a necessidade de extrair manualmente as características das imagens, uma vez que estas são automaticamente aprendidas pelo algoritmo. Apesar dos avanços e oportunidades na área de aprendizagem profunda, a aplicação de RNCs para promover avaliações de SEC tem sido pouco explorada.

Esta tese visa desenvolver uma classificação automatizada de fotografias de redes sociais (mais precisamente, das plataformas Flickr e Wikiloc) da área protegida, Peneda-Gerês (norte de Portugal), que se espera útil para a avaliação dos SEC, bem como para providenciar soluções inovadores à comunidade científica. Para tal, implementou-se duas arquiteturas de RNCs – VGG16 e ResNet152 –, em conjunto com três abordagens: duas baseadas em cenários de aprendizagem por transferência (um com os pesos do Places365 e outro com os pesos do ImageNet) e uma baseada nos pesos obtidos ao treinar apenas sobre o nosso conjunto de dados. A transferibilidade e generalização dos modelos também foram testadas utilizando fotografias do Flickr da área protegida Sierra Nevada (sul de Espanha).

Os modelos implementados com ambos os cenários de aprendizagem por transferência obtiveram resultados mais precisos do que os implementados com os pesos obtidos ao treinar apenas sobre o nosso conjunto de dados. Isto sugere que a aprendizagem por transferência constitui uma solução viável para treinar quando um pequeno conjunto de dados está sob análise. Além disso, das duas arquiteturas de rede, a ResNet152 alcançou um desempenho ligeiramente melhor do que a VGG16. O mesmo verificou-se para a ImageNet e Places365, onde a ImageNet conduziu a um melhor desempenho do modelo, provavelmente devido ao facto do treino ter sido realizado com um maior número de imagens. A transferibilidade e capacidade de generalização dos modelos quando em contacto com dados novos, não foram tão precisas quanto seria esperado, o que pode estar relacionado com as características/elementos das fotografias que são muito distintas entre os conjuntos de dados. Especificamente, em Sierra Nevada, cores frias e neutras, como o branco, cinza e azul, predominam, enquanto que na Peneda-Gerês, cores quentes e frias, tais como o verde, azul e castanho, são as mais comuns. No geral, os resultados revelaram que os métodos de aprendizagem profunda podem oferecer contribuições significativas para auxiliar na avaliação dos SEC.

Palavras-chave: bioinformática, redes neuronais convolucionais, conservação digital, iEcologia, experiências baseadas na natureza, aprendizagem por transferência.

Agradecimentos

O desenvolvimento desta dissertação contou com o apoio e suporte de várias pessoas, com as quais convivi e troquei experiências e conhecimentos ao longo de todo o meu percurso pessoal e académico.

Primeiramente, gostaria de agradecer à Doutora Ana Sofia Vaz e ao Doutor Francesco, pela sua dedicação, orientação, empenho e apoio enquanto orientadora e coorientador, foram deveras inalcançáveis. Sem ambos a realização deste projeto não teria sido possível.

Agradeço igualmente às instituições que me acolheram, CIBIO-InBIO (Centro de Investigação em Biodiversidade e Recursos Genéticos, Instituto de telecomunicações (entidade financiada pela FCT) e FCUP (Faculdade de Ciências da Universidade do Porto), pela confiança e oportunidade, bem como por terem providenciado todas as condições necessárias à condução e execução desta dissertação. I also thank Ricardo Moreno-Llorca, Andrea Ros-Candeira and Domingo Alcaraz-Segura, from IISTA-CEAMA (University of Granada, Spain), for their work on the dataset of “Sierra Nevada”. (i.e., data mining and image classification).

Por fim, um enorme agradecimento à minha Mãe, Irmãos, Amigos e Namorado, por toda a confiança depositada em mim, bem como por todo o apoio e incentivo que me proporcionaram ao longo de todo o meu percurso académico e, em particular, durante a realização desta dissertação.

Contents

Abstract i

Resumo iii

Agradecimentos v

List of Tables viii

List of Figures ix

1. Introduction 1

2. Thesis motivation and research goals 5

3. Background 6

3.1. Brief overview on cultural ecosystem services 6

3.2. Humans, ecosystems and social media 7

3.3. The rise of iEcology and digital conservation 8

3.4. Artificial intelligence: computer vision and convolutional neural networks 9

3.4.1. Supervised learning 10

3.4.2. Artificial neural networks 10

3.4.3. Convolutional neural networks 11

3.4.4. Training 14

3.4.5. Underfitting and Overfitting 15

4. State of the art 17

4.1. Assessing cultural ecosystem services 17

4.2. Machine and deep learning for cultural services 18

5. Implementation details 22

5.1. Test areas 22

5.2. Mining data from social media 24

5.3. Data pre-processing 29

5.3.1. Peneda-Gerês dataset 29

5.3.2. Sierra Nevada dataset 30

5.4. Automated Image Classification Methodology 31

5.4.1. Convolutional neural networks architectures 32

5.4.2. Transfer learning ... 34
5.4.3. Training ... 35
5.4.4. Data augmentation ... 38
5.4.5. Performance metrics ... 38
5.4.6. Transferability and generalization ... 39

6. Results ... 41

6.1. Nature vs. Human classification ... 41
6.1.1. Performance without data augmentation ... 41
6.1.2. Performance with data augmentation ... 44

6.2. Multilabel classification ... 46
6.2.1. Performance without data augmentation ... 46
6.2.2. Performance with data augmentation ... 49

6.3. Analysis of the best performing architectures ... 52
6.3.1. Nature vs. Human classification ... 52
6.3.2. Multilabel classification ... 55

6.4. Transferability and generalization ... 58
6.4.1. Nature vs. Human classification ... 59
6.4.2. Multilabel classification ... 62

7. Discussion and Conclusions ... 66

7.1. Nature vs. Human and multilabel classification ... 66

7.2. Transferability and generalization ... 67

7.3. Limitations and future work ... 67

Bibliography ... 69

Appendix ... 78

List of Tables

5.1. Classes considered to classify each social media photograph according to its main focus. 29

5.2. Example of a confusion matrix used to compare the manual and automatic classification of the social media photographs into the “Nature” and “Human” labels. 39

6.1 Performance metrics of the two transfer learning scenarios and the weights trained only with our dataset for the “Nature” vs. “Human” classification. ACC – Accuracy, TNR – True negative rate/Sensitivity, TPR – True Positive Rate/Specificity, F1 – f1-score. 52

6.2. Confusion matrix for ResNet152 with Places365 weights. 0 – Nature, Human – 1. 53

6.3. Confusion matrix for ResNet152 with ImageNet weights. 0 – Nature, Human – 1.......... 54

6.4. Confusion matrix for ResNet152 with the weights trained only with our dataset. 0 – Nature, Human – 1. 54

6.5. Performance metrics of the two transfer learning scenarios and the weights trained only with our dataset for the multilabel classification. ACC – Accuracy, TNR – True negative rate/Sensitivity, TPR – True Positive Rate/Specificity, F1 – f1-score. 55

6.6. Confusion matrix for ResNet152 with Places365 weights. 0 – Species, 1 – Landscape, 2 – Nature, 3 – Human activities, 4 – Human structures, 5 – Posing. 56

6.7. Confusion matrix for ResNet152 with ImageNet weights. 0 – Species, 1 – Landscape, 2 – Nature, 3 – Human activities, 4 – Human structures, 5 – Posing. 57

6.8. Confusion matrix for ResNet152 with ImageNet weights. 0 – Species, 1 – Landscape, 2 – Nature, 3 – Human activities, 4 – Human structures, 5 – Posing. 64

List of Figures

1.1. Illustration of the relationship between ecosystem services and human well-being (source: MEA, 2005). 2

3.1. Representation of the iEcology conceptual framework (source: Jarić et al. 2020)…………9

3.2. Architecture of a feedforward neural network.. 12

3.3. Linear convolution process with a mean filter mask. Figure extracted from Jeong et al. (2011) 13

5.1. Representation of one of the study areas, “Peneda-Gerês” (on the left), and its location in southwestern Europe (on the right). The figure on the left also illustrates the position of the main roads (brown) and rivers (blue). Figure extracted from Vaz et al. (2019)………………………………………………………………………………………23

5.2. Representation of one of the study areas, “Sierra Nevada” (at the bottom), and its location in the Iberian Peninsula (on the left) and at the Andalusia region (southern Spain). The figure on the bottom also illustrates the main land use and land cover (LULC) types. Figure extracted from Moreno-Llorca et al. (2020a).. 23

5.3. Examples of images belonging to the Outdoor (a) and Indoor (b) classes...................... 24

5.4. Examples of images belonging to the Nature (a) and Human (b) classes………………25

5.5. Examples of images belonging to the classes of: a) Species, b) Landscape, c) Nature, d) Human activities (the picture was edited in order to protect the identity of the persons posing in the picture), e) Human structures, f) Posing (the picture was edited in order to protect the identity of the person posing in the picture). 26

5.6. Examples of images belonging to the Outdoor (a) and Indoor (b) classes...................... 27

5.7. Examples of images belonging to the Nature (a) and Human (b) classes. 27

5.8. Examples of images belonging to the classes of: a) Species, b) Landscape, c) Nature, d) Human activities, e) Human structures, f) Posing (the picture was edited in order to protect the identity of the persons posing in the picture)…………………………………………...28

5.9. Distribution of the number of outdoor photographs for each class in the “Peneda-Gerês” dataset... 30

5.10. Distribution of the number of outdoor photographs for each class in the “Peneda-Gerês” dataset... 31

5.11. Workflow of the image classification methodology.. 32

5.12. Configuration of the VGG16 model architecture. Retrieved from Nash et al., 2018. 33

5.13. Configuration of a residual network with 34 parameter layers (3.6 billion FLOPs). Retrieved from He et al., 2016. 34

5.14. Behaviour of ResNet152 (a) and VGG16 (b) with Places365 weights, in terms of accuracy, validation accuracy, loss and validation loss. 36

5.15. Behaviour of ResNet152 (a) and VGG16 (b) with ImageNet weights, in terms of accuracy, validation accuracy, loss and validation loss. 37

6.1. Accuracy of the VGG16 and ResNet152 model performance for the two transfer learning scenarios and the weights obtained by training only over our dataset. The bars reported in the plot represent standard deviations……………………………………………………..42

6.2. Sensitivity of the VGG16 and ResNet152 model performance for the two transfer learning scenarios and the weights obtained by training only over our dataset. The bars reported in the plot represent standard deviations.. 42

6.3. Specificity of the VGG16 and ResNet152 model performance for the two transfer learning scenarios and the weights obtained by training only over our dataset. The bars reported in the plot represent standard deviations.. 43

6.4. F1-score of the VGG16 and ResNet152 model performance for the two transfer learning scenarios and the weights obtained by training only over our dataset. The bars reported in the plot represent standard deviations.. 43

6.5. Accuracy of the VGG16 and ResNet152 model performance for the two transfer learning scenarios and the weights obtained by training only over our dataset. The bars reported in the plot represent standard deviations.. 44

6.6. Sensitivity of the VGG16 and ResNet152 model performance for the two transfer learning scenarios and the weights obtained by training only over our dataset. The bars reported in the plot represent standard deviations.. 45

6.7. Specificity of the VGG16 and ResNet152 model performance for the two transfer learning scenarios and the weights obtained by training only over our dataset. The bars reported in the plot represent standard deviations.. 45

6.8. F1-score of the VGG16 and ResNet152 model performance for the two transfer learning scenarios and the weights obtained by training only over our dataset. The bars reported in the plot represent standard deviations.. 46

6.9. Accuracy of the VGG16 and ResNet152 model performance for the two transfer learning scenarios and the weights obtained by training only over our dataset. The bars reported in the plot represent standard deviations. 47

6.10. Sensitivity of the VGG16 and ResNet152 model performance for the two transfer learning scenarios and the weights obtained by training only over our dataset. The bars reported in the plot represent standard deviations. 47

6.11. Specificity of the VGG16 and ResNet152 model performance for the two transfer learning scenarios and the weights obtained by training only over our dataset. The bars reported in the plot represent standard deviations. 48

6.12. F1-score of the VGG16 and ResNet152 model performance for the two transfer learning scenarios and the weights obtained by training only over our dataset. The bars reported in the plot represent standard deviations. 49

6.13. Accuracy of the VGG16 and ResNet152 model performance for the two transfer learning scenarios and the weights obtained by training only over our dataset. The bars reported in the plot represent standard deviations. 49

6.14. Sensitivity of the VGG16 and ResNet152 model performance for the two transfer learning scenarios and the weights obtained by training only over our dataset. The bars reported in the plot represent standard deviations. 50

6.15. Specificity of the VGG16 and ResNet152 model performance for the two transfer learning scenarios and the weights obtained by training only over our dataset. The bars reported in the plot represent standard deviations 51

6.16. F1-score of the VGG16 and ResNet152 model performance for two transfer learning scenarios and the weights obtained by training only over our dataset. The bars reported in the plot represent standard deviations. 51

6.17. Examples of photographs where the labels were swapped by ResNet152 with Places365 weights. 0 – Nature, 1 –Human. 53

6.18. Examples of photographs where the labels were swapped by ResNet152 with ImageNet weights. 0 – Nature, 1 –Human. 54

6.19. Examples of photographs where the labels were swapped by ResNet152 with the weights obtained by training only over our dataset. 0 – Nature, 1 – Human. 55

6.20. Examples of photographs with labels for which the algorithm most failed the classification in ResNet152 with Places365 weights. 0 – Species, 1 – Landscape, 2 – Nature, 3 – Human activities, 4 – Human structures, 5 – Posing 57

6.21. Examples of photographs with labels for which the algorithm most failed the classification in ResNet152 with ImageNet weights. 0 – Species, 1 – Landscape, 2 – Nature, 3 – Human activities, 4 – Human structures, 5 – Posing. 58

6.22. Accuracy of the ResNet152 model performance for the two transfer learning scenarios and the weights obtained by training only over our dataset. 59

6.23. Sensitivity of the ResNet152 model performance for the two transfer learning scenarios and the weights obtained by training only over our dataset. 60

6.24. Specificity of the ResNet152 model performance for the two transfer learning scenarios and the weights obtained by training only over our dataset. 60

6.25. F1-score of the ResNet152 model performance for the two transfer learning scenarios and the weights obtained by training only over our dataset. 61

6.26. Examples of photographs where the labels were swapped by the models. 0 – Nature, 1 – Human, a), b) – ResNet152 with Places365 weights, c), d) – ResNet152 with ImageNet weights, e), f) – ResNet152 with weights trained only with our dataset (the e) picture was edited in order to protect the identity of the persons posing in the picture). 62

6.27. Accuracy of the ResNet152 model performance for the two transfer learning scenarios and the weights obtained by training only over our dataset. 62

6.28. Sensitivity of the ResNet152 model performance for the two transfer learning scenarios and the weights obtained by training only over our dataset. 63

6.29. Specificity of the ResNet152 model performance for the two transfer learning scenarios and the weights obtained by training only over our dataset. 63

6.30. F1-score of the ResNet152 model performance for the two transfer learning scenarios and the weights obtained by training only over our dataset. 64

6.31. Examples of photographs with labels for which the algorithm most failed the classification in ResNet152 with ImageNet weights. 0 – Species, 1 – Landscape, 2 – Nature, 3 – Human activities, 4 – Human structures, 5 – Posing (the picture was edited in order to protect the identity of the person posing in the picture). 65

List of abbreviations

ES	Ecosystem services
CES	Cultural ecosystem services
AI	Artificial intelligence
CNNs	Convolutional neural networks
ANNs	Artificial neural networks
MLPs	Multilayer perceptrons
CCE	Categorical cross-entropy
MSE	Mean squared error
SGD	Stochastic gradient descent
VGG	Very deep convolutional network
ResNet	Residual neural network
SPAs	Special protected areas
SIC	Site of community importance
API	Application programming interface
ReLU	Rectified linear unit
ACC	Accuracy
TPR	Sensitivity/True Positive rate
TNR	Specificity/True Negative rate
$\mathbf{F_1}$	F1-score

Chapter 1

Introduction

Ecosystem services (ES) represent the benefits that people can obtain from nature, which are currently subdivided into three main categories: provisioning, regulating and cultural services (MEA, 2005; Figure 1.1). ES result from the outputs that are supplied by ecosystems and from the benefits that are demanded by people to achieve a good quality of life. Provisioning services include material benefits directly obtained from nature, such as food, water and energy, while regulating services comprise the results from regulatory processes of ecosystems that contribute to climate regulation, pest control and water quality, among others. Cultural ecosystem services (CES), include nature-based experiences and preferences and constitute the non-material benefits that people can experience from nature, such as recreation and ecotourism, as well as those pertaining to spiritual, religious, aesthetic or heritage values, among others (MEA, 2005; Hausmann *et al*., 2018). In recent years, out of the three categories of ES, the study of CES has become extremely relevant, for instance encouraging people to engage in activities related to the conservation and maintenance of the environment and ecosystems, or the improvement of areas associated with urban planning and landscape design (Richards & Tunçer, 2018; Fish *et al*., 2016a). However, the study of CES has been particularly complex and challenging in the scope of decision-making, comparatively to the other ES, due to its intangible and often subjective nature that results from intellectual (e.g. aesthetics) or physical (e.g. recreation) interactions between humans and the environment (Cheng *et al*., 2019).

CES have been traditionally evaluated through revealed and stated preference methods from the social sciences (e.g. social surveys), which can be restricted in terms of temporal and spatial coverage, especially if a large study area is under consideration (Yoshimura & Hiura, 2017). Due to fast improvements in computational power and data storage capacity during the last years, the emergent fields of Digital Conservation (van der Wal & Arts, 2015), iEcology (Jarić *et al*., 2020) and conservation culturomics (Ladle *et al*., 2016) have brought new opportunities to address CES. These disciplinary fields refer to the use of digital (big) data and technology to understand human-nature interactions and to provide evidence in favour of nature conservation and of the sustainable use and management of ecosystems (van der Wal & Arts, 2015; Toivonen *et al*., 2019). It considers the use of digital information and data produced and shared by people, for instance through their mobile devices (e.g. smartphones, digital cameras, etc.) and in social media platforms (e.g. geo-tagged photographs and tweets). The incorporation of social media data in the study of conservation-related events, such as CES, has brought, for instance, the possibility of assessing the presence of people in protected areas, as well as to identify human-nature interactions hotspots, which can add a significant contribution to the understanding of landscape values, human activities in nature, visitation preferences of people, among others. (Tenkanen *et al*., 2017).

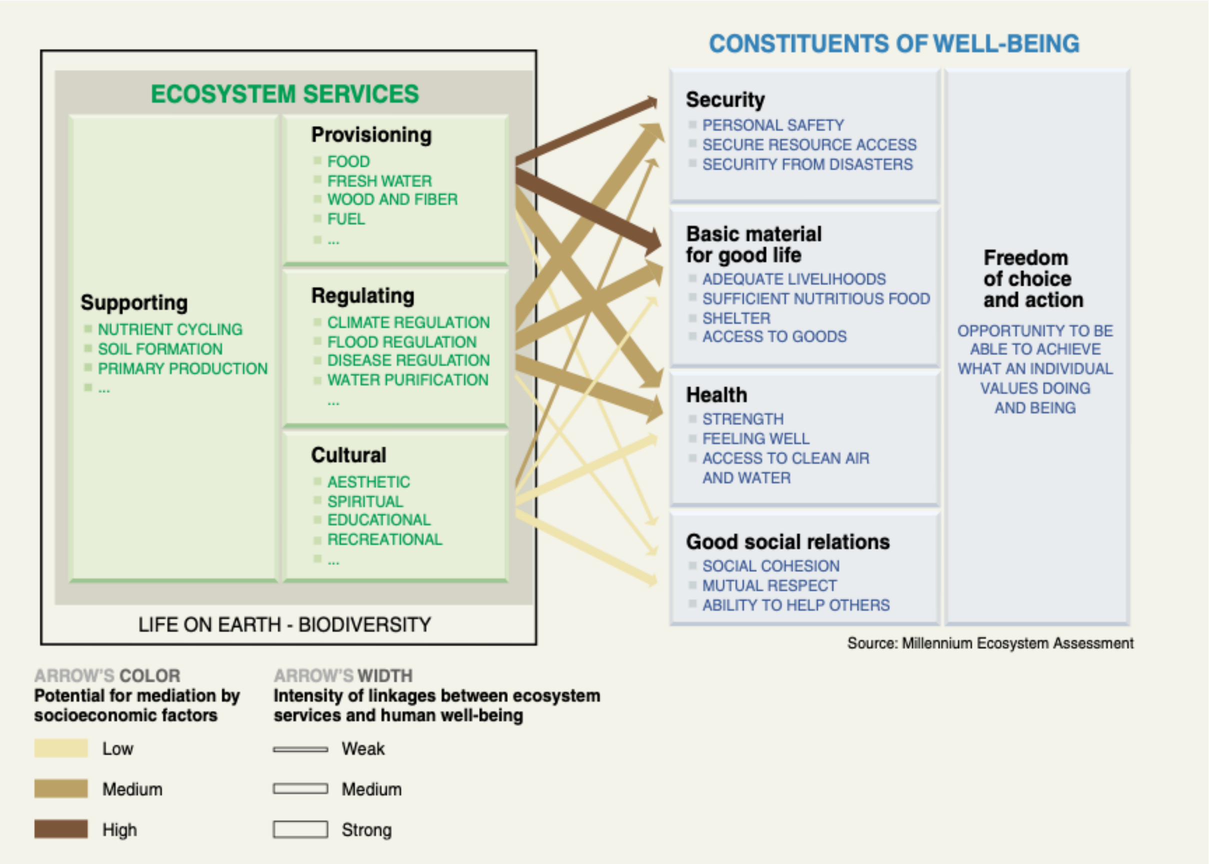


Figure 1.1. Illustration of the relationship between ecosystem services and human well-being (source: MEA, 2005).

Nowadays, computer science and related fields have been highly invested in the use and combination of methods that incorporate social media analytics (Sherren *et al*., 2017). Social media platforms represent a very significant fraction of all social digital data, constituting an efficient method to collect big data that provide information on people's interactions with each other and with their environment (Di Minin *et al.*, 2015). For instance, Flickr (often used for sharing nature-related content) and Instagram (normally associated with day-to-day content) are two media-sharing platforms with a rich visual (e.g. photos) and textual (e.g. tags and comments) contents. Furthermore, these platforms respectively have about 1.7 and 40 million photos shared by their users at a daily basis, making them candidate tools for the identification, mapping and monitoring of physical, visual and sensory features of ecosystems and nature

(Heikinheimo *et al*., 2018; Fu & Rui, 2017). However, an approach that combines different data from social media with advanced analytics, besides spatial analysis, remains very uncommon. Thus, the investment in methods that can identify features of ecosystems and nature through the content analysis of shared photos (or text), can constitute an asset to support the evaluation of CES, particularly, related to aesthetics and recreation or ecotourism (Fu & Rui, 2017; Richards & Tunçer, 2018). Furthermore, the action of sharing photographs in social media can provide digital proxies of spatial preferences, since people usually share pictures that they assume to be worthy of sharing visually, which gives a sense of the places that are considered to be valuable for visitation (Gliozzo *et al*., 2016).

Most social media content analyses considered in the context of CES are based on the manual classification of photos or texts shared by social media users (Cheng *et al*., 2019). Examples include the work developed by Hausmann *et al*. (2018), who evaluated people's preferences for biodiversity based on the manual interpretation of Flickr and Instagram's photographs. Inevitably, the manual classification of big photographic data is too time consuming and costly, particularly when it comes to large areas under analysis. In this context, advances in automated techniques for image classification have been showing great relevance to address CES (and other ES; Willcock *et al*., 2018; Gosal *et al*., 2019).

One of the biggest challenges in the modelling and assessment of ES is deciding on either a simple approach or a complex approach. Complex approaches are time consuming and require intensive data, therefore being more challenging to implement and hard to upscale, while simple approaches allow for speedier assessments. However, complex approaches imply more accurate and locally specific results, and simple approaches sacrifice accuracy and credibility of its results (Martínez-López *et al*., 2019). The emergence of the artificial intelligence (AI) field brought a wide range of possibilities in broadest areas, with ES being no exception, since it allows to reduce assessment complexity and cost for the user, offering a simple, yet precise, approach (Villa, 2009). An AI algorithm constitutes every set of rules that is followed in order to perform a given task, allowing to perceive its environment, a singularity that is acquired by selecting actions that maximize the probability of successfully reaching particular goals (Vinuesa *et al*., 2020).

One of the most common AI subfields in computer vision analysis is machine learning, which, in general, focuses on the development of computer (or machine) programs that can access data and use that data in order to self-learn task performance. Within machine learning, deep learning has been gaining ground in the areas of digital image processing, constituting the current state-of-the-art. Deep learning is defined as the set of algorithms that are based on artificial neural network architectures (Lusch *et al*., 2018), such as Convolutional Neural Networks (CNNs), and that allow the automated identification and analysis of labelled data in the context of visual imagery classification, natural language processing (NLP), video analysis, time series forecasting, among many others (Najafabadi *et al*., 2015). Advances in deep learning have shown to be very effective in many research areas, such as species identification (Ferreira *et al*., 2019), health care (Renna *et al*., 2019) or safety surveillance (Castillo *et al*., 2019). However, those algorithms have been seldom applied to ES in general, and more specifically to CES evaluation from social media (Willcock *et al*., 2018). Currently developed tools that can be useful to support CES analysis, such as the Google Cloud Vision[1] (https://cloud.google.com/vision), are not freely available (Richards & Tunçer, 2018; Mulfari *et al*., 2016; Gosal *et al*., 2019).

In this context, the development of scientific tools grounded on AI and machine learning algorithms should constitute an asset to CES evaluations. Therefore, this will allow the incorporation of big data into interdisciplinary models to provide holistic solutions to complex socio-ecological issues under user-friendly, cost-effective and publicly accessible ways.

[1] The Google Cloud Vision is an application programming interface that provides powerful pre-trained machine learning classification models, specialized in detecting objects and faces, read printed and handwritten text, etc., by means of assigning labels to a given image.

Chapter 2

Thesis motivation and research goals

Cultural ecosystem services (CES) are an essential concept of human health and well-being, contributing, among others, to foster social cohesion, which makes the assessment and quantification of these services essential for people’s welfare when in contact with the environment and nature, supporting decision making in national parks, protected areas, among others. CES are often considered as intangible and subjective, that makes the measurement and quantification of these services difficult. However, CES evaluations can be indirectly obtained through the evaluation of people’s experiences and preferences towards nature. For instance, such preferences can be evaluated from texts and pictures shared by people on social media platforms. During recent years, CES have been mainly inferred from the manual interpretation of social media pictures, which is time and resource consuming. Thus, the investment in the development of new techniques that reduce both the evaluation time and cost to infer on CES from social media data, has become a priority in the digital conservation field.

This project aims to develop an automated classification of social media photographs that can be useful for CES evaluation and for providing innovative solutions to the scientific community. Specifically, this study aims to answer the following questions: (1) can deep learning algorithms be developed to support an automated classification of social media photographs in the context of CES? (2) how can those algorithms and models be improved in order to promote statistically reliable image classifications? and (3) at which point can those algorithms and models be replicable and applied to other geographical contexts than those for which they have been trained? To achieve this, deep learning algorithms are here developed and tested, more specifically Convolutional Neural Networks and transfer learning strategies, in digital photographs from social media platforms Flickr and Wikiloc at the “Peneda-Gerês” protected area (Northern Portugal). The spatial transferability of those algorithms is also tested using Flickr’s photos from the protected area of “Sierra Nevada” (Southern Spain).

Chapter 3

Background

3.1. Brief overview on cultural ecosystem services

In the last decade, there has been a general acceptance that ecosystem services (ES) have become an important tool for decision-making on several ecological and social challenges. ES have the capacity to reflect the benefits that people can acquire from nature, as well as to increase public awareness on environmental protection and sustainability (Plieninger *et al*., 2015). Although the outcomes of ES evaluation can sustain practical applications and policy making (e.g. urban planning or landscape design), the low availability of data constitutes a significant obstacle to the assessment of ES. When it comes to the cultural ecosystem services (CES) domain - i.e. the non-material benefits people obtain from nature - this barrier becomes more evident, for both quantitative and qualitative data, since there isn't a clear boundary between the different CES categories, which can often result in double-counting problems (e.g., benefits for recreation can be frequently confused with aesthetic values, educational values, and spiritual and religious values) (Brown *et al*., 2016).

Multiple methods have been developed to assist on CES valuation. These methods may have a monetary and non-monetary nature (Hirons *et al*., 2016), and account with revealed preference and stated preference approaches. For instance, the revealed preference in monetary methods consists in observing actual markets for CES valuation, while the stated preference consists in building and simulating a market and questioning respondents to declare their willingness to pay, receive or give up of some benefits which can emerge from CES. For non-monetary methods, in turn, the revealed preference is based in monitoring behaviour or analysing pre-existent information, such as written texts and advertisements, in order to determine, indirectly, people's preferences for CES (e.g. through questionnaires). As CES values can be particularly challenging and not always be assessed through monetary methods, interviews and questionnaires represent the methods most frequently used for CES assessment (Cheng *et al*., 2019), which are often time consuming particularly when targeting large audiences. In this context, cost-effective alternatives to traditional methods, such as social media analytics, have been increasingly rising at the interface between ecology and computer sciences (see section 3.2).

3.2. Humans, ecosystems and social media

The current information age, in which an increasing volume of data (big data) is progressively produced through user activities in virtual networks and platforms, has highly contributed to the emergence of new research avenues in various fields of science, with ES being no exception (Toivonen *et al*., 2019). With the arising of social media and digital resources, new communication opportunities have emerged, providing a rich source for studying people's activities in nature, and therefore offering new opportunities for people (via computers) to classify species in pictures, identify patches of forests in satellite images, evaluate the sentiments and emotions expressed by social media users, among others. The emergence of social media has also allowed people to express their opinions and thoughts, report on newsworthy events, debate conservation issues or simply discuss or search topics of interest, offering real-time insights into significant events, actions that were not possible with traditional media, since people actively contribute to data for research in a structured manner (Pathak *et al*., 2017; Toivonen *et al*., 2019).

Social media platforms constitute the web-based services that enable individuals, groups, communities and organizations to interact, cooperate and connect, providing resources for people to create, share and engage on freely available and easily accessible content (McCay-Peet & Quan-Haase, 2017). The information is characterized by large and variable volumes of data being stored, which makes filtering and cleaning very important tasks particularly when dealing with data that is inaccurately georeferenced or generated by bots (Di Minin *et al*., 2018). It can be divided into five main elements: (1) user information (e.g., full name, username, number of followers, home location), (2) data content (e.g. text, image, video, sound), (3) timestamp (date and time of the post), (4) geotag (automatic or user-defined location for the post), (5) reactions of the remaining users (e.g. comments and likes), making social media data similar to other types of reliable geographic information (Toivonen *et al*., 2019). In this context, the information withdrawn from social media has offered new approaches for studying visitation patterns in conservation areas, preferences and activities of protected area visitors, monitor public reactions to conservation-related events, as well as mapping CES, among others (Tenkanen *et al*., 2017; Gliozzo *et al*., 2016; Lunstrum, 2017).

These data is largely shared thanks to the progressive and increasingly frequent use of smartphones, that provide the capacity to record people's locations through the mobile network operator services and mobile applications such as social media platforms, constituting proxies for identifying changes in people's distribution, as well as for understanding the movement patterns of the users (Frank et al., 2014). In fact, over the last years, we have been witnessing the emergence of new transdisciplinary areas of research, such as conservation culturomics, digital conservation or iEcology (Jarić *et al*., 2020). In these areas, researchers take advantage of digital technology and publicly available data from large audiences to support the surveillance and management of nature's contributions to people at several scales (Jarić *et al*., 2020; Toivonen *et al*., 2019; see section 3.3).

3.3. The rise of iEcology and digital conservation

Digital conservation has recently risen as a subarea of conservation science that is dedicated to the use of innovative data sources, such as social media data, satellite Earth observations and other (big) digital data sets, in order to analyse and mitigate biodiversity issues and environmental challenges (Arts *et al*., 2015). It constitutes a set of digital interactive services that enable the creation and/or distribution of information, opinions, and other types of communication/interactions, through virtual networks and groups. Since user-generated big data can give insights about human-nature interactions, such as people's interests on nature, conservation debates or online discussions (among others), social media data may offer novel cost-efficient methods for CES monitoring and assessment.

The field of iEcology (Figure 3.1) and conservation culturomics represent an opportunity to understand human and nature interactions in the digital realm, being dedicated to the study of ecological informatics, using innovative data sources generated online by human society, being the target data not purposefully produced to address ecological and environmental challenges (Jarić *et al*., 2020; Ladle *et al*., 2016). Nowadays, the most common applications of these methods are closely related with the study of ecosystem and habitat dynamics, species occurrences, as well as their spatiotemporal trends (e.g. trait dynamics, evolutionary trends, biogeographic patterns, biotic and abiotic interactions, among others). The incorporation of methods and tools in the iEcology and conservation culturomics arenas could effectively be beneficial for studies in other ecology and environmental sciences fields. For that, technological advances, namely on artificial intelligence, offer numerous opportunities (see section 3.4).

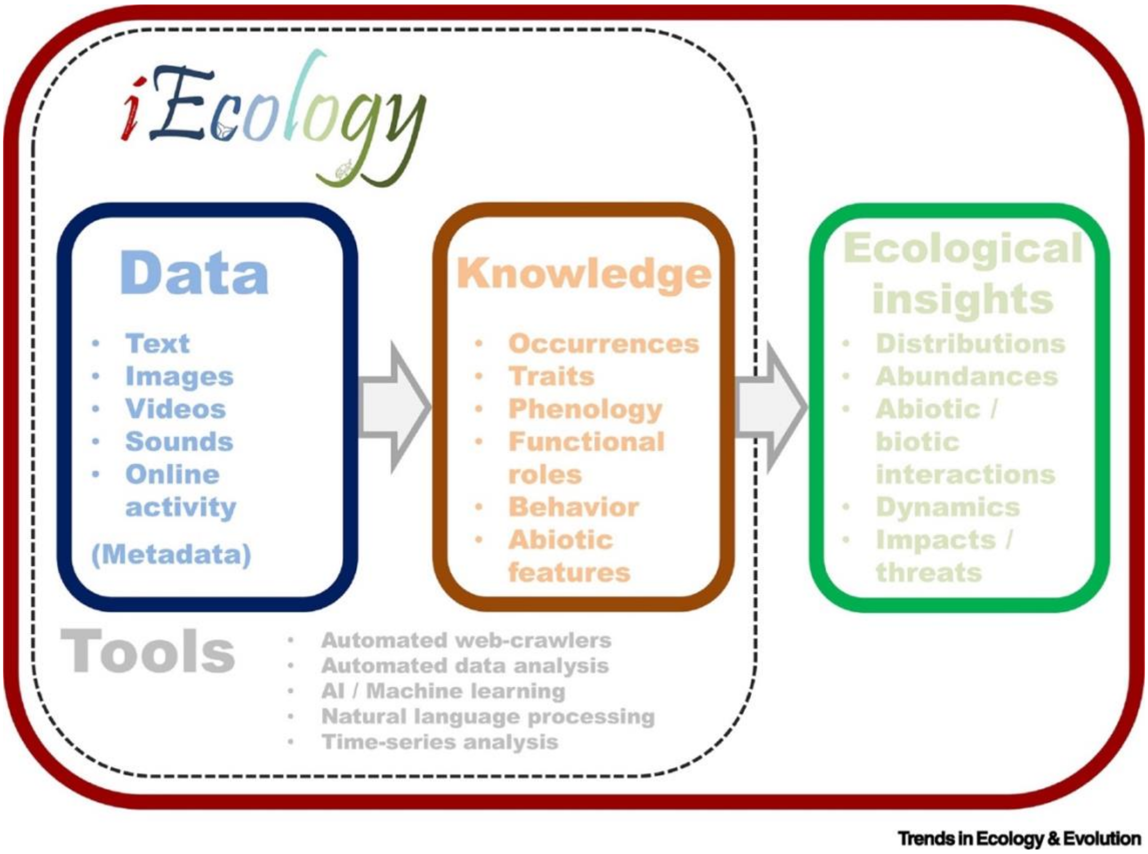


Figure 3.1. Representation of the iEcology conceptual framework (source: Jarić et al. 2020)

3.4. Artificial intelligence: computer vision and convolutional neural networks

In the computer science arena, computer vision is closely related with artificial intelligence, having first arisen with the main goal of mimicking the human visual system for endowing robots with intelligent behaviour (Szeliski, 2010). Computer vision constitutes a wide-range and interdisciplinary field that is essentially based on the study of automatic processing and understanding of digital images and videos. Computer vision main tasks consist in acquiring, processing, analysing, understanding and extracting high-dimensional data from the real world, in order to produce numerical or symbolic information that can be useful for decision-making (Klette, 2014). In other words, computer vision encompasses the process of transforming visual images (the input of the retina) into descriptions of the real world that can be used to manage a specific action (complete scene understanding). Deep learning has brought state-of-the-art methods and tools for solving challenging computer vision tasks, such as classifying the content of digital photographs, through the identification of objects and their outlines, providing descriptions of the entire image or their parts and enabling the monitoring of species and ecosystems (Rawat & Wang, 2017; Lee *et al*., 2019).

3.4.1. Supervised learning

In the generality of machine learning and deep learning algorithms, the process of learning can be subdivided into three learning paradigms: supervised, unsupervised and reinforcement. Among the three, supervised learning is one of the most common approaches in the digital image processing field.

In this learning approach, a model is trained from annotated data, based on example input-output pairs, that can be used to solve either a classification or regression problem. It has the capacity of inferring a function from a set of labelled training examples, that can be further used for mapping new and unseen data. The main goal of supervised learning consists in approximating the mapping function $f(x)$ in a way that, in the presence of new input data x, it becomes possible to predict the output variables y for that specific data (Eq. 2.1):

$$y = f(x) \qquad \text{Eq. (2.1)}$$

This process is achieved through the adjustment of the inter connection weight combinations with the aid of error signals (Sathya & Abraham, 2013). Compared to conventional methods, supervised learning, when applied to deep learning methods, is subject to scalability problems, demanding huge volumes of labelled data to generalize properly. Also, these techniques have low capacity of generalization to multiple domains and tasks (Chum *et al*., 2019).

3.4.2. Artificial neural networks

Most deep learning models used for visual understanding are based on artificial neural networks (ANNs). ANNs were initially proposed with the main goal of mimicking the performance of a biological brain. They are based on a set of connected units or nodes named artificial neurons: mathematical operations that have the capacity to transmit a signal, as well as to receive one or more inputs through connections (edges) with other artificial neurons (process that resembles the biological synapses). The neurons are normally organized into multiple layers, where the connections among the neurons only occur between the immediately preceding and following layers. The transmission of the signal occurs in the first layer (input layer) and continues to the last layer (output layer), typically after navigating through the layers, multiple times. This signal corresponds to a real number that is computed by a selected non-linear activation function (e.g., rectified linear unit (ReLU), sigmoid, softmax, among others).

Also, each neuron and connection normally have an associated weight that undergoes adjustments throughout the learning process, depending on whether it increases or decreases the total strength of the signal, and consequently, improves the accuracy of the result. Whenever the weighted sum of inputs surpasses the threshold value of a certain neuron, it stays activated and sends the signal through a transfer function that is responsible for passing it to the neighbouring neurons (Zou *et al*., 2008). Similarly, biases are also added to the neurons as an additional input into the next layer. In these models, each level learns the process of transforming input data (extracting higher level features) into a slightly more abstract, composite and complex representation (output) (LeCun *et al*., 2015; Goodfellow *et al*., 2016). When working with image data, the raw input is a matrix of pixels, where the model automatically learns which features (e.g. edges, colour) should be considered in each level.

3.4.3. Convolutional neural networks

Within deep learning techniques, CNNs constitute one of the most used algorithms in the visual imagery scope. Similarly, CNNs constitute an electronic system that is capable of learning to identify similarities between patterns of information, in a manner that closely resembles a biological brain. However, they were specially developed for processing structural data that have a sequential (1D) or grid-like structure (2D, 3D), constituting a class of deep neural networks where linear operations (matrix multiplications) among nodes are substituted by mathematical operations named convolutions. CNNs, as being a type of feedforward neural networks (Figure 3.2), have connections between the nodes that do not form a cycle or loop, where information moves in only one direction (forward), from the input nodes, through the hidden nodes (if existing) and, finally, to the output nodes. These can be interpreted as regularized versions of multilayer perceptrons (MLPs), where each neuron in n^{th} layer is connected to all the neurons in the next $(n+1)$ th layer (fully connected networks). In order to compute the output, Rosenblatt (1958) presented the weights concept, which is responsible for representing the importance of the respective inputs to the output. This type of networks has as main objective to approximate some function f^{*}. For instance, for a common classifier, $y = f^{*}(x)$ maps an input x to a category y. Feedforward networks, in its turn, describe a mapping $y = f(x;\theta)$, in order to learn the value of the parameter θ that leads to the best function approximation (Goodfellow *et al*., 2016).

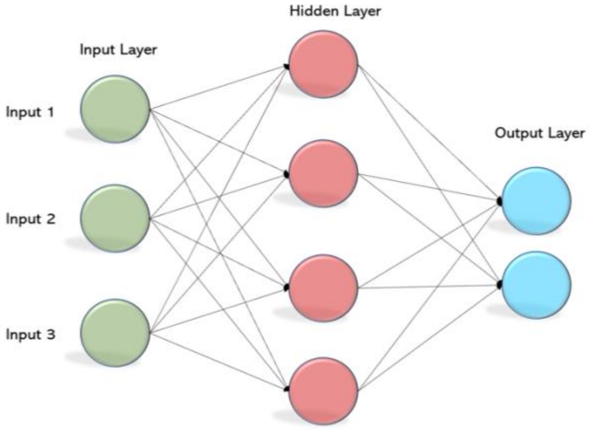


Figure 3.2. Architecture of a feedforward neural network.

Convolutional layers in CNNs imply sparse, local interactions between neurons and parameter sharing, which guarantee to provide equivariant representations of the input data (Goodfellow *et al*., 2016). The sparse interactions are accomplished by considering convolution kernels smaller than the input, which allow to learn local patterns using small but meaningful features, reduce the number of parameters to be stored and the memory requirements. Parameter sharing, in its turn, refers to the use of the same parameter for more than one function in a model (each component of the kernel is normally used at every location of the input) and equivariant representations implies that a transformation in the input will be also translated in a transformation in the output. Since a photograph can have a million pixels, all these features optimize and automate its classification, distinguishing these networks from other machine learning algorithms (such as support vector machines (SVMs) or k-nearest neighbours (k-NN)).

Besides convolutional layers, CNNs also comprise three different types of layers: pooling, fully connected and output layers. Pooling layers are responsible for reducing the dimensions of the data, through the downsampling of the feature maps, individually. This is achieved by sliding a two-dimensional kernel over each channel of the feature map, reducing the features that lie inside that specific area. Also, pooling layers can be either global (over all of the neurons) or local (over small clusters), max (using the maximum value of the cluster) or average (using the average value of the cluster). Fully connected layers, in turn, have the main function of connecting the inputs from one layer to every activation unit of the next layer. Lastly, the most used type of output layer in CNNs is the one that uses the softmax as the activation function. This function is used to normalize the output to a probability distribution over the predicted output labels. Currently, it constitutes the state-of-the-art for most CNNs models, due to its simplistic interpretation (Liu *et al*., 2016).

The use of convolutions in the neural network architecture allows avoiding the necessity of manually extracting features from the images, as these are automatically learned by the portion of the network that contains convolutional layers (Nielsen *et al*., 2017). For this reason, CNNs only demand a minimum level of pre-processing, making the use of these techniques preferred over other image classification algorithms in several application scenarios. A discrete convolution of two one-dimensional signals x and k can be described as (Eq. 2.2):

$$(x * k)(i) = \sum_{j=-\infty}^{\infty} x(j) \times k(i-j) \qquad \text{Eq. (2.2)}$$

Where * symbolizes the convolution operation, x the input, k the kernel, i the location where the convolution is calculated and $j,$ a value that is responsible for indicating which input and kernel elements will be multiplied. When considering CNNs, the convolution output is often mentioned as a feature map. In each convolutional layer, the convolution operation is applied between the input and the kernel to produce a feature map (Figure 3.3). More precisely, for every convolutional layer, multiple feature maps are generated by applying a convolution operation between the input and the kernel. The kernel slides along the input, stops at every possible position and, lastly, calculates the dot product between the input and the kernel. A convolution normally requires specifying a few parameters to be implemented, namely the kernel size, number of kernels to be applied, stride size (also referred to as increment, how much we slide at each step), among others.

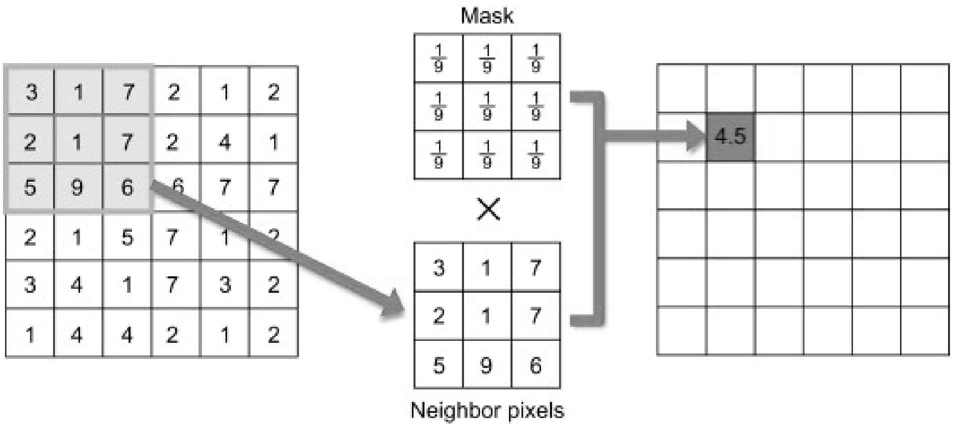


Figure 3.3. Linear convolution process with a mean filter mask. Figure extracted from Jeong *et al*. (2011)

3.4.4. Training

CNNs are usually trained using iterative, gradient-based optimizers that drive the cost function to a very low value (non-convex loss functions). Thus, training with CNNs implies the presence of a loss/cost function that allows to compute the model error and, consequently, to optimize the training process. A loss/cost function is responsible for mapping the values of one or more variables into a real number that has some "cost" associated, contributing to the knowledge of the weights and biases during the training. In an optimization problem, the goal is to find the best values that minimize the selected loss/cost function.

Categorical cross-entropy (CCE) and mean squared error (MSE) constitute the two most common loss/cost functions in the state-of-the-art, regarding to training with CNNs. MSE is preferred for regression analyses, while CCE is normally more appropriated for classification problems, where one example can be assigned to a specific category with probability 1 and to other categories with probability 0. CCE is defined as follows (Eq. 2.3):

$$CCE = -\sum_{i}^{C} t_i \log(s_i) \qquad \text{Eq. (2.3)}$$

where t_i corresponds to the ground truth vector and s_i to the predicted probability vector (softmax output). Also, normally softmax or sigmoid functions constitute the two most recommended activation functions to use in conjunction with the categorical cross-entropy cost/loss function, since both can provide the appropriate conditions for the CCE function to fulfill its main purpose of comparing two probability distributions (Kanai *et al.*, 2018).

However, gradient based learning can be very time consuming when training a significant number of inputs, since finding a global minimum (or a very low value) can be quite a difficult task, given that the loss function is not convex, in general. Accordingly, the iterative method of stochastic gradient descent (SGD) can be useful, since its performance consists in randomly picking up a small number of randomly chosen training inputs (a mini-batch gradient descent with a batch size of one). In other words, SGD operates by replacing the actual gradient (estimated from the entire dataset) by an estimate thereof (computed from a randomly selected subset of the data) (Goodfellow *et al.*, 2016), in order to calculate the gradient of the loss function CCE with respect to the parameters θ given a learning rate η and each individual training example x^i and corresponding label y^i (Eq. 2.4):

$$\theta_{t+1} = \theta_t - \eta . \nabla_\theta CCE\left(x^{(i)}; y^{(i)}; \theta_t\right) \qquad \text{Eq. (2.4)}$$

The architecture of the SGD algorithm implies the definition of a few essential hyperparameters (constant parameters whose value is established before the learning process), being the batch size and the number of epochs two of the most relevant. The batch size corresponds to the number of training samples considered in one iteration. An iteration constitutes a single gradient update (i.e. model's weights update) in the training, while the number of epochs, in turn, represent the number of times an entire dataset is passed forward and backward through the model. Lastly, the learning rate is classified as a tuning parameter that is responsible for determining the step size at each iteration, in order to adjust for errors and, consequently, converge towards a minimum of a loss function.

The algorithm responsible for applying gradient descent to the training of the majority of convolutional neural networks is called backpropagation (Rumelhart *et al*., 1986). This algorithm can be divided into two phases: a first one, frequently known as forward propagation, where the input propagates over the network to generate an output $\hat{y}$, with the algorithm keeping a stack of function calls, as well as their computed parameters, and a second one, known as back-propagation, that enables the back propagation of the information from the cost through the network. This process is only possible due to the chain rule of calculus, a formula that is used to compute the derivatives of composite functions. Backpropagation begins with the loss value and runs backwards from the final layers to the initial layers, uses the chain rule in the computations of gradient values and, lastly, calculates the contribution each parameter had to the final loss value. In feedforward networks, backpropagation can either be manifested in terms of matrix multiplication or adjoint graphs (more common) (Nielsen, 2015; Goodfellow *et al*., 2016).

3.4.5. Underfitting and Overfitting

One of the main drawbacks associated with deep learning algorithms is the requirement of high volumes of training data for models to learn and generalize well when in the presence of unseen data, a condition that normally is not verified in conservation science datasets (Toivonen *et al*., 2019). This frequently leads to overfitting, a negative condition where the model learns and memorizes the detail and noise present in the training data to the extent that it fails to reliably fit new and unseen data, as the noise or random fluctuations may vary between different data (Jabbar & Khan, 2015). The resulting model is overfitted, and therefore comprises more parameters than the ones that can be explained by the training data. The inverse, underfitting, occurs when the model is unable to capture the underlying structure of the data, which often results in low generalization and unreliable predictions. Unlike the overfitted model, the underfitted one does not contain the parameters that should be present in a properly specified model. Both irregularities in the model's performance are negative and not desirable when training an accurate model.

New techniques, such as transfer learning and data augmentation have been used to surpass and alleviate shortcomings in data volume, and consequently avoid overfitting and underfitting. Transfer learning constitutes the process of learning representations on larger datasets, through the reuse of a model that was initially developed for a similar task (e.g., classification) to the one pretended (Weiss et al., 2016). This can be achieved by training the classifier over a distinct dataset (possibly with different classes), being useful mainly due to the fact that the first layers of CNNs learn to recognize low-level features (e.g. edges, simple geometrical shapes) that, in turn, can constitute beneficial information for different tasks.

Data augmentation, in turn, is the process of increasing the amount of training data. It was selected as a feasible solution to deal with the limitations associated with the size of the dataset, since deep learning algorithms normally require large amounts of training data to fit models able to generalize properly. Data augmentation is obtained *via* the introduction of transformations and slight distortions of the original training data that does not imply significant semantic changes in the information contained by the data, i.e., change of class. This procedure can therefore reduce overfitting when training the model (Sladojevic *et al*., 2016).

Chapter 4

State of the art

4.1. Assessing cultural ecosystem services

Human interests constitute the main drivers of the rapid worldwide loss of biodiversity that has been occurring in the last years (Maxwell *et al*., 2017). Therefore, understanding human-nature interactions becomes extremely relevant to address the biodiversity crisis, as well as to encourage prevention actions and develop new mitigation strategies. The quantitative analyses of social media data remain scarce in conservation-related sciences, with ecosystem services (ES) and, especially, cultural ecosystem services (CES), being no exception. Previous studies on CES have been based predominantly in questionnaires and interviews, which are time consuming and low cost-benefit.

Examples of studies based in questionnaires/interviews include the one developed by Bryce *et al.* (2016), where it was evaluated the benefits of cultural ecosystem services in 151 marine sites from the United Kingdom, through the presentation of an online questionnaire (containing 15 subjective well-being indicators) to a group of recreational drivers or anglers. Similarly, Fish *et al*. (2016b), also studied CES in the Northern Devon Nature Improvement Area (NDNIA) from south west England, through a combination of methods involving structured questionnaire surveys, qualitative mapping (based in the questionnaires), group discussion (based in the questionnaires and mapping) and participatory arts-based research process (based in the questionnaires, mapping and group discussion). Schmidt *et al*. (2016), in turn, explored the sociocultural CES value of Edinburg urban green areas and Pentland Hills, situated in Edinburg, Scotland. The assessment was based on two structured surveys: face-to-face interviews and online surveys. Likewise, Dou *et al.* (2017) proposed a method for quantifying cultural ecosystem services through human perceptions, by making questionnaires and expert interviews, combined with monetary ecosystem services valuation, in six metropolitan areas of Beijing, China.

Also, Hausmann *et al*. (2018) introduced a method for content analysis of georeferenced photos from social media platforms (e.g., Flickr and Instagram) to infer on visitors' preferences in protected areas, through statistical methods. The results didn't reveal significant differences between the analyses of surveys content and social media content, which launches social media data as a potential source of reliable information for assessing environment challenges.

Lastly, Moreno-Llorca *et al.* (2020a) found that social media data and social surveys can provide complementary information on CES assessments pertaining to tourism.

In recent years, the emergence of the iEcology, Digital conservation and conservation culturomics fields brought a wide range of possibilities to address cultural ecosystem services, taking advantage of social media data (including text, video and image). However, most of the studies in this scope still rely on manual content analysis of photographs, which, similarly to the questionnaires/interviews, is time and resource consuming. Examples of studies using social media data include the one developed by Eid & Handal (2018), where illegal hunting in Jordan was studied using social media data, in order to assess its impacts on wildlife; the one proposed by Hausmann *et al*. (2018), based in the manual classification of pictures from Instagram and Flickr, with the purpose of understanding tourists' preferences for nature-based experiences in protected areas; and finally the one from Moreno-Llorca *et al*. (2020b) which used the manual labelling of social media pictures to understand tourist profiles and visitation preferences.

Nevertheless, some advances are being made in the automated image analysis that allow an automated CES assessment. Richards & Tunçer (2018), for example, developed a method for assessing ecosystem services, based on the automating content analysis of social media photographs from Flickr, using the machine learning algorithm provided by Google Cloud Vision. Also, Gosal *et al*. (2019) explored multiple recreational beneficiaries in social media photographs from Flickr, using the machine learning algorithm provided by Google Cloud Vision to classify its content. However, these efforts make use of existing platforms (e.g., Google cloud vision) that are not freely available to the scientific community, making them a low cost-benefit alternative.

4.2. Machine and deep learning for cultural services

Deep learning approaches have become popular methods due to their ability to capture nonlinear patterns and have been established as the state-of-the-art for several image classification tasks, achieving better results compared to other machine learning classifiers (such as support vector machine and k-nearest neighbours), therefore, being widely used in the most diverse areas (Fu & Rui, 2017). Deep learning has been used in other similar works, such as the one developed by Wang *et al*. (2016), where CNNs, more precisely, the AlexNet network architecture and the Caffe open-source software framework, were used to track natural events from social media data. These authors achieved good results that corroborate the ability of CNNs to address image classification tasks and also highlighted the potential of using social media data to tackle conservation events. Other analogous work is the one proposed by Seresinhe *et al*. (2017), where it was adopted a transfer learning approach that uses convolutional neural networks, more precisely, AlexNet, VGG16, GoogleNet and ResNet152, trained over the Places (http://places2.csail.mit.edu/index.html) dataset to quantify the beauty of outdoor places, by analysing over 200 000 photographs of Great Britain that were retrieved from the online game *Scenic-Or-Not.* Overall, this work achieved results with great potential to quantify the scenicness of outdoor spaces, with the VGG16 network architecture constituting the setup with the best performance, followed by GoogleNet and ResNet152, and lastly, AlexNet. Similarly, Koylu *et al*. (2019), proposed a method in which CNNs with kernel density estimation, more precisely "You Only Look Once" (YOLO) real-time object detection, were implemented in order to identify bird images and infer birdwatching activity patterns from geo-tagged social media photos. This work achieved positive results that, once again, support the use of CNNs to address social media content classification tasks in the context of conservation events.

Nowadays, the assessment of CES using social media data is fundamentally based on three main machine learning and/or deep learning methodologies: a first one (more common), that is essentially focused on spatial and spatio-temporal analysis combined with content evaluation of human-nature interactions, through social media locations and timestamps (Fisher *et al*., 2018); a second one, based on social media geotags, text, image and video content for monitoring biodiversity and natural/landscape features (Dylewski *et al*., 2017); a third one, focused on text analysis of online discussions, perceptions and reactions to conservation-related events, news and management activities (Wu *et al*., 2018). One example is the work developed by Gosal & Ziv (2020), where the landscape aesthetics of the northern English Protected Area of the Yorkshire Dales National Park were studied through the use of social media photographs, analysed using the machine learning Google's Cloud Vision application programming interface (API), in combination with paired-comparison surveys, probability modelling, machine learning based with text annotations, natural language processing and regression analysis.

Other works have been developed through the last years in the areas of digital conservation, iEcology and conservation culturomics which offer many opportunities for CES evaluation. As an example, Hafemann *et al.* (2014) presented a method for forest species, that was implemented using CNNs to analyse two datasets (one with macroscopic images and other with microscopic images of Brazilian forest species). These CNNs were based on state-of-the-art models for object classification, presenting a structure (an input layer, a set of convolutional layers and pooling layers, a locally connect layer and a fully connected output layer) that enhances this task. Hafemann *et al*. (2014) achieved great results, comparable to the state-of-the-art ones, emphasizing the potential of CNNs for object detection in the scope of forest species recognition. Also, Salman *et al*. (2016) proposed a deep learning approach based on CNNs, more precisely, on a K-layered convolutional neural network, in conjunction with conventional machine learning algorithms: k-NNs and SVMs. The main goal consisted in the classification of images of fishes in unconstrained underwater environments, where the results revealed a superior classification performance for the CNNs coupled with k-NN and SVMs, when compared to classifying with only k-NN and SVMs.

Similarly, Richards & Tunçer (2018) proposed an approach for automating the assessment of cultural ecosystem services from social media photographs, using the Google Cloud Vision software to analyse and group, through hierarchical clustering, over 20,000 photographs from Singapore. This automated classification method was then compared to a manual classification. These authors achieved accurate results, with great potential to quantify CES and to support urban planning, especially in large areas, where the manual classification is not cost-effective. Lee *et al*. (2019) proposed a method for analysing large amounts of social media photographs (from the Mulde river basin in Saxony, Germany) and for deriving indicators of socio-cultural usage of landscapes, through cluster detection with CNNs, using the computer vision and artificial intelligence enterprise platform Clarifai. This platform uses deep CNNs to identify and analyse image and video content. Lee *et al*. (2019) reached satisfactory and reliable results that provide a source of knowledge to the spatial explicit monitoring of CES activities, as well as for landscape management and planning, even in the presence of large volumes of data.

Likewise, Gosal *et al*. (2019) studied multiple recreational beneficiaries across the Camargue region in Southern France, in order to predict and map beneficiaries' types and choices through the combined analysis of machine learning techniques, natural language processing (latent semantic analysis (LSA)) and self-organizing maps (SOM) of 20,000 social media photographs retrieved from Flickr. These photographs were automatically annotated in descriptive terms using Google's Cloud Vision API. The results revealed great potential for decision-making and urban planning, supporting the development of smarter strategies by park managers and constituting a more cost-effective method that can be used at the expense of surveys in the field.

Willi *et al*. (2019) proposed an automatic method, based on the transfer learning strategy, for identifying animal species in camera trap images using deep learning, where CNNs, more precisely, the ResNet18 network architecture, were implemented using the TensorFlow platform, in order to distinguish between animal species, humans, vehicles and empty photographs. The results revealed a finest classification performance for the models trained with transfer learning, when compared to the models trained from scratch, which leverages the potential of this technique. Also, the results demonstrated to be a viable, and less time consuming, alternative to manual classification of images. Finally, Hausmann *et al.* (2020)

studied the visitor’s sentiment, using automated natural language processing to analyse Instagram posts geolocated inside four national parks in South Africa.

Chapter 5

Implementation details

5.1. Test areas

The test area (Figure 5.1) is “Peneda-Gerês” and includes the National Park and protected areas under the Natura 2000 network (special protected areas, SPAs, and a site of community importance, SIC, in Northern Portugal). The area covers over 950 km^2, presenting a temperate Atlantic to sub-Mediterranean climate, with a mean annual temperature of 13-15 ºC, a total annual precipitation that, normally, surpasses the 2000 mm, an altitude that varies between 100 and 1548 m a.s.l. and, predominantly, a granite bedrock. Apart from its rich biodiversity and mountain landscapes with native scrublands, grasslands and *Quercus* woodlands, “Peneda-Gerês” also contains a diverse archaeological and historical heritage (like megalithic monuments and signs of Roman occupation, traditional celebrations and land-use practices), which makes it a very popular area for recreation and other socio-cultural activities (Santarem *et al*., 2015). “Peneda-Gerês” was selected as the test area to develop an automated classification of social media photographs since it constitutes one of the main conservation areas in Portugal, presenting also a natural capital that underlie many CES, namely through recreational and touristic activities.

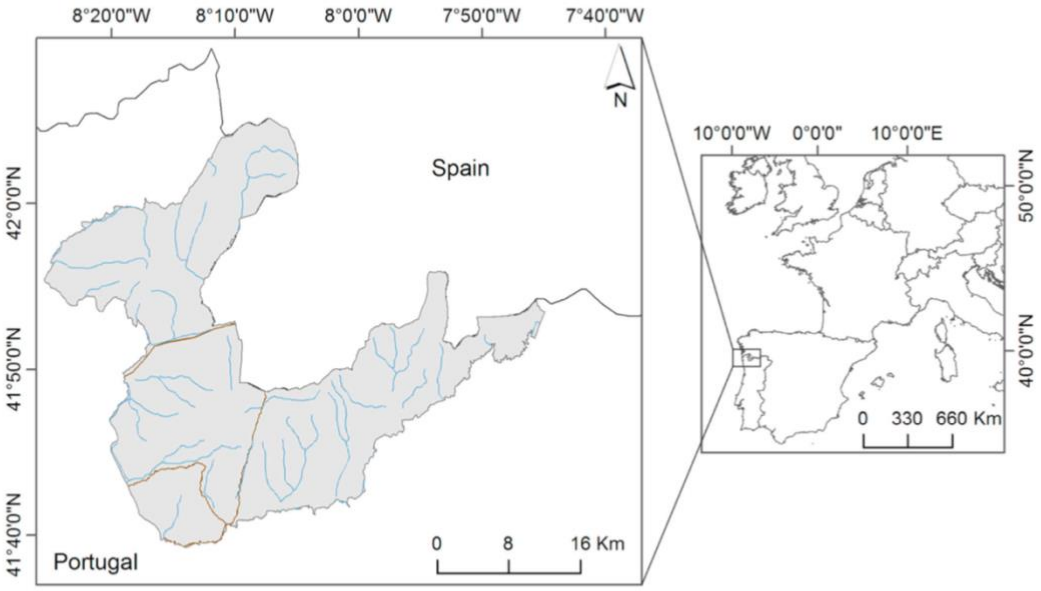


Figure 5.1. Representation of one of the study areas, “Peneda-Gerês” (on the left), and its location in southwestern Europe (on the right). The figure on the left also illustrates the position of the main roads (brown) and rivers (blue). Figure extracted from Vaz *et al.* (2019).

In this study it was also considered an additional area (Figure 5.2), the UNESCO Biosphere Reserve “Sierra Nevada”, in order to test the reproducibility of the automated classification. “Sierra Nevada” spreads over 1,722 km^2, being a major mountainous region of Andalusia (Granada and Almería provinces), in southern Spain (elevation between 860 and 3,482 m a.s.l.). Hosting more than 80 endemic plant species and more than 2,300 taxa of vascular flora in total, “Sierra Nevada” is considered one of the most important biodiversity hotspots in the Mediterranean region. Besides containing several species listed in the European Union Habitats and Birds directives, its socio-economy is supported by several social-cultural activities, including rural tourism and sports (Ros-Candeira *et al*., 2020). Furthermore, this area holds several protection regimes (Natural and National Parks, Natura 2000 Special Protection Area and Special Area of Conservation, Biosphere Reserve) and is part of the European Long-Term Ecosystem Research Infrastructure.

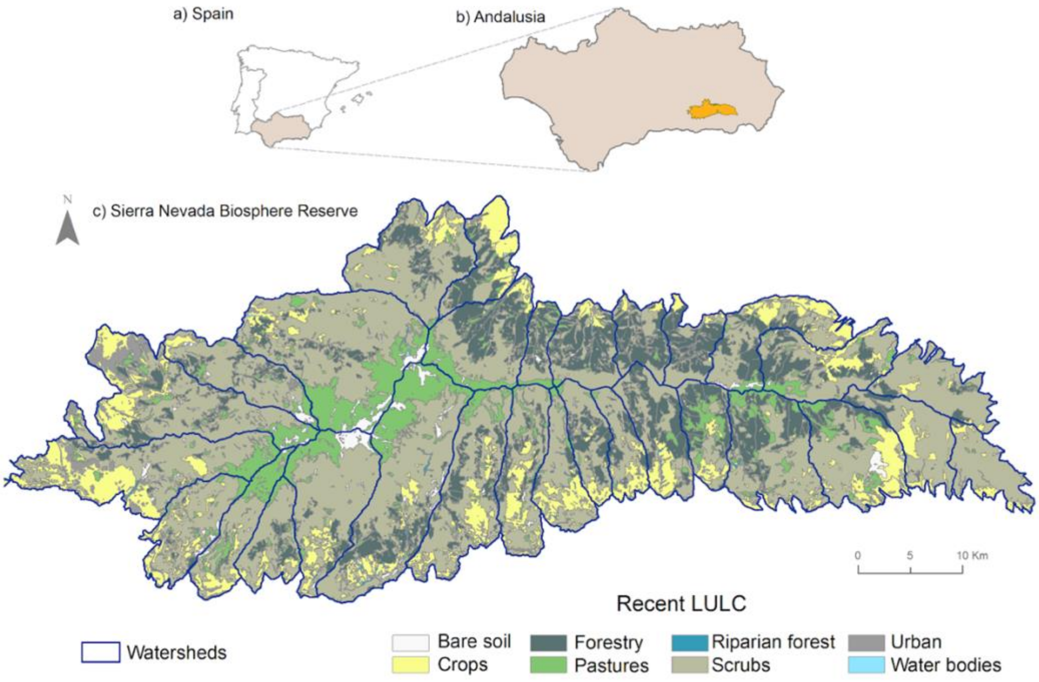


Figure 5.2. Representation of one of the study areas, “Sierra Nevada” (at the bottom), and its location in the Iberian Peninsula (on the left) and at the Andalusia region (southern Spain). The figure on the bottom also illustrates the main land use and land cover (LULC) types. Figure extracted from Moreno-Llorca *et al*. (2020a).

5.2. Mining data from social media

In this study, social media data was collected from the Flickr (https://www.flickr.com/) and Wikiloc (https://www.wikiloc.com/) platforms. Flickr was selected because of its high temporal coverage and due to the fact that the users who benefit from this platform are usually more "nature-oriented", being the photographs that users upload more related to their surrounding environment in the wild (which encompass the scope of this study). Wikiloc, in its turn, was selected because it contains photographs of nature trails that were uploaded or shared by people and that are directly related to touristic and recreational activities in the wild (e.g. hiking, cycling). Specifically, it was considered the geographically referenced social media photographs published by Flickr and Wikiloc users from 2003 to 2017 inside "Peneda-Gerês". In respect to the General Data Protection Regulation 2016/679, social media data protected by users' rights was not downloaded nor analysed. Public data that would potentially contain personal information from social media users was kept anonymous through the study.

The photographic dataset was retrieved through the use of the freely available Flickr's Application Programming Interface (API), indicating a time window and a bounding box with a pair of coordinates (in our case: minimum latitude: 41.653104; maximum lat.: 42.083595; min. longitude: -8.426270; max. lon.: -7.754076) around "Peneda-Gerês". This information was then saved as an excel file with the following attributes: user-id, date taken, latitude, longitude, picture uniform resource locator (url).

Then, a first classification was performed by dividing the photographs of the dataset into "Indoor" and "Outdoor" classes (Figure 5.3).

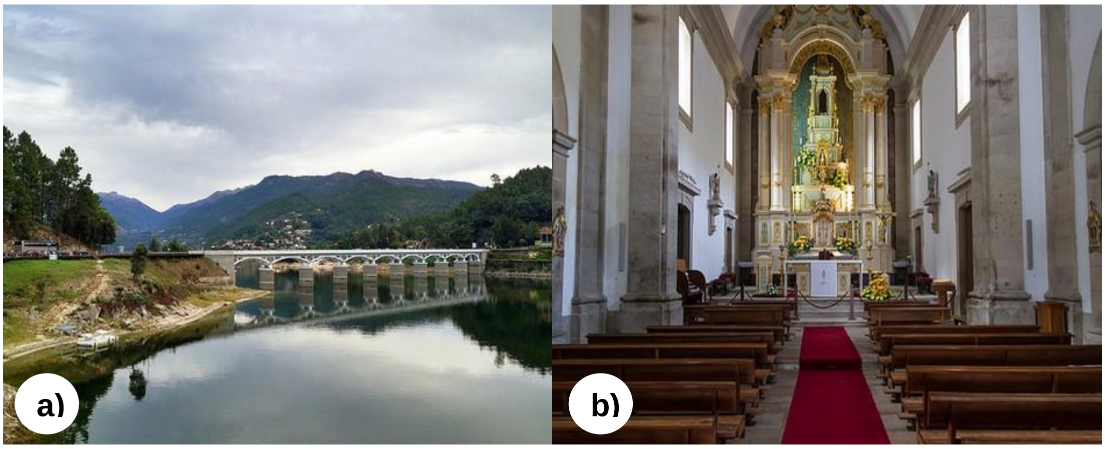


Figure 5.3. Examples of images belonging to the Outdoor (a) and Indoor (b) classes.

Only the “Outdoor” pictures were included in this study, since CES are directly connected to nature and environment, which in turn are related to the outside/outdoor. The “Outdoor” images were further divided into two main classes, “Nature” and “Human”, depending on whether the image was dominated by natural or man-made elements (Figure 5.4).

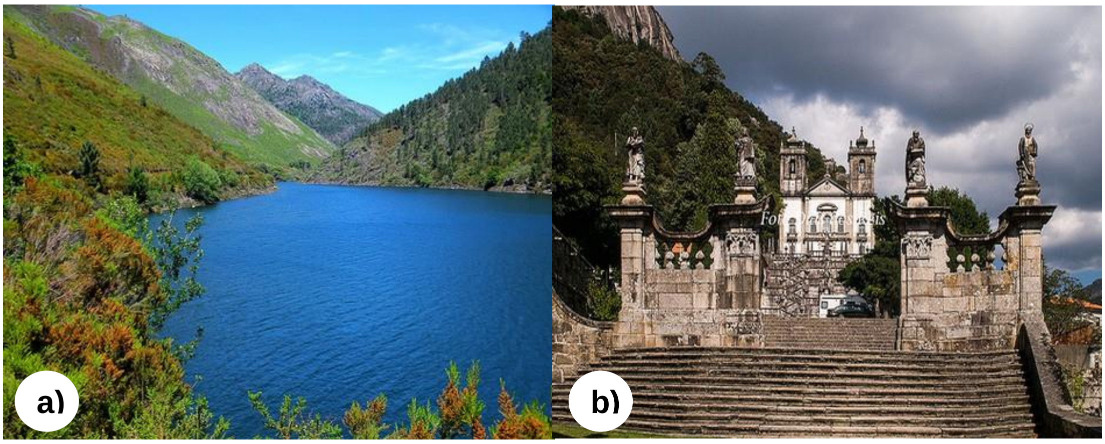


Figure 5.4. Examples of images belonging to the Nature (a) and Human (b) classes.

Lastly, a finer classification for outdoor images was also provided, which encompasses the following six classes: “Species”, “Landscape”, “Nature”, “Human activities”, “Human structures” and “Posing” (Figure 5.5). “Species” pictures respectively pertained to close-up shots of animals or plants in the wild, translating CES pertaining to biodiversity appreciation (Goodness *et al*., 2016). “Landscape” pictures show wide-open shots of nature in the wild, often with a visible horizon most often representing people’s enjoyment of landscape aesthetics (Richards & Friess, 2015). “Human activities” include pictures where people engage in by recreational activities (Richards & Friess, 2015), for instance related to sports such as ski or cycling. “Human structures” include those pictures where man-made structures dominate in the wild, e.g. historical monuments and churches, capturing situations of cultural heritage and spiritual enrichment (Blicharska *et al*., 2017). “Posing” refers to pictures with people looking at the camera, with recognizable faces, testifying social enjoyment and sense of identity (Riechers *et al*., 2016). Finally, “Nature” pictures capture natural elements with no particular feature (such as species) but with an intermediate shot (differing from wide-open shots attributed to landscapes), expressing the appreciation of nature by people (Richards & Friess, 2015).

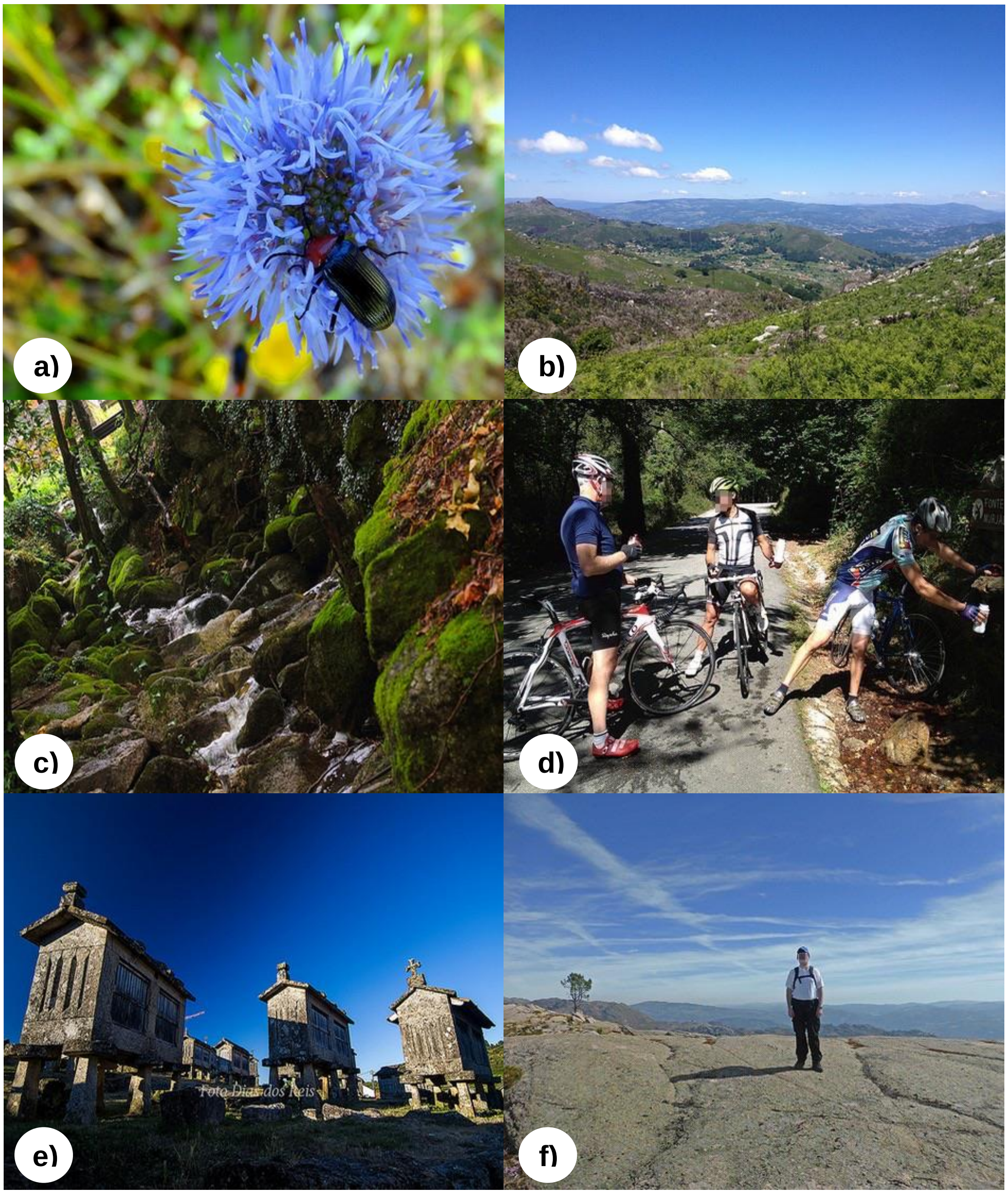


Figure 5.5. Examples of images belonging to the classes of: a) Species, b) Landscape, c) Nature, d) Human activities (the picture was edited in order to protect the identity of the persons posing in the picture), e) Human structures, f) Posing (the picture was edited in order to protect the identity of the person posing in the picture).

The data collection and manual classification processes were performed previously as part of a previous study (see Vaz *et al*., 2019).

For “Sierra Nevada”, we followed the exact same procedures mentioned above, but unlike the verified for “Peneda-Gerês” (2003 to 2017), the time frame selected for this area was from 2004 to 2017. Data collection and manual classification processes were also done as part of a previous study (Ros-Candeira *et al.*, 2020). Similar to what was tailed for “Peneda-Gerês”, a first classification was performed by dividing the photos of the dataset into “Indoor” and “Outdoor” classes (Figure 5.6). Only the “Outdoor” pictures were included, for the same reasons mentioned for the “Peneda-Gerês” dataset.

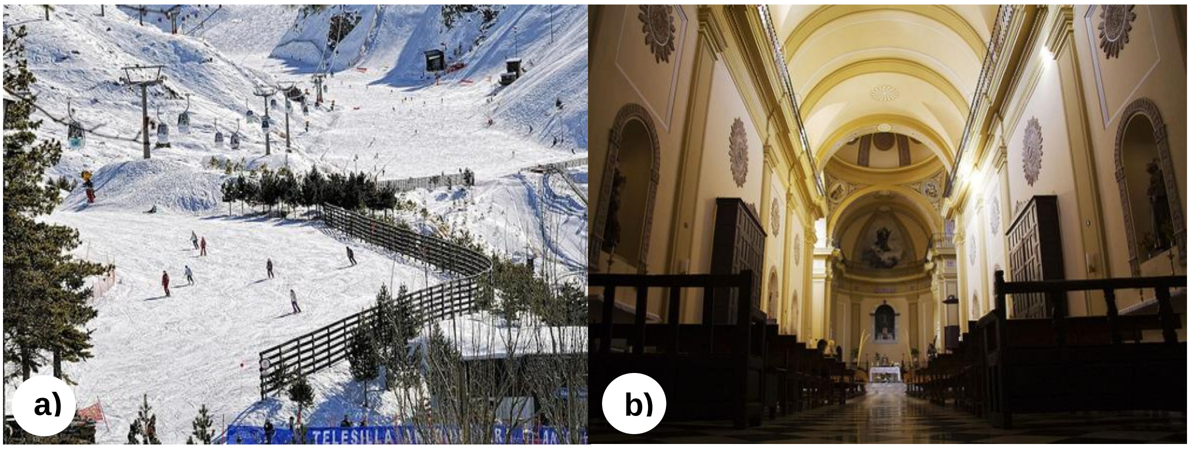


Figure 5.6. Examples of images belonging to the Outdoor (a) and Indoor (b) classes.

This was followed, again, by the division into two main classes, “Nature” and “Human” (Figure 5.7).

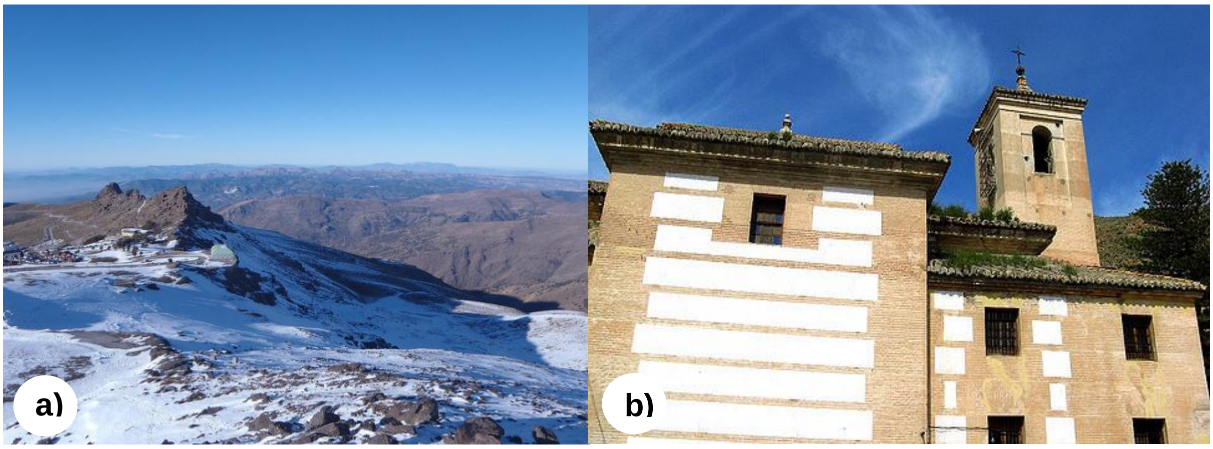


Figure 5.7. Examples of images belonging to the Nature (a) and Human (b) classes.

Lastly, a finer classification for “Outdoor” images (Figure 5.8) was also provided (“Human activities”, “Human structures”, “Landscape”, “Nature”, “Posing” and “Species”), following the exact same method of manual classification for both of the datasets ("Peneda-Gerês" and "Sierra Nevada").

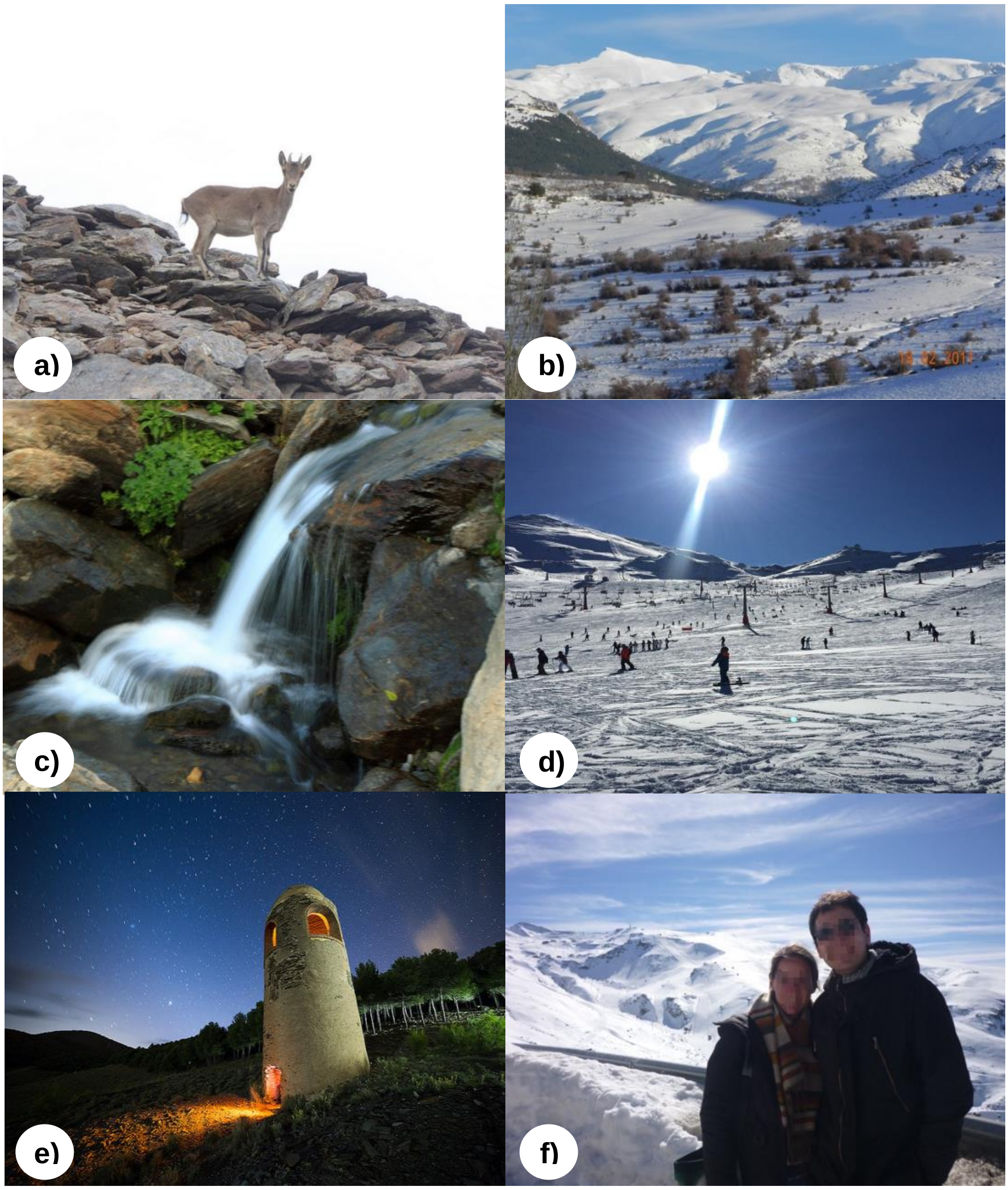


Figure 5.8. Examples of images belonging to the classes of: a) Species, b) Landscape, c) Nature, d) Human activities, e) Human structures, f) Posing (the picture was edited in order to protect the identity of the persons posing in the picture).

The final dataset of "Peneda-Gerês" included a total of 1861 pictures, while the one of "Sierra Nevada" comprised a total of 881 photographs. The description of each class in both datasets is displayed in Table 5.1.

Table 5.1. Classes considered to classify each social media photograph according to its main focus.

Category/class	Description
1st classification level	
Nature	The photo shows a dominance of natural features (e.g. forests, rivers) with no or very little human influence
Human	The photo shows a dominance of human/build features (e.g. houses, humans, cattle)
2rd classification level	
Species	Trees or parts of trees (e.g. flowers or leaves) as main subject
Landscape	Pictures showing wide views of an area, with visible horizon
Nature	The photo shows a dominance of natural features (e.g. forests, rivers) with no or very little human influence
Human activities	People engaged in recreational activities (e.g. hiking and biking), including related objects (e.g. canoes and bicycles)
Human structures	Pictures showing human infrastructures (e.g. houses or monuments)
Posing	People looking at the camera, with recognizable faces

5.3. Data pre-processing

5.3.1. Peneda-Gerês dataset

The data initially collected was pre-processed in order to suit the needs of this study. To achieve that, filtering and transformation steps, such as cleaning, selection and normalization, were applied to the data. The photographs belonging to the "Indoor" class were excluded from the study, as well as the photos without corresponding labels and the labels without matching photos. Initially, the "Peneda-Gerês" dataset comprised a total of 1861 photographs and labels. In the end, the datasets remained with only 1778 for "Peneda-Gerês". The frequency of pictures in each class is displayed in Figure 5.9.

Since the platforms selected for this study required some specific pre-processing steps, all the photographs were resized to the same resolution, that was computed taking into account the mean resolution of the set. Later, the pictures were normalized and scaled, and the respective labels were converted into binary class matrices.

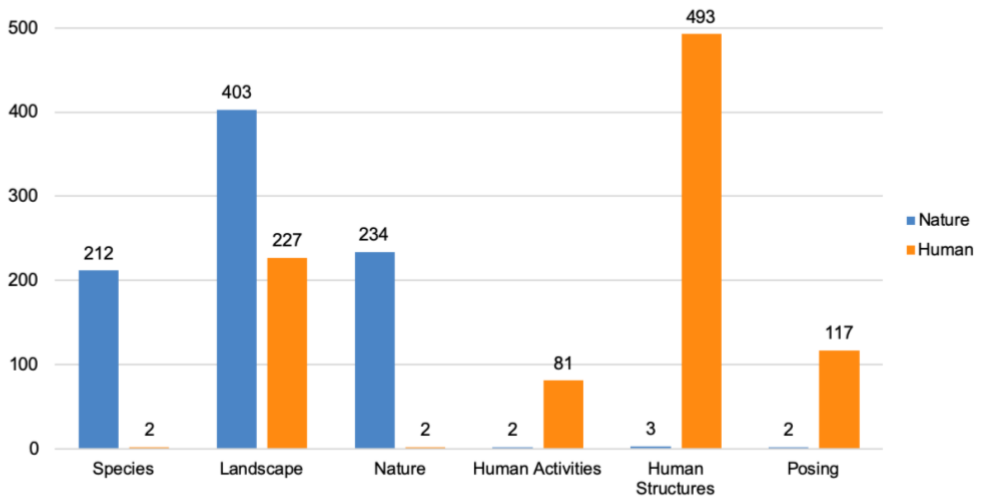


Figure 5.9. Distribution of the number of outdoor photographs for each class in the “Peneda-Gerês” dataset.

5.3.2. Sierra Nevada dataset

The data corresponding to the study area “Sierra Nevada” was pre-processed following the same steps implemented in the “Peneda-Gerês” dataset. Initially, the “Sierra Nevada” dataset was composed of 880 photos and 889 labels. After cleaning the data, only 745 photographs remained. The frequency of pictures in each class is displayed in Figure 5.10.

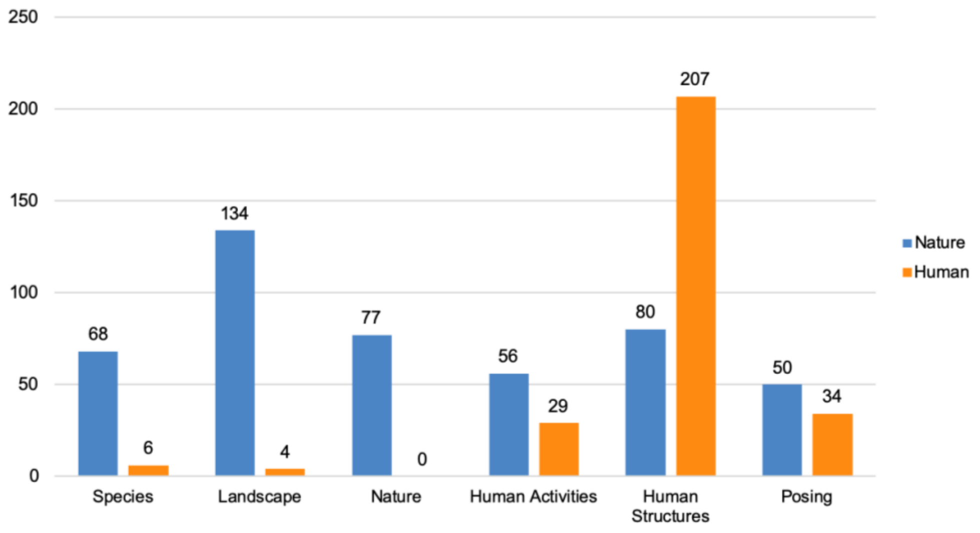


Figure 5.10. Distribution of the number of outdoor photographs for each class in the "Peneda-Gerês" dataset.

5.4. Automated Image Classification Methodology

First, we performed an initial classification of the content of social media photographs based on "Nature" and "Human" labels, followed by a second classification based on "Human activities", "Human structures", "Landscape", "Nature", "Posing" and "Species" labels. To achieve that, two different convolutional neural networks architectures were implemented in order to verify which one was indeed the most appropriate and suitable for our study. In this regard, it was used the Keras (https://keras.io/) platform with TensorFlow (https://www.tensorflow.org/) backend, which constitute two of the most used libraries for building and training deep learning models, especially CNNs (Figure 5.11). The algorithms were implemented using a freely available GPU in the Google colab environment (https://colab.research.google.com/), as well as the programming language Python (https://www.python.org/).

The proposed image classification methods were evaluated over the dataset described in section 5.2 using the k-fold cross-validation, a method that corresponds to the random partition of the original dataset into k equal sized subsets. One of these subsets is then retained for testing the model, while the remaining k-1 subsets are used as the training set. This procedure is then repeated so that each subset is used only once for testing the model. In this project, a 5-fold-cross validation method was adopted, since 5 is the number of partitions most used/cited in the literature, considering the computational resources and the running time. To achieve that, the original dataset was randomly partitioned into 5 equal sized subsets.

Also, a random seed of 7 was established, a value is used to initialize a pseudorandom number generator and ensure reproducibility. The performance metrics were computed as the mean of the performance metrics obtained over the 5 different folds. Also, during the training, in each of the 5 folds, part of the training data, more specifically, 10%, was retained to perform model validation, in order to determine the best training parameters (validation accuracy and loss). The chosen training parameters were the ones that guaranteed the highest accuracy over the validation set. Since we are coping with a small dataset, two deep learning associated approaches were implemented to improve the generalization of the model and avoid overfitting: transfer learning and data augmentation.

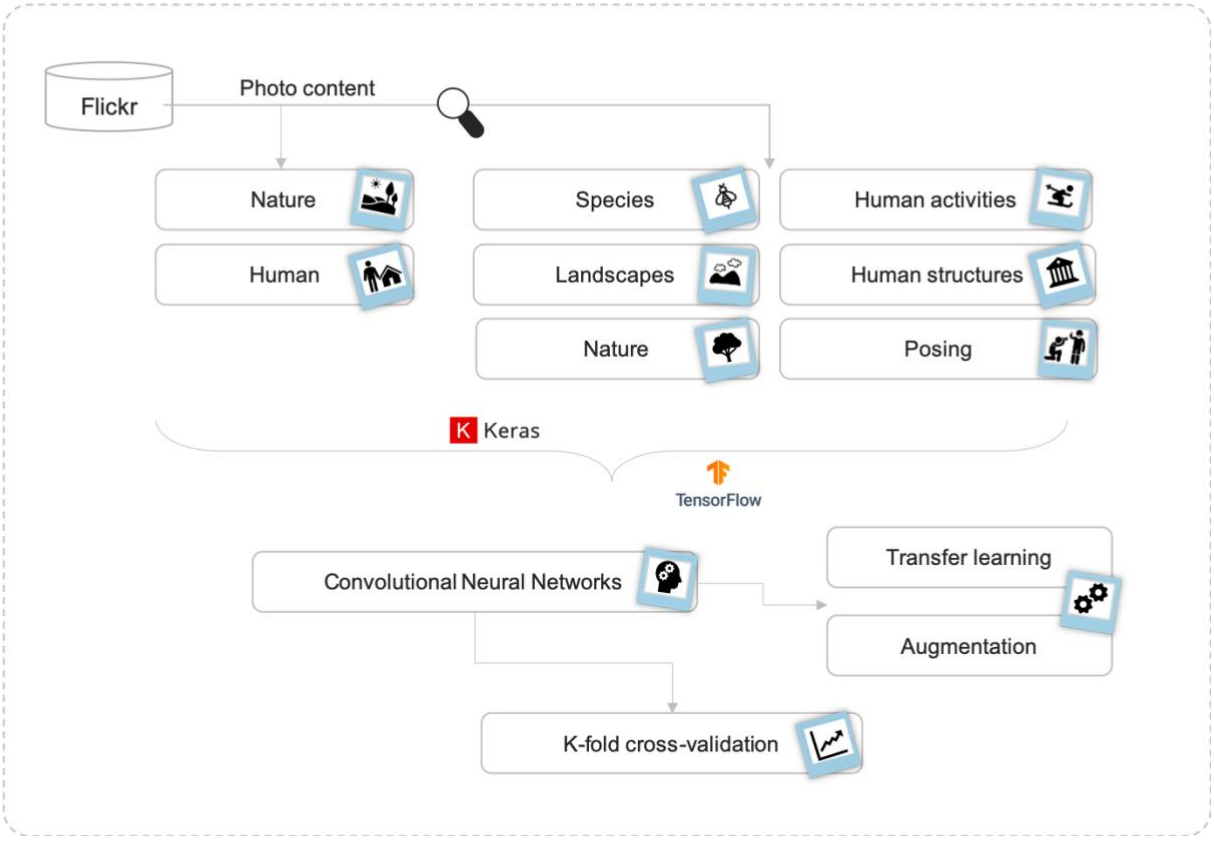


Figure 5.11. Workflow of the image classification methodology.

5.4.1. Convolutional neural networks architectures

As mentioned above, two different convolutional neural networks architectures were implemented in this study: VGG16 and ResNet152. The VGG16 convolutional neural network, which describes very deep convolutional networks for large-scale image recognition, was initially proposed by Simonyan & Zisserman (2014) as an improvement of the AlexNet, that was achieved by replacing large kernel-sized filters (11 and 5 in the first and second convolutional layer, respectively) with multiple 3×3 kernel-sized filters one after another. This allows to simulate large filters without losing the advantages and benefits of a small filter. VGG16 was submitted to ImageNet Large Scale Visual Recognition Challenge 2014 (ILSVRC2014), where it achieved 92.7% test accuracy, placing it among the top 5 best performances in ImageNet (Simonyan & Zisserman, 2014). Besides the convolutional layers, VGG16 (Figure 5.12) has 5 max-pooling layers, 3 fully connected layers and a softmax layer (output layer). Initially proposed for large-scale image recognition, VGG16 also demonstrated to be very effective for image classification tasks.

During training, each image is passed through a stack of convolutional layers, where two different types of filters are applied: 3x3, which allows to capture the notion of left/right, up/down, centre, and 1x1, which provides a linear transformation of the input channels. Then, max pooling is applied to the images through the use of 2x2 kernels, followed by three fully connected layers and a softmax layer, that provides a probability distribution for each class.

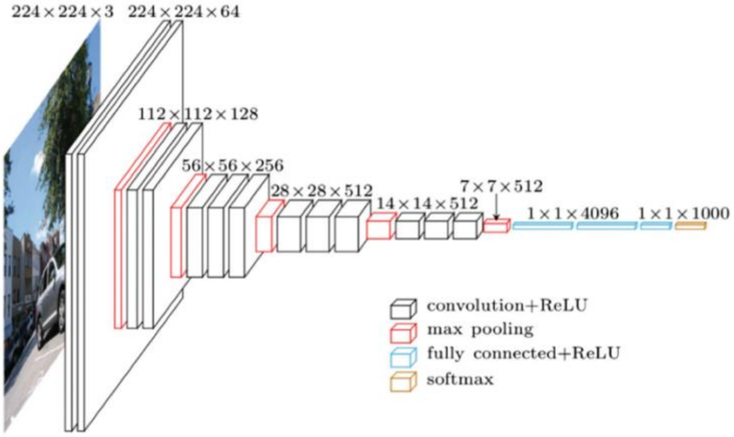


Figure 5.12. Configuration of the VGG16 model architecture. Retrieved from Nash *et al.*, 2018.

The ResNet152 (residual neural network), in turn, represents very deep convolutional networks, that incorporate the differential of introducing a structure called residual learning unit, capable of avoiding the degradation of deep neural networks (He *et al*., 2016). These units allow the network to jump over some layers, stopping these layers from changing the values of the gradient and, consequently, contributing to avoid the problem of vanishing gradients, since activations from a previous layer are reused until the next layer learns its weights. Normally, the ResNet architectures are implemented (Figure 5.13) considering double or triple layer jumps that comprise nonlinearities (ReLU) and batch normalization in between. In other words, each layer is connected to the next layer and to the layers about 2-3 hops away (He *et al*., 2016). ResNet came in first place at the ILSVRC 2015 in the areas of classification, detection and location of images, as well as in the Microsoft Common Objects in Context (MS COCO) 2015 detection, and segmentation.

During the training, the model learns which layers are effectively contributing to improve its performance, as well as those that compromise its effectiveness. If the layers improve the performance, are maintained in the model, while the others become identity mappings.

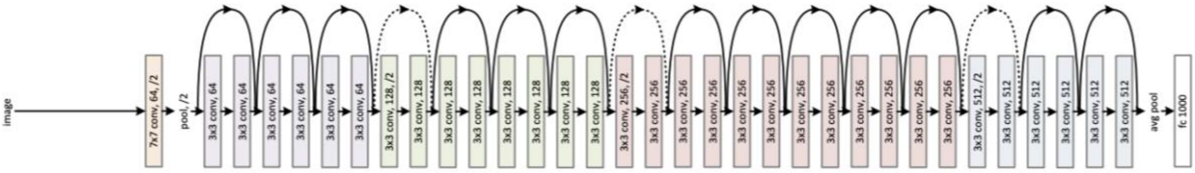


Figure 5.13. Configuration of a residual network with 34 parameter layers (3.6 billion FLOPs). Retrieved from He et al., 2016.

Both CNNs' architectures are significantly slow to train, taking a lot of storage memory, that is associated with the high number of layers (more precisely, 16 and 152) presented in these networks. However, they were selected for this study due to its easy and simple implementation and high performance with great results in image classification tasks, which makes these types of networks the state-of-the-art for several computer vision analysis. VGG16, for example, has the pre-trained weights freely available online, while residual neural networks have the advantage of allowing the consideration of deeper architectures without incurring in vanishing gradient problems.

5.4.2. Transfer learning

Transfer learning is the selection of a dataset with similar features to those present in the dataset under study, as a way to extract pre-trained weights. For both of the networks architectures mentioned above three different weights were considered: one extracted from networks trained in the database "Places" (http://places2.csail.mit.edu/), more precisely, in the dataset "Places365" (https://github.com/CSAILVision/places365), one withdrawn from networks trained in the database "ImageNet" (http://www.image-net.org/) and one that represents the weights obtained by training only over our dataset.

The database "Places" comprises around 10 million scene photographs, labelled with 434 scene semantic categories, covering about 98 percent of the places a human can encounter in the world, including, in a way, photographs with similar elements to the ones under study (Zhou *et al*., 2017). The Places365 dataset, in turn, is the latest subset of the database Places, containing around 1.8 million scene photographs, labelled with 365 scene semantic categories. The ImageNet database constitutes a large-scale hierarchical image database, that has several applications in the broadest areas, comprising more than 14 million cleanly annotated images spread over around 21,000 categories and providing a significantly representative and diverse coverage of the image world (Mettes *et al*., 2016). Both of these databases were also selected due to their freely available online resources (weights and models), that facilitated our work.

Regarding the details of the transfer learning strategy implemented, all the convolutional layers were kept frozen when training over our dataset, while the remaining 3 (for VGG16) and 1 (for ResNet152) fully connected layers were trained with our dataset. Moreover, an additional dense layer with 128 units and a rectifier linear unit activation function was also included before the output layer, which was modified in order to have 2 or 6 units.

5.4.3. Training

Regarding the chosen optimizers and training parameters, Adam was the selected optimizer, an algorithm for first-order gradient-based optimization of stochastic objective functions, since it is one of the most used optimizers in the scope of deep learning and was indeed the most suitable for the classification models under study (Kingma & Ba, 2014). For batch size, an hyperparameter of gradient descent responsible for controlling the number of training samples to be considered before the model's internal parameters are updated (Brownlee, 2016), it was chosen a mini batch size of 10. This value was selected mainly because of the small size of our data and, especially, due to limitations on available memory usage.

For the six approaches mentioned above - VGG16 with the weights from Places365, VGG16 with the weights from ImageNet, VGG16 with the weights trained only over our dataset, ResNet152 with the weights from Places365, ResNet152 with the weights from ImageNet and ResNet152 with the weights obtained by training only over our dataset -, the parameters were tuned over the validation set (in each fold of the proposed 5-fold cross-validation setup, 10% of the training data was reserved for validation), in order to improve the performance of the models. This tuning consisted essentially in the variation of the learning rate and number of epochs (hyperparameter of gradient descent responsible for controlling the number of complete passes through the training set). At an initial stage, for both of the architectures and for the three approaches, it was considered the keras default learning rate (0.001), as well as 100 epochs.

An early stop approach was also implemented, which is a method used to prevent overfitting when training with an interactive algorithm, such as gradient descent, being widely used in the most diverse fields of deep learning due to its simplicity of understanding and implementation (Nielsen *et al*., 2017). To achieve that, it was generated a validation set, setting aside 10% of the training data, in order to use the validation loss as the stopping criteria. In other words, at the end of each epoch, the validation loss is computed on the validation data, terminating when it stops improving (i.e., when it stops decreasing). A patience value of 16 was also established, adding a delay to the trigger in terms of the number of epochs (16) on which we expect to perceive no improvement.

For the first binary classification (“Nature” and “Human”) and for both neural networks (VGG16 and ResNet152), it was verified that the keras default learning rate had a very poor performance when fitting the model. With that under consideration, the learning rate was tuned in order to find the value that was most suitable for our model. For VGG16, the best performance was verified when considering a learning rate of 0.000001 while, for ResNet152, 0.0001 was the most accurate learning rate. Both network architectures (VGG16 and ResNet152) were implemented considering 100 epochs. The behaviour of both networks when training with the Places365 weights is displayed in Figure 5.14.

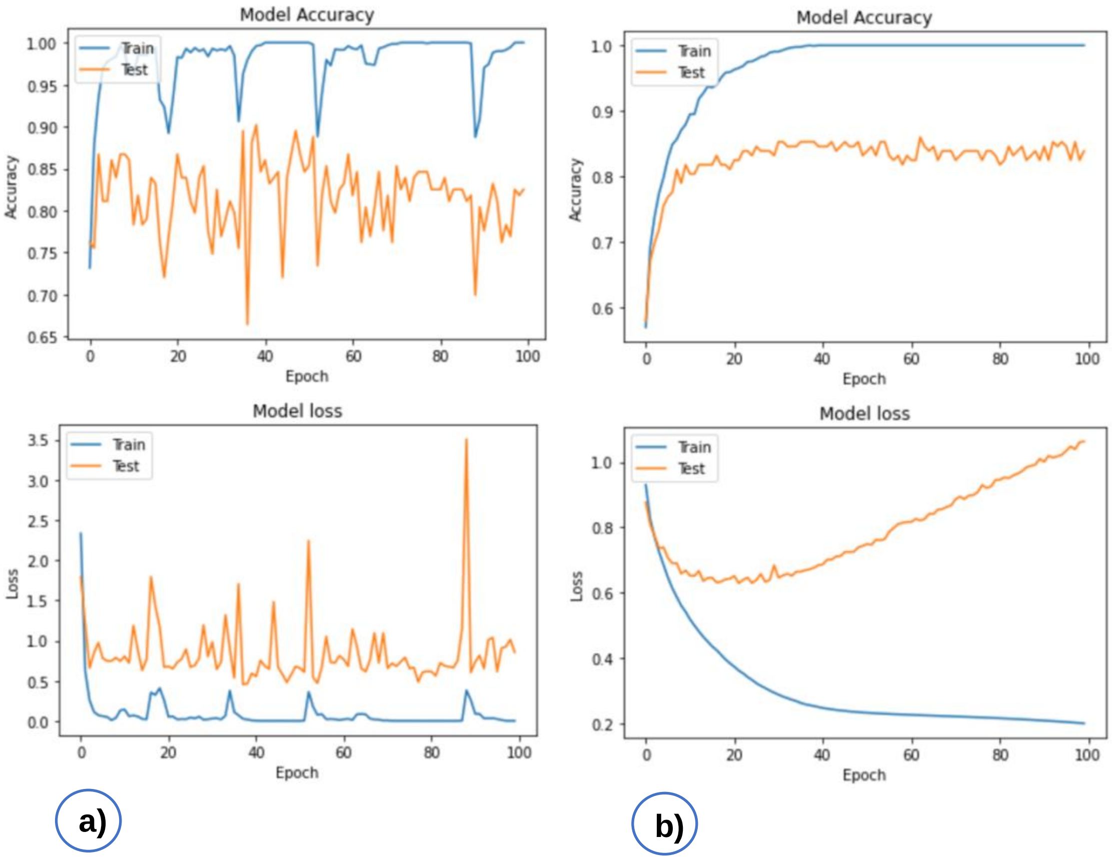


Figure 5.14. Behaviour of ResNet152 (a) and VGG16 (b) with Places365 weights, in terms of accuracy, validation accuracy, loss and validation loss.

For the second classification (multilabel classification), similarly to that verified for the “Nature” vs. “Human” classification task, for both of the neural networks (VGG16 and ResNet152) the Keras default learning rate (0.001) had a very poor performance when fitting the model. With that under consideration, and analogously to the procedure followed for the first classification, the learning rate was tuned in order to find the value that was most suitable for our model. For VGG16, the best performance was verified when considering a learning rate of 0.000001 while, for ResNet152, it was 0.0001 the most accurate learning rate (same values obtained for the first classification task). The number of epochs considered for this classification task was also 100. The behavior of both networks after training with the ImageNet weights is displayed in Figure 5.15.

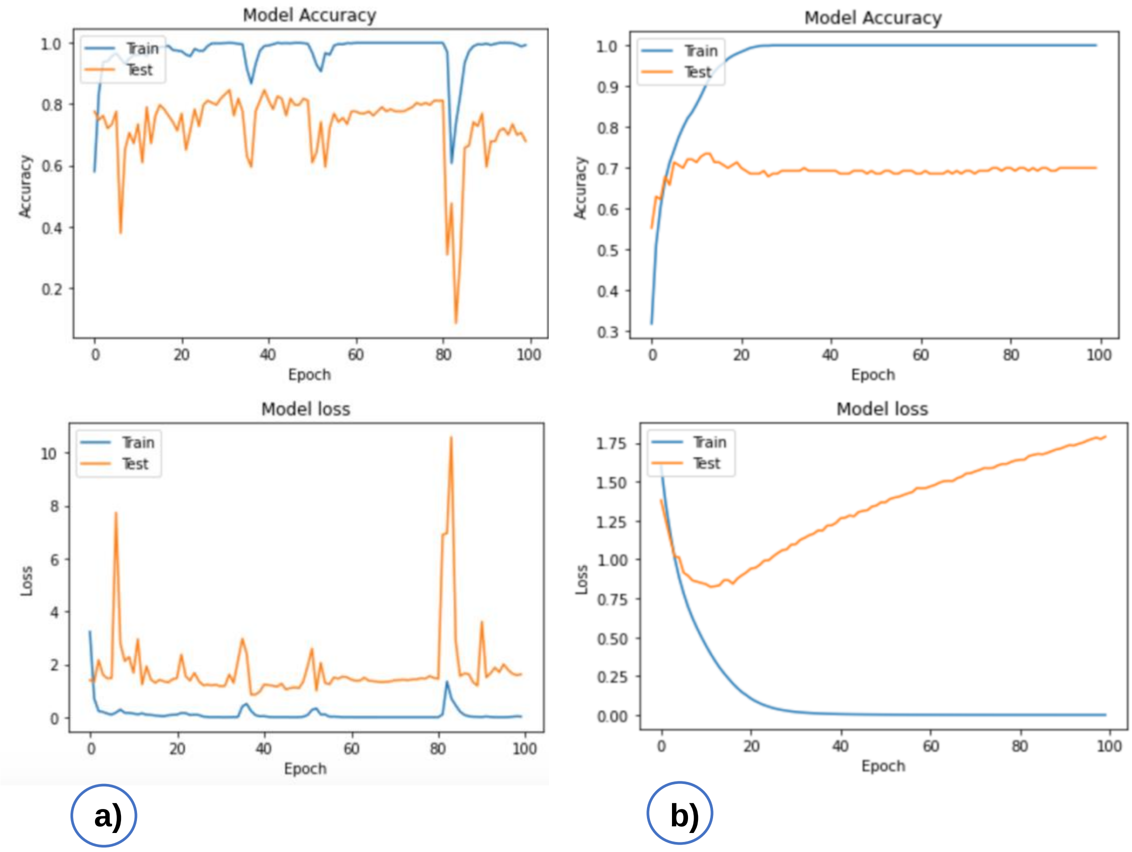


Figure 5.15. Behaviour of ResNet152 (a) and VGG16 (b) with ImageNet weights, in terms of accuracy, validation accuracy, loss and validation loss.

5.4.4. Data augmentation

In order to implement data augmentation, 5 transformations (including horizontal flip, width shift, height shift and zoom) were implemented for each of the images belonging to the training set, which resulted in a total of 8532 or 8538 transformations, depending on the fold. The images in the validation set were not included in this process, in order to avoid biased results. The total number of transformations applied to each photograph (5 per image) was selected taking into account the overall running time of the algorithm, as well as the available computational memory.

In the first implementation, it was observed that, for VGG16, the model accuracy and loss had fully converged after 50 epochs, having been decided, because of that, to use only 50 epochs to build the VGG16 model after data augmentation. The number of epochs established for the ResNet152 model was also 50, due to computing resource management.

5.4.5. Performance metrics

The metrics selected to evaluate the model's performance were classification accuracy (ACC), sensitivity (TPR, true positive rate or recall), specificity (TNR, true negative rate) and F1-score (F_1, f-score or f-measure), since they constitute the three most frequently analyzed metrics when considering a classification problem (Tharwat, 2020). In order to compute these metrics, it was first calculated a confusion matrix, which is a specific table arrangement that allows visualizing the performance of an algorithm, normally a supervised learning one, by opposing instances in a predicted class with instances in an actual class (Powers, 2011). Also, this matrix allows to report the number of true positives, false positives, true negatives and false negatives. In the "Nature" vs. "Human" classification, the terms negative and positive refer to "Nature" and "Human" categories, respectively. The images that were correctly classified as "Nature" were assumed to be true negatives (A), while the ones that were correctly classified as "Human" were labeled as true positives (D). The photographs that were classified as "Nature" having "Human" as its real category are described as false negatives (C) and the ones classified as "Human" with "Nature" being its actual label are represented as false positives (B).

Accuracy represents the proximity of the measurement results to the true value (Eq 4.1), while specificity indicates the proportion of labels that were correctly classified for the category that was previously selected as the negative (Eq 4.2).

$$ACC = \frac{D + A}{D + A + B + C} \qquad \text{Eq. (4.1)}$$

$$TNR = \frac{A}{A + B} \qquad \text{Eq. (4.2)}$$

Finally, sensitivity corresponds to the proportion of labels that were correctly classified for the category that was previously selected as the positive (Eq 4.3) and F1-score computes the harmonic mean of the precision and recall, being a measure of a test's accuracy (Eq 4.4):

$$TPR = \frac{D}{D + C} \qquad \text{Eq. (4.3)}$$

$$F_1 = \frac{2D}{2D + B + C} \qquad \text{Eq. (4.4)}$$

In the "Nature" vs. "Human" classification, these metrics were computed as the mean of the performance metrics obtained over the 5 different folds. For the multilabel classification these metrics were calculated taking into account the macro average, in which the value for each metric is first calculated independently for every label/class, being posteriorly used to compute the unweighted mean. Details on these metrics, including on their calculations and interpretation are shown in Table 5.2.

Table 5.2. Example of a confusion matrix used to compare the manual and automatic classification of the social media photographs into the "Nature" and "Human" labels.

		Predicted label	
		Nature (Negative)	Human (Positive)
Actual label	Nature (Negative)	A	B
	Human (Positive)	C	D

5.4.6. Transferability and generalization

The generalization ability of the model when in contact with a new and unseen dataset was also evaluated in order to obtain results with reinforced statistical significance. Since the models trained with the ResNet152 architecture had an outstanding performance, with finest results for both of the classifications, especially for the second one, we opted to only test the transferability and generalization of these models. To achieve that, it was considered the same parameters, transfer learning and data augmentation strategies as those used for the tests carried out within the “Peneda-Gerês” dataset. All of the photographs in the “Peneda-Gerês” dataset were used as the training set, while the ones in “Sierra Nevada” dataset were assigned as the test set.

Chapter 6

Results

6.1. Nature vs. Human classification

In this section, we consider the results obtained by the different models in the classification of “Nature” vs. “Human” images. We first consider the case when data augmentation is not applied to the training dataset, followed by the case where data augmentation is applied.

6.1.1. Performance without data augmentation

When comparing the classification accuracy of the two transfer learning scenarios and of the weights obtained by training only over our dataset (Figure 6.1), ImageNet had, overall, a slightly finer performance for the two architectures under study (Accuracy: 83.58 for both of them), followed by Places365 and finally by weights trained only with our dataset. An exception was found for Places365 in VGG16, that resulted in a higher accuracy value (83.75). Also, it was observed that VGG16 had a better performance, in general, when compared to ResNet152, showing higher accuracy values, except for ImageNet (where the two networks had the same accuracy value, 83.58).

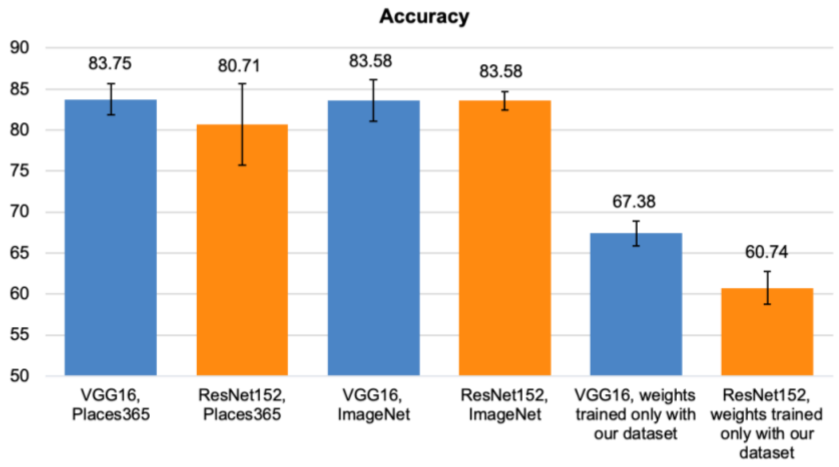


Figure 6.1. Accuracy of the VGG16 and ResNet152 model performance for the two transfer learning scenarios and the weights obtained by training only over our dataset. The bars reported in the plot represent standard deviations.

With regards to sensitivity (Figure 6.2) and taking into account the two transfer learning scenarios as well as the weights obtained by training only over our dataset, it was verified that Places365 had, in general, finer sensitivity results for the two architectures under study (82.7 and 75.69), with the exception of ImageNet for VGG16 (81.05), followed by ImageNet and weights trained only with our dataset. Also, it was observed that VGG16 had better sensitivity results when compared to ResNet152.

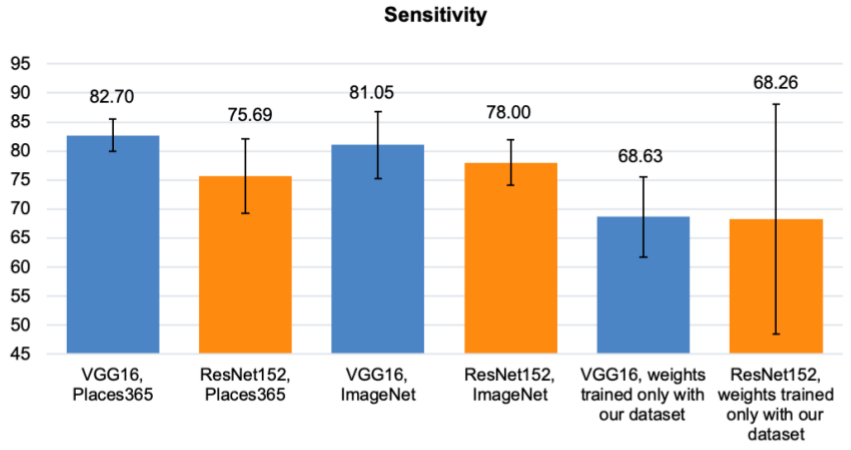


Figure 6.2. Sensitivity of the VGG16 and ResNet152 model performance for the two transfer learning scenarios and the weights obtained by training only over our dataset. The bars reported in the plot represent standard deviations.

For specificity (Figure 6.3), when comparing the two transfer learning scenarios and the weights obtained by training only over our dataset, it was observed that ImageNet had better specificity results for the two architectures under study (86.49 and 89.42), followed by Places365 and by weights trained only with our dataset. Likewise, it was verified that ResNet152 had, overall, finer specificity results, when compared to VGG16, except for the ones considering the weights trained only with our dataset (66.54 for VGG16 and 51.47 for ResNet152).

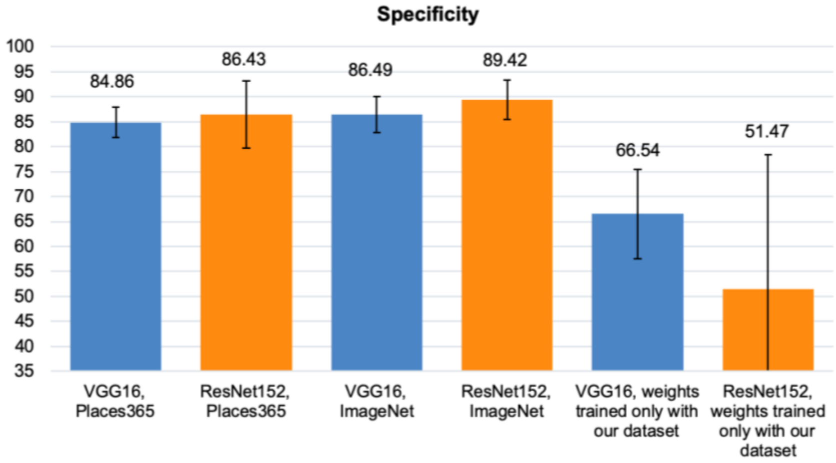


Figure 6.3. Specificity of the VGG16 and ResNet152 model performance for the two transfer learning scenarios and the weights obtained by training only over our dataset. The bars reported in the plot represent standard deviations.

With regard to F1-score (Figure 6.4) and taking into account the two transfer learning scenarios and the weights obtained by training only over our dataset, it was verified that ImageNet had, overall, finer F1-score results for the two architectures under study (83.52 and 83.06), followed by Places365 and weights trained only with our dataset, with the exception of Places365 in VGG16, that resulted in a higher f1-score value (84.00). Similarly, it was observed that VGG16 had better F1-score results when compared to ResNet152.

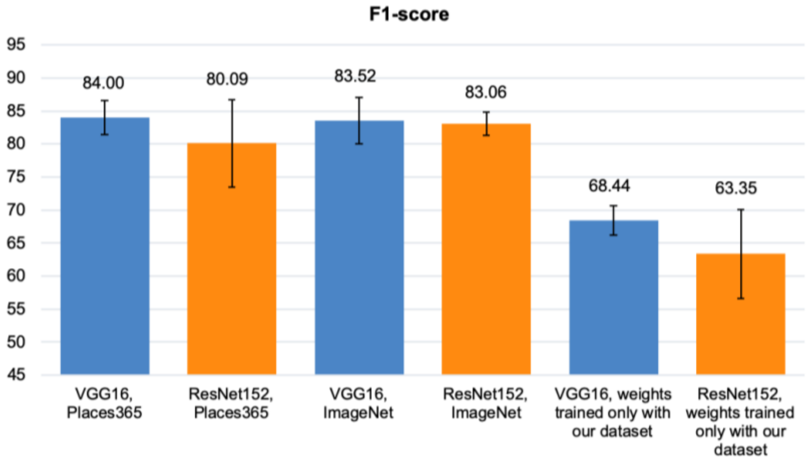


Figure 6.4. F1-score of the VGG16 and ResNet152 model performance for the two transfer learning scenarios and the weights obtained by training only over our dataset. The bars reported in the plot represent standard deviations.

6.1.2. Performance with data augmentation

After the data augmentation process (Figure 6.5) it was observed that, similarly to the verified for the performance without data augmentation, ImageNet had, overall, a higher accuracy for the two architectures under study (86.11 vs 87.18), followed by Places365 and by weights trained only with our dataset, with the exception of Places365 in VGG16, that resulted in an equally high accuracy value (87.01). Also, it was verified that, for Places365, VGG16 had a better performance when compared to ResNet152 (87.01 vs 86.00), while for the remaining scenarios, the ResNet152 model was more accurate than the one for VGG16.

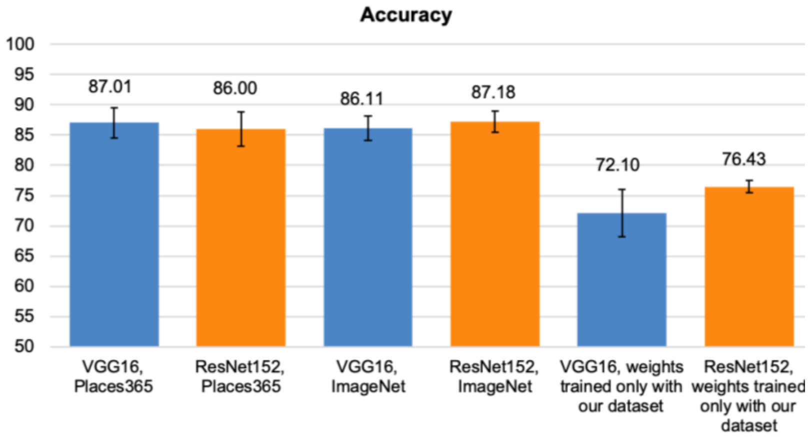


Figure 6.5. Accuracy of the VGG16 and ResNet152 model performance for the two transfer learning scenarios and the weights obtained by training only over our dataset. The bars reported in the plot represent standard deviations.

For sensitivity after augmentation (Figure 6.6), when comparing the two transfer learning scenarios and the weights obtained by training only over our dataset, it was verified that ImageNet had, overall, better sensitivity results for the two architectures under study (86.71 and 86.78), followed by Places365 and weights trained only with our dataset, with the exception of Places365 in VGG16, that resulted in a higher sensitivity value (88.48). Likewise, it was observed that ResNet152 had slightly finer sensitivity results when compared to VGG16, except for Places365, where VGG16 showed the best result (88.48 vs 83.40).

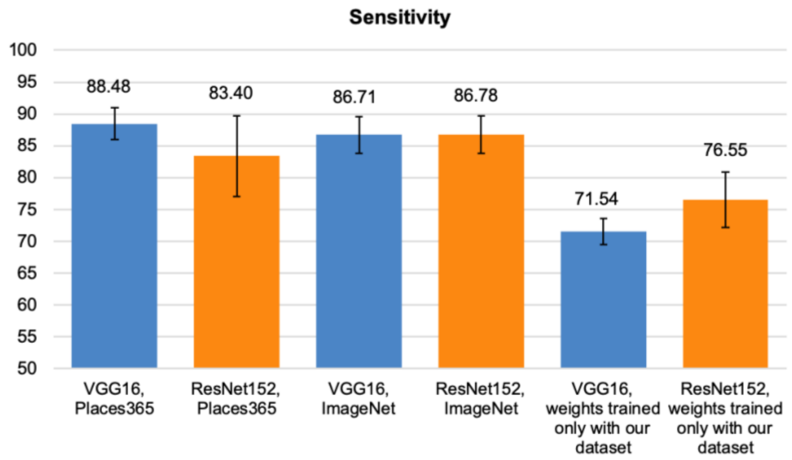


Figure 6.6. Sensitivity of the VGG16 and ResNet152 model performance for the two transfer learning scenarios and the weights obtained by training only over our dataset. The bars reported in the plot represent standard deviations.

With regard to specificity after augmentation (Figure 6.7) and taking into account the two transfer learning scenarios and the weights obtained by training only over our dataset, it was observed that Places365 had finer specificity results for the two architectures under study (85.54 and 88.46), followed by ImageNet and by weights trained only with our dataset. Similarly, it was verified that ResNet152 had better specificity results when compared to VGG16, for all the scenarios under study.

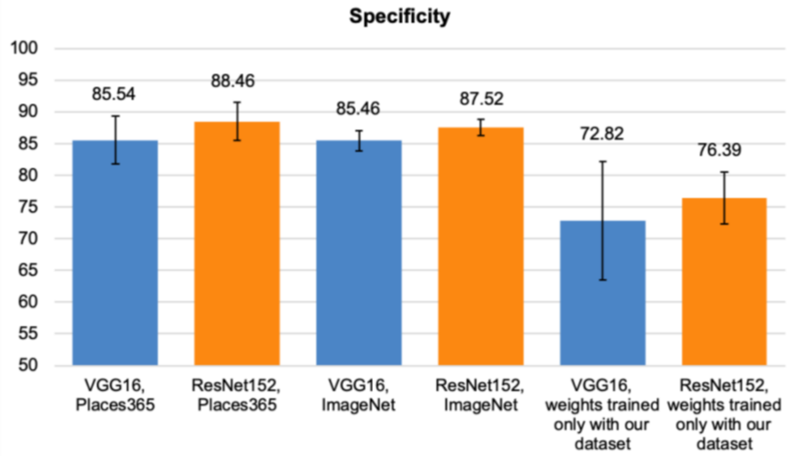


Figure 6.7. Specificity of the VGG16 and ResNet152 model performance for the two transfer learning scenarios and the weights obtained by training only over our dataset. The bars reported in the plot represent standard deviations.

For F1-score after augmentation (Figure 6.8), when comparing the two transfer learning scenarios and the weights obtained by training only over our dataset, it was verified that

ImageNet had slightly better F1-score results for the two architectures under study (86.53 and 87.44), followed by Places365 and weights trained only with our dataset. Also, it was observed that ResNet152 had finer F1-score results when compared to VGG16, except for Places365, where VGG16 showed the best result (87.53 vs 85.89).

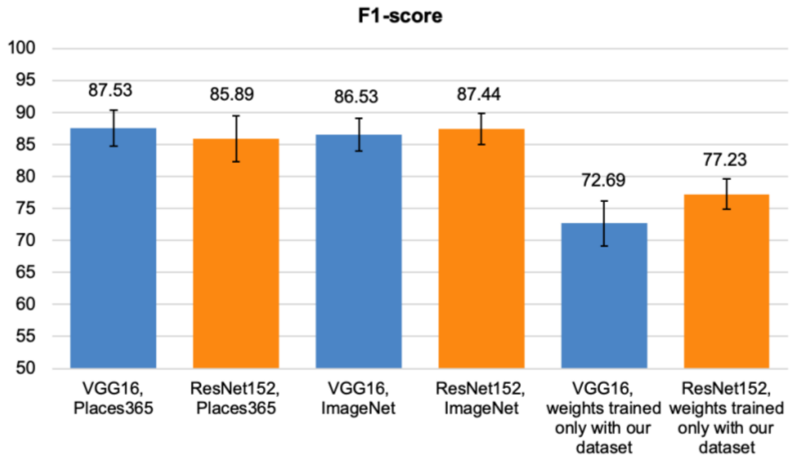


Figure 6.8. F1-score of the VGG16 and ResNet152 model performance for the two transfer learning scenarios and the weights obtained by training only over our dataset. The bars reported in the plot represent standard deviations.

6.2. Multilabel classification

In this section, we consider the results obtained by the different models in the classification of images labelled as "Species", "Landscape", "Nature", "Human activities", "Human structures" and "Posing". We first consider the case when data augmentation is not applied to the training dataset, followed by the case where data augmentation is applied.

6.2.1. Performance without data augmentation

When comparing the two transfer learning scenarios and the weights obtained by training only over our dataset (Figure 6.9), Places365 had, overall, a slightly finer performance for the two architectures under study (69.74 and 76.49), followed by ImageNet and weights trained only with our dataset. Also, it was observed that ResNet152 had a better performance, in general, when compared to VGG16, showing higher accuracy values, except for the weights trained only with our dataset (53.66 vs 48.76).

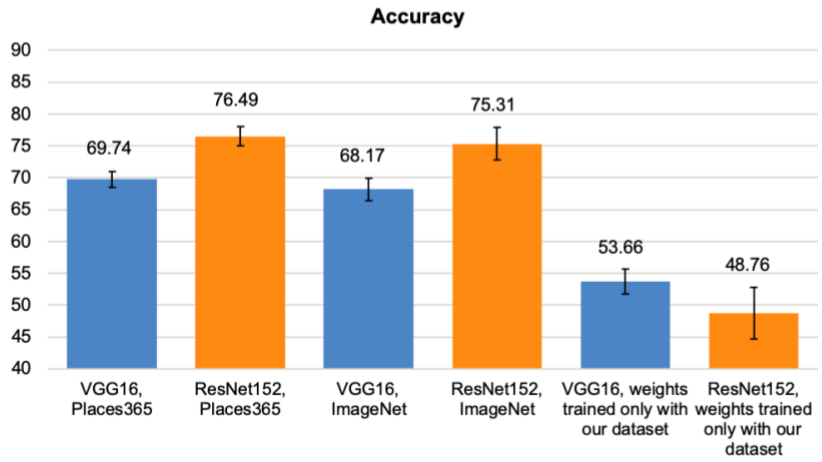


Figure 6.9. Accuracy of the VGG16 and ResNet152 model performance for the two transfer learning scenarios and the weights obtained by training only over our dataset. The bars reported in the plot represent standard deviations.

With regard to sensitivity (Figure 6.10) and taking into account the two transfer learning scenarios and the weights obtained by training only over our dataset, it was verified that Places365 had, in general, finer sensitivity results for the two architectures under study (55.60 and 67.52), followed by ImageNet and by weights trained only with our dataset. Also, it was observed that ResNet152 had better sensitivity results when compared to VGG16, except for the weights trained only with our dataset, that resulted in a higher value (35.68 vs 32.55).

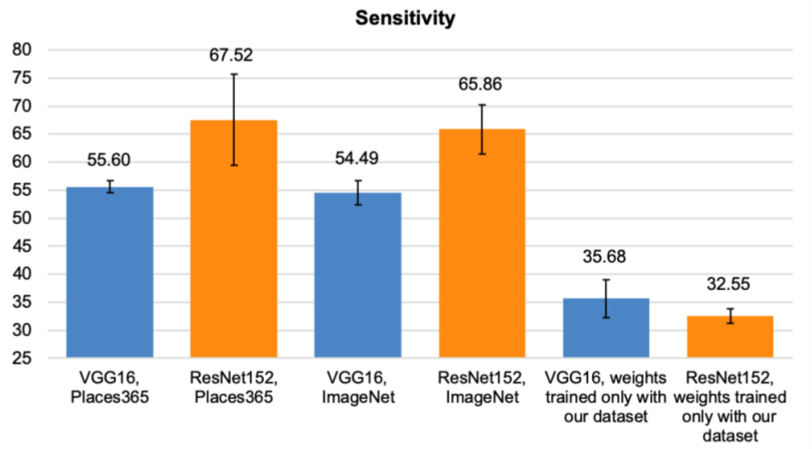


Figure 6.10. Sensitivity of the VGG16 and ResNet152 model performance for the two transfer learning scenarios and the weights obtained by training only over our dataset. The bars reported in the plot represent standard deviations.

For specificity (Figure 6.11), when comparing the two transfer learning scenarios and the weights obtained by training only over our dataset, it was observed that Places365 had higher specificity results for the two architectures under study (93.33 and 94.81), followed by ImageNet and weights trained only with our dataset. Likewise, it was verified that ResNet152 had, overall, finer specificity results, when compared to VGG16, except for the ones considering the weights trained only with our dataset (89.50 for VGG16 and 88.32 for ResNet152).

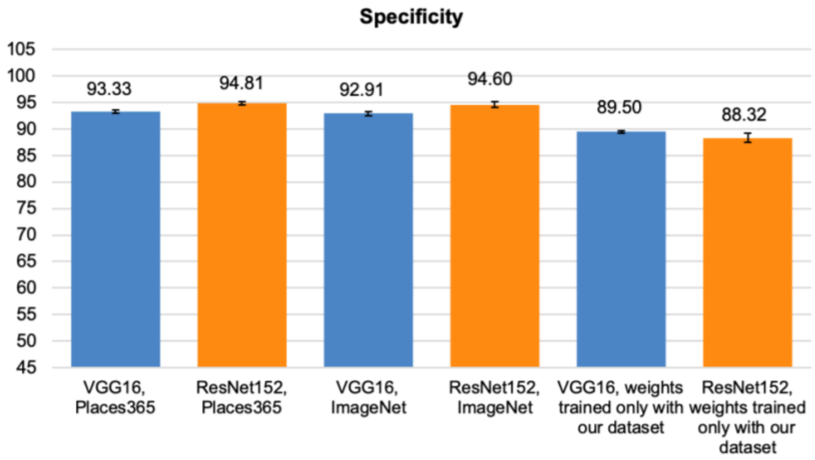


Figure 6.11. Specificity of the VGG16 and ResNet152 model performance for the two transfer learning scenarios and the weights obtained by training only over our dataset. The bars reported in the plot represent standard deviations.

With regard to F1-score (Figure 6.12) and taking into account the two transfer learning scenarios and the weights obtained by training only over our dataset, it was verified that Places365 had, overall, finer F1-score results for the two architectures under study (55.91 and 68.44), followed by ImageNet and by weights trained only with our dataset. Similarly, it was observed that ResNet152 had better F1-score results when compared to VGG16, except for the weights trained only with our dataset (34.27 vs 30.98).

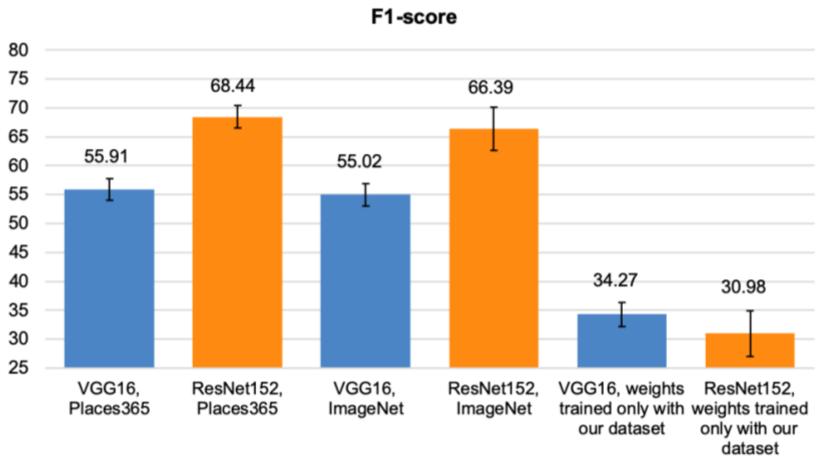


Figure 6.12. F1-score of the VGG16 and ResNet152 model performance for the two transfer learning scenarios and the weights obtained by training only over our dataset. The bars reported in the plot represent standard deviations.

6.2.2. Performance with data augmentation

After the data augmentation process (Figure 6.13) it was observed that, similarly to what was verified for the case without data augmentation, Places365 weights had, in general, a slightly better performance for the two architectures under study (74.07 and 76.89), followed by ImageNet and weights trained only with our dataset. Also, it was verified that ResNet152 had a finer performance when compared to VGG16, for all the scenarios under study.

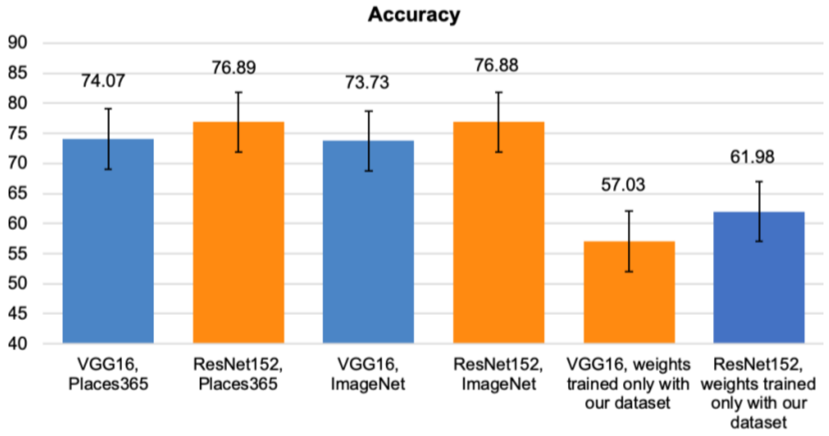


Figure 6.13. Accuracy of the VGG16 and ResNet152 model performance for the two transfer learning scenarios and the weights obtained by training only over our dataset. The bars reported in the plot represent standard deviations.

For sensitivity (Figure 6.14), when comparing the two transfer learning scenarios and the weights obtained by training only over our dataset, it was verified that Places365 and ImageNet had similar sensitivity results for the two architectures under study (64.35 and 67.92 for Places365; 63.85 and 68.09 for ImageNet), followed by weights trained only with our dataset. Similarly, it was observed that ResNet152 had better sensitivity results when compared to VGG16.

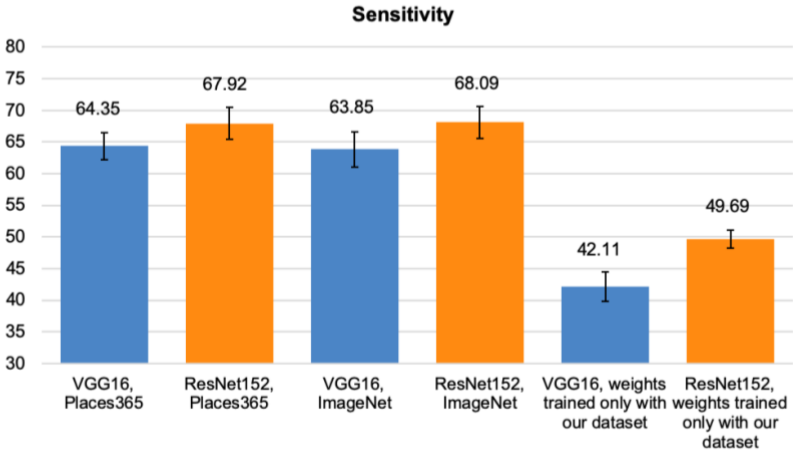


Figure 6.14. Sensitivity of the VGG16 and ResNet152 model performance for the two transfer learning scenarios and the weights obtained by training only over our dataset. The bars reported in the plot represent standard deviations.

With regard to specificity (Figure 6.15) and taking into account the two transfer learning scenarios and the weights obtained by training only over our dataset, it was observed that Places365 and ImageNet had similar specificity results for the two architectures under study (94.30 and 94.90 for Places365; 94.32 and 94.88 for ImageNet), followed by weights trained only with our dataset. Also, it was verified that ResNet152 had finer specificity results when compared to VGG16.

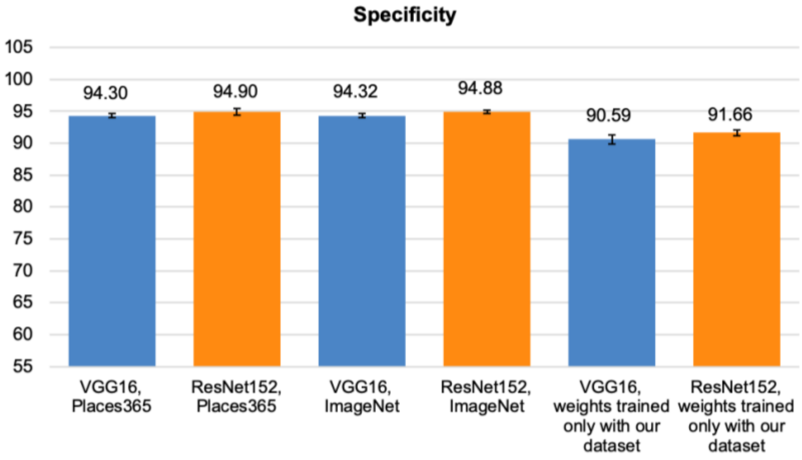


Figure 6.15. Specificity of the VGG16 and ResNet152 model performance for the two transfer learning scenarios and the weights obtained by training only over our dataset. The bars reported in the plot represent standard deviations.

For F1-score (Figure 6.16), when comparing the two transfer learning scenarios and the weights obtained by training only over our dataset, it was verified that Places365 had, overall, slightly finer F1-score results for the two architectures under study (64.84 and 68.92), followed by ImageNet and weights trained only with our dataset. Similarly, it was observed that ResNet152 had better F1-score results when compared to VGG16, for all the scenarios under study.

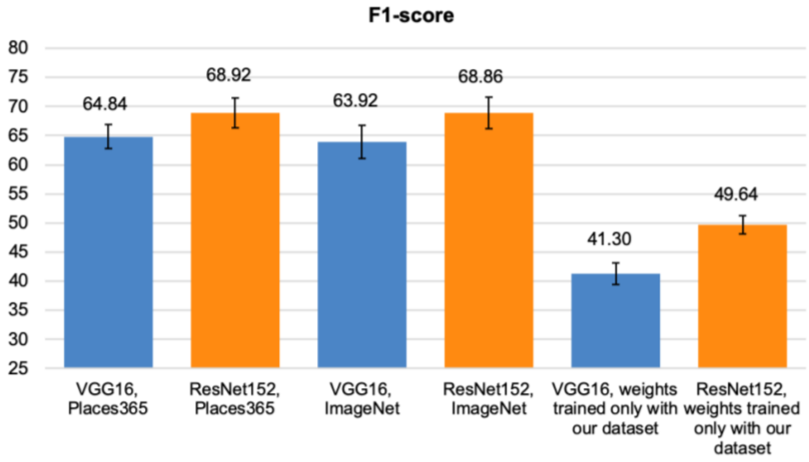


Figure 6.16. F1-score of the VGG16 and ResNet152 model performance for two transfer learning scenarios and the weights obtained by training only over our dataset. The bars reported in the plot represent standard deviations.

6.3. Analysis of the best performing architectures

In this section, we consider the best performing architectures. We first consider the classification of “Nature” vs. “Human” images, followed by the classification of images labelled as “Species”, “Landscape”, “Nature”, “Human activities”, “Human structures” and “Posing”.

6.3.1. Nature vs. Human classification

Regarding the results previously presented (Table 6.1), it was verified that data augmentation actually improved the performance of the transfer learning models developed, since all the evaluation metrics had, overall, higher results after this procedure. Also, for the “Nature” vs. “Human” classification, it was observed that VGG16 and ResNet152 had similar performances for the two transfer learning scenarios and the weights obtained by training only over our dataset.

Table 6.1 Performance metrics of the two transfer learning scenarios and the weights trained only with our dataset for the “Nature” vs. “Human” classification. ACC – Accuracy, TNR – True negative rate/Sensitivity, TPR – True Positive Rate/Specificity, F_1 – f1-score.

	Without Augmentation				With Augmentation			
	ACC	TNR	TPR	F_1	ACC	TNR	TPR	F_1
VGG16, Places365	83.75	82.7	84.86	84	87.01	88.48	85.54	87.53
ResNet152, Places365	80.71	75.69	86.43	80.09	86.00	83.40	88.46	85.89
VGG16, ImageNet	83.58	81.05	86.49	83.52	86.11	86.71	85.46	86.53
ResNet152, ImageNet	83.58	78.00	89.42	83.06	87.18	86.78	87.52	87.44
VGG16, weights trained only with our dataset	67.38	68.63	66.54	68.44	72.10	71.54	72.82	72.69
ResNet152, weights trained only with our dataset	60.74	68.26	51.47	63.35	76.43	76.55	76.39	77.23

These results were achieved by computing the mean of the values in the confusion matrices obtained for each of the 5 folds. When comparing the two transfer learning scenarios and the weights obtained by training only over our dataset, the best setup, with finest and highest results according to the selected evaluation metrics (accuracy, sensitivity, specificity and F1-score), for two (ImageNet and weights trained only with our dataset) out of the three approaches studied, was the ResNet152.

In the following, we show in detail the classification performance achieved by these architectures via the corresponding confusion matrices (Table 6.2, Table 6.3 and Table 6.4).

Table 6.2. Confusion matrix for ResNet152 with Places365 weights. 0 – Nature, Human – 1.

		Predicted Values	
		0	1
Actual Values	0	152	20
	1	30	154

Examples of photographs where the labels were swapped by the algorithm (ResNet152 with Places365 weights) are illustrated in Figure 6.17. The presence of similar elements in the photographs belonging to each class (e.g., sky, vegetation), as well as some colours (e.g., grey, blue, green), could have been the reason for these mismatches in the classification.

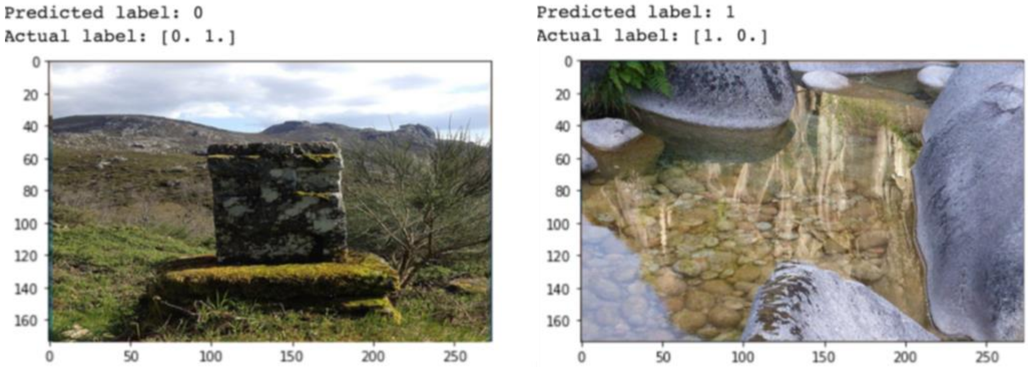


Figure 6.17. Examples of photographs where the labels were swapped by ResNet152 with Places365 weights. 0 – Nature, 1 – Human.

Table 6.3. Confusion matrix for ResNet152 with ImageNet weights. 0 – Nature, Human – 1.

		Predicted Values	
		0	1
Actual Values	0	150	21
	1	24	160

Examples of photographs where the labels were swapped by the algorithm (ResNet152 with ImageNet weights) are displayed in Figure 6.18. Again, as verified for the previous setup, the presence of similar elements in the photographs belonging to each class (e.g., sky, vegetation), as well as some colours (e.g., grey, blue, green), could have been the reason for these mismatches in the classification.

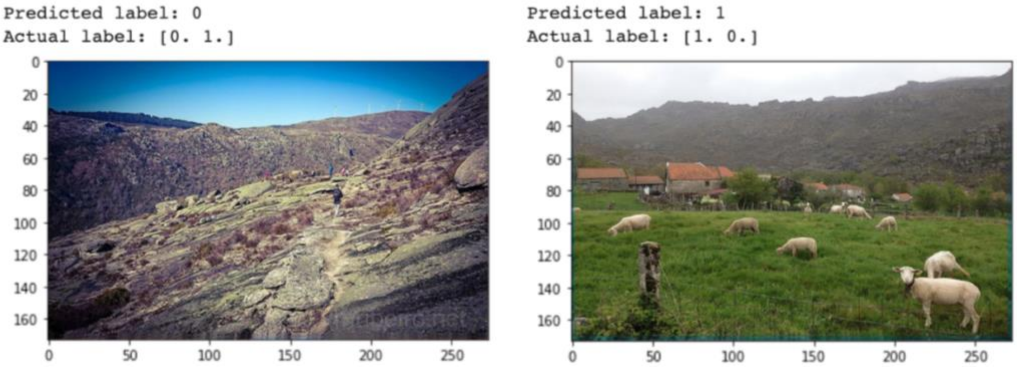


Figure 6.18. Examples of photographs where the labels were swapped by ResNet152 with ImageNet weights. 0 – Nature, 1 – Human.

Table 6.4. Confusion matrix for ResNet152 with the weights trained only with our dataset. 0 – Nature, Human – 1.

		Predicted Values	
		0	1
Actual Values	0	129	40
	1	44	143

Examples of photographs where the labels were swapped by the algorithm (ResNet152 with the weights trained only with our dataset) are illustrated in Figure 6.19. Once again, as verified for the previous setups, the presence of similar elements in the photographs belonging to each class (e.g., sky, vegetation), as well as some colours (e.g., grey, blue, green), could have been the reason for these mismatches in the classification.

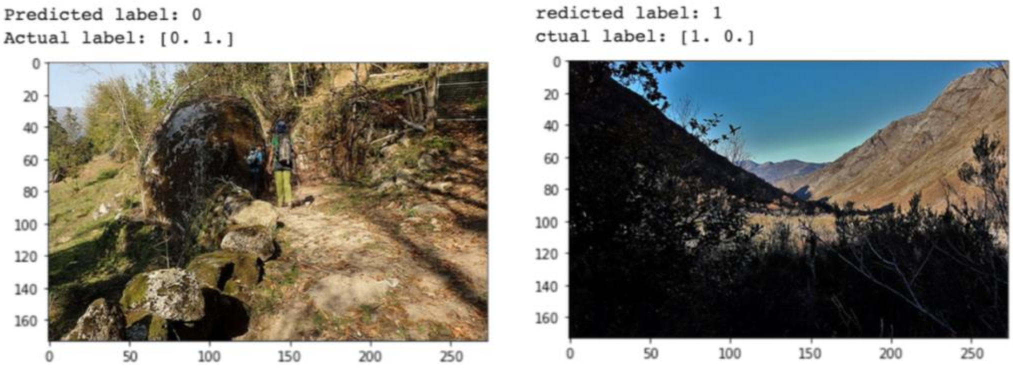


Figure 6.19. Examples of photographs where the labels were swapped by ResNet152 with the weights obtained by training only over our dataset. 0 – Nature, 1 – Human.

6.3.2. Multilabel classification

Regarding the results previously presented (Table 6.5), ResNet152 proved to be the finest and more accurate network architecture for the dataset under study, as well as for the two transfer learning scenarios and the weights obtained by training only over our dataset. Also, for the ResNet152, data augmentation did not increase the performance as much as for the VGG16.

Table 6.5. Performance metrics of the two transfer learning scenarios and the weights trained only with our dataset for the multilabel classification. ACC – Accuracy, TNR – True negative rate/Sensitivity, TPR – True Positive Rate/Specificity, F_1 – f1-score.

	Without Augmentation				With Augmentation			
	ACC	TNR	TPR	F_1	ACC	TNR	TPR	F_1
VGG16, Places365	69.74	55.60	93.33	55.91	74.07	64.35	94.30	64.84
ResNet152, Places365	76.49	67.52	94.81	68.44	76.89	67.92	94.90	68.92
VGG16, ImageNet	68.17	54.49	92.91	55.02	73.73	63.85	94.32	63.92
ResNet152, ImageNet	75.31	65.86	94.60	66.39	76.88	68.09	94.88	68.86

VGG16, weights trained only with our dataset	53.66	35.68	89.50	34.27	57.03	42.11	90.59	41.30
ResNet152, weights trained only with our dataset	48.76	32.55	88.32	30.98	61.98	49.69	91.66	49.64

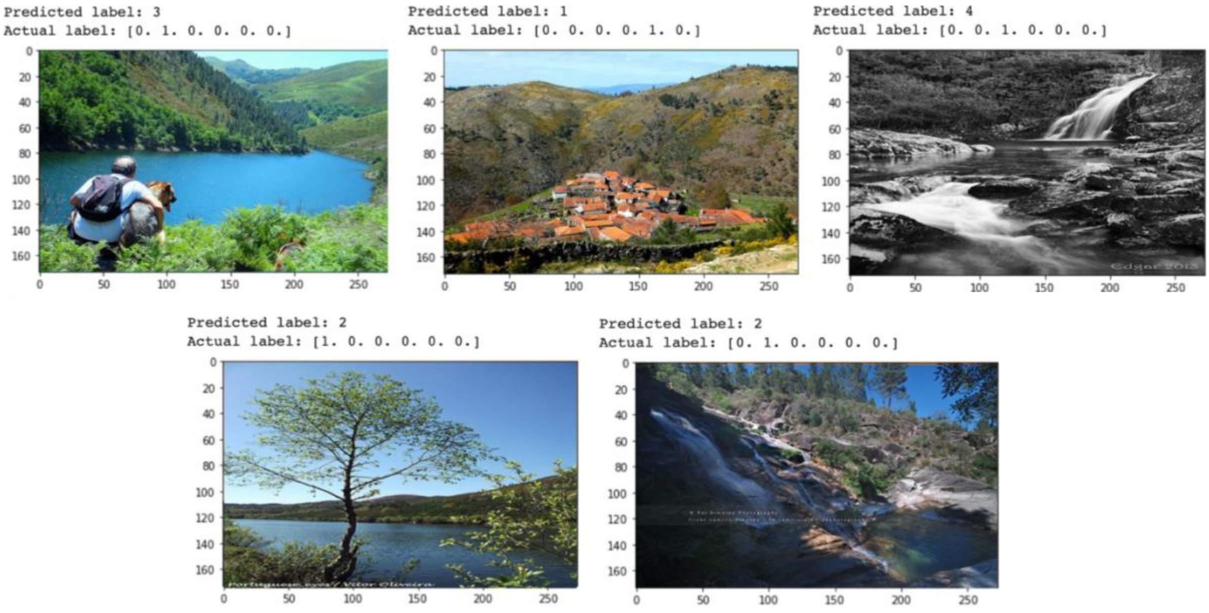
For a multi class problem it is normal that specificity frequently results in a higher value, when compared to the remaining evaluation metrics, as it measures the proportion of factual negatives that are correctly identified as such. Dealing with multiple classes leads to a predominance of the negative class which normally results in the inflation of the specificity values.

In Table 6.6 and Table 6.7, we report the confusion matrices associated with the classification results obtained with the models that performed best over the multilabel classification task, i.e., the ResNet152 architecture with transfer learning from the Places365 and ImageNet dataset, and with data augmentation. Observing Table 6.6, it was verified that “Human activities” with “Landscape”, “Human structures” with “Landscape” and “Nature”, and “Nature” with “Species”, “Landscape” and “Human structures”, were the pairs or sets of classes more indistinguishable to the algorithm, often being confused and swapped by the model.

Table 6.6. Confusion matrix for ResNet152 with Places365 weights. 0 – Species, 1 – Landscape, 2 – Nature, 3 – Human activities, 4 – Human structures, 5 – Posing.

		Predicted Values					
		0	1	2	3	4	5
Actual Values	0	31	3	3	1	4	1
	1	3	105	6	1	10	2
	2	4	6	34	1	3	0
	3	1	6	1	5	2	2
	4	1	9	4	1	84	1
	5	1	2	0	2	2	16

Examples of photographs with labels for which the algorithm (ResNet152 with Places365 weights) most failed the classification according with the confusion matrix are displayed in Figure 6.20. The similarity between some classes (e.g., “Nature”, “Landscape”), that contain common elements (e.g., sky, vegetation) and colours (e.g., blue, green), could have been the reason for the mismatches obtained in the classification.



```
Predicted label: 3
Actual label: [0. 1. 0. 0. 0. 0.]
```

```
Predicted label: 1
Actual label: [0. 0. 0. 0. 1. 0.]
```

```
Predicted label: 4
Actual label: [0. 0. 1. 0. 0. 0.]
```

```
Predicted label: 2
Actual label: [1. 0. 0. 0. 0. 0.]
```

```
Predicted label: 2
Actual label: [0. 1. 0. 0. 0. 0.]
```

Figure 6.20. Examples of photographs with labels for which the algorithm most failed the classification in ResNet152 with Places365 weights. 0 – Species, 1 – Landscape, 2 – Nature, 3 – Human activities, 4 – Human structures, 5 – Posing.

Analysing the Table 6.7, and similarly to the verified for Places365, it was observed that "Human activities" with "Landscape", "Human structures" with "Landscape" and "Nature", and "Nature" with "Species", "Landscape" and "Human structures", were the pairs or sets of classes more indistinguishable to the algorithm, being, because of that, frequently confused and swapped by the model.

Table 6.7. Confusion matrix for ResNet152 with ImageNet weights. 0 – Species, 1 – Landscape, 2 – Nature, 3 – Human activities, 4 – Human structures, 5 – Posing.

		Predicted Values					
		0	1	2	3	4	5
Actual Values	0	30	3	5	1	2	1
	1	2	107	6	1	7	1
	2	1	6	38	0	1	0
	3	1	5	2	4	1	3
	4	1	13	6	1	78	1

5	1	3	1	2	2	16

Examples of photographs with labels for which the algorithm (ResNet152 with ImageNet weights) most failed the classification according with the confusion matrix are illustrated in Figure 6.21. Once again, the similarity between some classes (e.g., “Nature”, “Landscape”), that share common elements (e.g., sky, vegetation) and colours (e.g., blue, green), could have been the cause for the mismatches obtained in the classification.

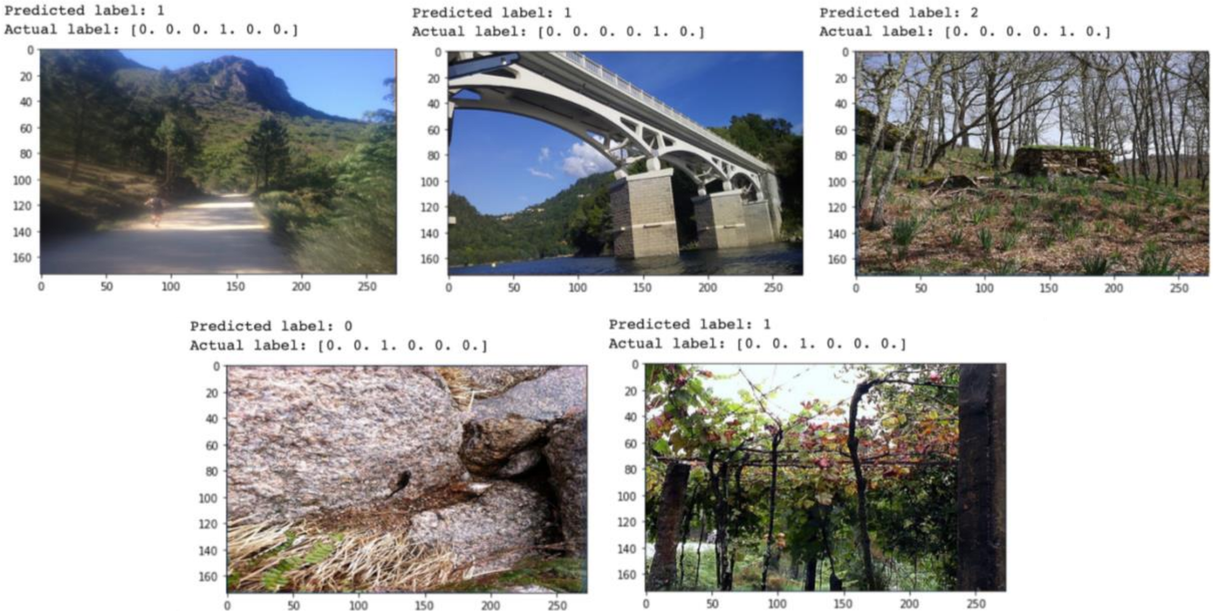

Predicted label: 1
Actual label: [0. 0. 0. 1. 0. 0.]
Predicted label: 1
Actual label: [0. 0. 0. 0. 1. 0.]
Predicted label: 2
Actual label: [0. 0. 0. 0. 1. 0.]
Predicted label: 0
Actual label: [0. 0. 1. 0. 0. 0.]
Predicted label: 1
Actual label: [0. 0. 1. 0. 0. 0.]
0
20
40
60
80
100
120
140
160
0
50
100
150
200
250

Figure 6.21. Examples of photographs with labels for which the algorithm most failed the classification in ResNet152 with ImageNet weights. 0 – Species, 1 – Landscape, 2 – Nature, 3 – Human activities, 4 – Human structures, 5 – Posing.

6.4. Transferability and generalization

This section reports the performance of the models, previously trained over the “Peneda-Gerês” dataset, when applied to images from “Sierra Nevada”, in order to understand the generalization capacity of these models. Only the ResNet152 architecture was implemented in these analyses, as this was the one that revealed the best performance within the “Peneda-Gerês” dataset for both classification tasks considered. We first consider the classification of “Nature” vs. “Human” images, followed by the classification of images labelled as “Species”, “Landscape”, “Nature”, “Human activities”, “Human structures” and “Posing”.

6.4.1. Nature vs. Human classification

When comparing the two transfer learning scenarios considered and weights obtained by training only over our dataset (Figure 6.22), it was observed that ImageNet had a finer performance for ResNet152 (Accuracy: 72.89), followed by Places365 and by the weights trained only with our dataset.

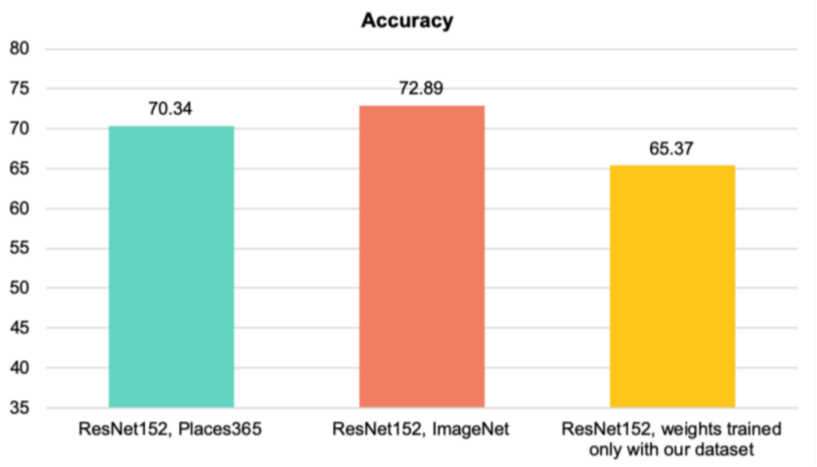


Figure 6.22. Accuracy of the ResNet152 model performance for the two transfer learning scenarios and the weights obtained by training only over our dataset.

With regards to sensitivity (Figure 6.23) and taking into account the two transfer learning scenarios and the weights obtained by training only over our dataset, it was verified that ImageNet had a better sensitivity result (85.36) for ResNet152, followed by Places365 and the weights trained only with our dataset.

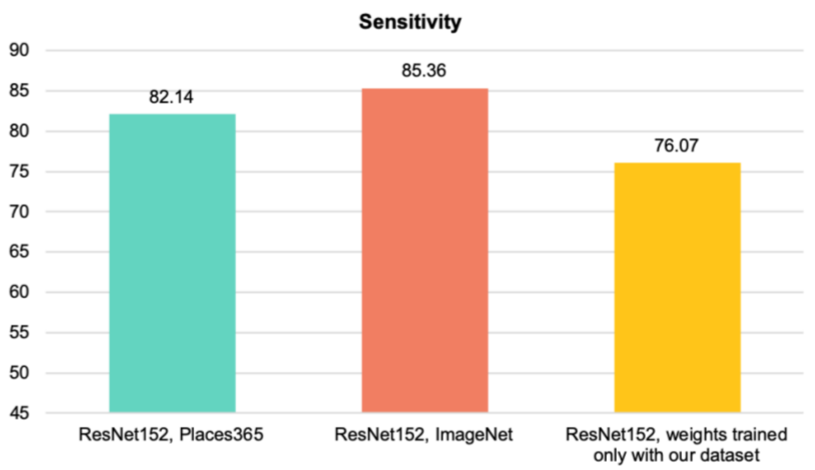


Figure 6.23. Sensitivity of the ResNet152 model performance for the two transfer learning scenarios and the weights obtained by training only over our dataset.

For specificity (Figure 6.24), when comparing the two transfer learning scenarios and the weights obtained by training only over our dataset, it was observed that ImageNet had a better specificity result for ResNet152 (65.38), followed by Places365 and by weights trained only with our dataset.

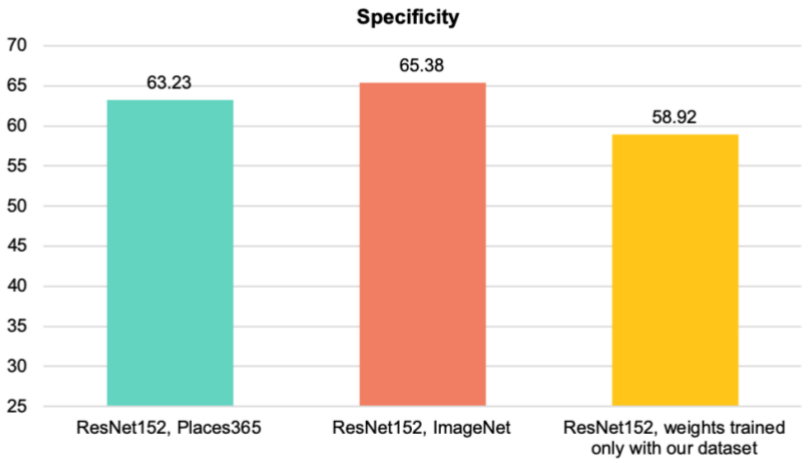


Figure 6.24. Specificity of the ResNet152 model performance for the two transfer learning scenarios and the weights obtained by training only over our dataset.

With regard to F1-score (Figure 6.25) and taking into account the two transfer learning scenarios and the weights obtained by training only over our dataset, it was verified that ImageNet had a finer F1-score result for ResNet152 (70.29), followed by Places365 and the weights trained only with our dataset.

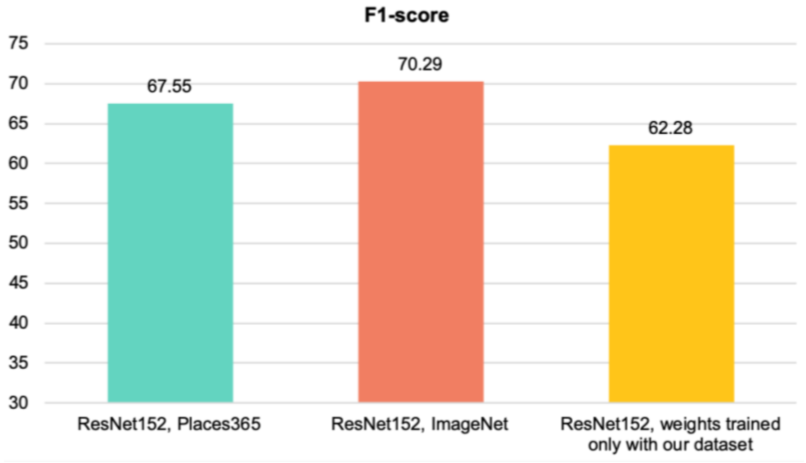


Figure 6.25. F1-score of the ResNet152 model performance for the two transfer learning scenarios and the weights obtained by training only over our dataset.

In this classification task, the performance of the models when applied to a new dataset, was lower than the obtained when working only with data from the “Peneda-Gerês” dataset. Also, when training with the weights trained only with our dataset, the results obtained for “Sierra Nevada” were worse (compared to transfer learning) than those obtained for the “Peneda-Gerês” dataset. Examples of photographs where the labels were swapped by the algorithms (ResNet152 with Places365, ImageNet and the weights trained only with our dataset) are displayed in Figure 6.26. The presence of similar elements in the photographs belonging to each class (e.g., sky, vegetation), as well as some colours (e.g., grey, blue, green), could have been the main reason for these mismatches in the classification.

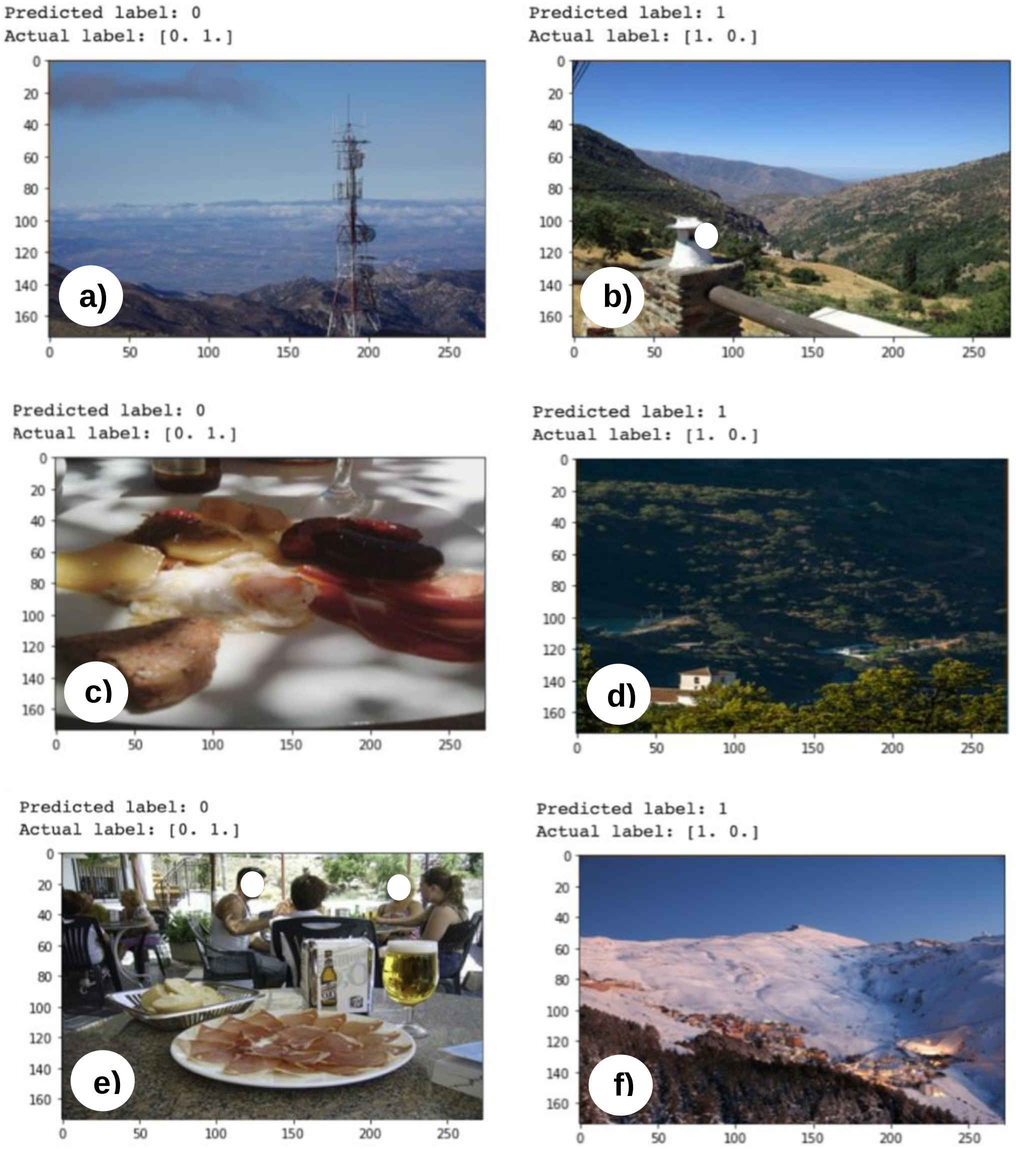


Figure 6.26. Examples of photographs where the labels were swapped by the models. 0 – Nature, 1 – Human, a), b) – ResNet152 with Places365 weights, c), d) – ResNet152 with ImageNet weights, e), f) – ResNet152 with weights trained only with our dataset (the e) picture was edited in order to protect the identity of the persons posing in the picture).

6.4.2. Multilabel classification

When comparing the two transfer learning scenarios and the weights obtained by training only over our dataset (Figure 6.27), Places365 had a slightly finer performance for ResNet152 (Accuracy: 55.03), followed by ImageNet and by weights trained only with our dataset.

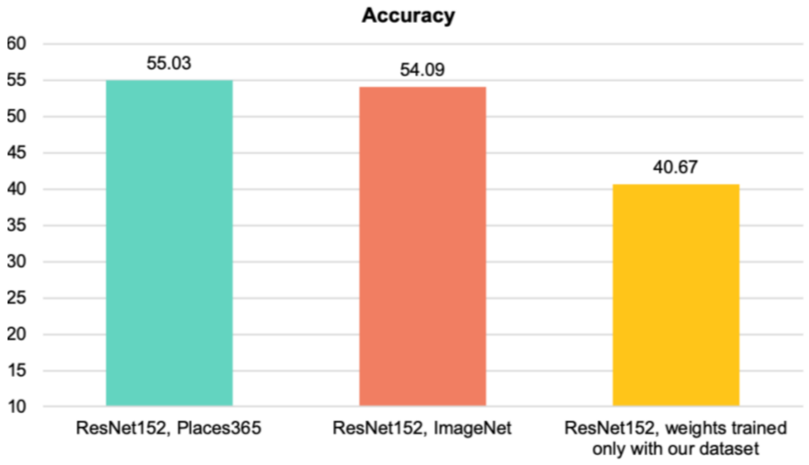


Figure 6.27. Accuracy of the ResNet152 model performance for the two transfer learning scenarios and the weights obtained by training only over our dataset.

With regard to sensitivity (Figure 6.28) and taking into account the two transfer learning scenarios and the weights obtained by training only over our dataset, it was verified that ImageNet had finer sensitivity results for ResNet152 (51.82), followed by Places365 and the weights trained only with our dataset.

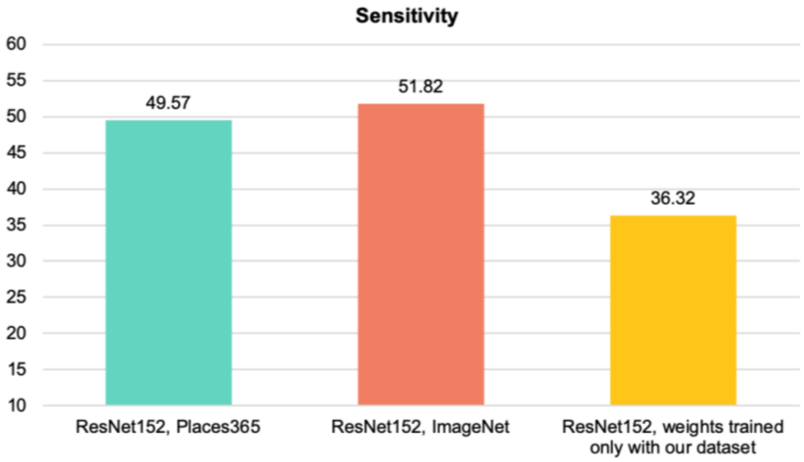


Figure 6.28. Sensitivity of the ResNet152 model performance for the two transfer learning scenarios and the weights obtained by training only over our dataset.

For specificity (Figure 6.29), when comparing the two transfer learning scenarios and the weights obtained by training only over our dataset, it was observed that ImageNet and Places365 had similar specificity results for ResNet152 (90.7 for ImageNet and 90.52 for Places365), followed by weights trained only with our dataset.

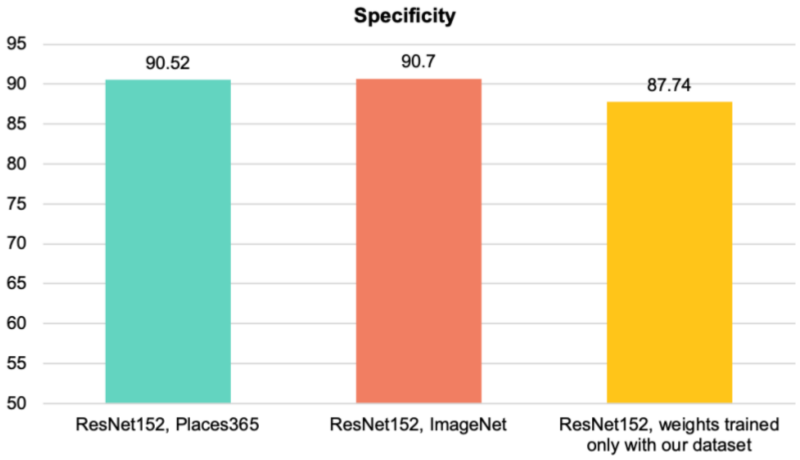


Figure 6.29. Specificity of the ResNet152 model performance for the two transfer learning scenarios and the weights obtained by training only over our dataset.

With regard to F1-score (Figure 6.30) and taking into account the two transfer learning scenarios and the weights obtained by training only over our dataset, it was verified that

ImageNet had finer F1-score results for ResNet152 (50.18), followed by Places365 and by weights trained only with our dataset.

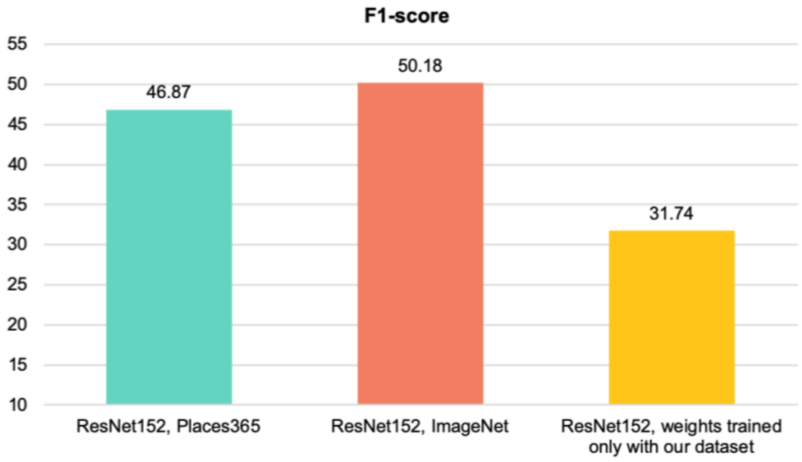


Figure 6.30. F1-score of the ResNet152 model performance for the two transfer learning scenarios and the weights obtained by training only over our dataset.

In Table 6.8, we report the confusion matrix associated with the classification results obtained for the ResNet152 architecture with ImageNet weights, and with data augmentation, since it was the setup that achieved the best performance.

Analysing the Table 6.8, it was observed that "Nature" with "Landscape", "Human activities" with "Landscape", "Human structures" with "Landscape" and "Posing" with "Human activities", were the pairs of classes more indistinguishable to the algorithm, often being confused and swapped by the model. In the "Nature" with "Landscape" pair, the algorithm failed more (45) than it hit the correct label (20).

Table 6.8. Confusion matrix for ResNet152 with ImageNet weights. 0 – Species, 1 – Landscape, 2 – Nature, 3 – Human activities, 4 – Human structures, 5 – Posing.

		Predicted Values					
		0	1	2	3	4	5
Actual Values	0	39	8	9	7	8	3
	1	0	132	3	2	1	0
	2	4	45	20	6	1	1
	3	3	24	3	32	5	18
	4	15	100	8	18	137	9

5	2	7	0	22	10	43

Examples of photographs with labels for which the algorithm (ResNet152 with ImageNet weights) most failed the classification according with the confusion matrix are displayed in Figure 6.31.

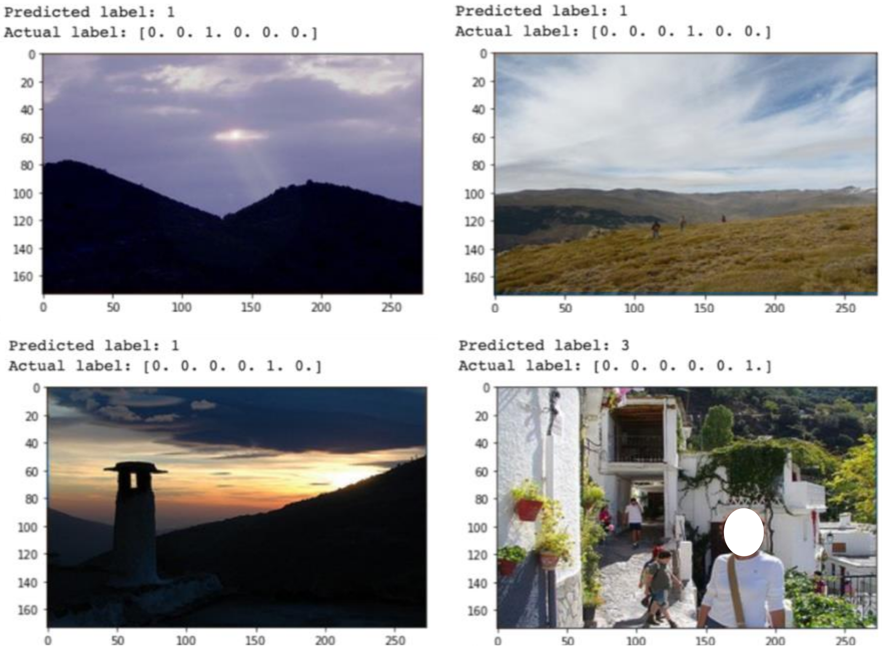

Predicted label: 1
Actual label: [0. 0. 1. 0. 0. 0.]
Predicted label: 1
Actual label: [0. 0. 0. 1. 0. 0.]
Predicted label: 1
Actual label: [0. 0. 0. 0. 1. 0.]
Predicted label: 3
Actual label: [0. 0. 0. 0. 0. 1.]
0
20
40
60
80
100
120
140
160
0
50
100
150
200
250

Figure 6.31. Examples of photographs with labels for which the algorithm most failed the classification in ResNet152 with ImageNet weights. 0 – Species, 1 – Landscape, 2 – Nature, 3 – Human activities, 4 – Human structures, 5 – Posing (the picture was edited in order to protect the identity of the person posing in the picture).

Chapter 7

Discussion and Conclusions

7.1. Nature vs. Human and multilabel classification

The results obtained for both of the classifications were in fact, very satisfactory, with accurate performances for both of the VGG16 and ResNet152 architectures. However, the lack of better results could be assigned to overfitting, since the dataset under analysis is relatively small.

When comparing the two transfer learning scenarios and the weights obtained by training only over our dataset for the “Nature” vs. “Human” classification, as well as for the multilabel classification task, it was expected that the model implemented with the Places365 weights would have a finer performance than the other two (with ImageNet weights and weights trained only with our dataset), since all the photographs contained in this dataset were exclusively related with landscapes and places in general, constituting the database that most resembles our dataset. Perhaps surprisingly, this was not the case for both VGG16 and ResNet152, as ImageNet was the database where the two transfer learning scenarios achieved better results. A possible explanation for this behaviour can reside in the observation that deep learning models achieve more accurate results when trained in the presence of large datasets. In fact, ImageNet, by containing a larger number of photographs (more than 14 million) than Places365 (around 1.8 million), has led to a better performance of the model. Similarly, it was already predictable that the two transfer learning scenarios would lead to more accurate results than the weights trained only with our dataset, for the same reason mentioned above: limitations in the size of the dataset.

In the multilabel classification task, since we were working with multiple classes, it was already expected to obtain less accurate results than the ones obtained for the “Nature” vs. “Human” binary classification. The greater the number of output nodes, the higher the complexity of the model, and the lower the effectiveness and reliability of the results. Many factors can influence these results, however the similarity between certain classes (see section 6.3.2) of the photographs (e.g., “Landscape” and “Nature”), that contain several elements in common (e.g., sky, sea, vegetation), constitutes one of the most significant.

Also, as mentioned above, for both classification tasks, data augmentation did not increase the performance of the models as much for the VGG16 as it did for the ResNet152. This suggests that the ResNet152 (through the transfer learning) was able to cope better with overfitting with respect to the VGG16, even without requiring data augmentation, which might be associated with the presence of residual connections in this architecture.

7.2. Transferability and generalization

Regarding the transferability and generalization capacity of the ResNet152 models, when in the presence of a completely new and unseen dataset, it was desirable to have similar results to those obtained for the “Peneda-Gerês” dataset. These results turned out to be less promising as initially expected, presenting lower performance rates, with accuracies values varying between 41% and 73%, depending on the classification task. A likely interpretation for this behaviour can be related to the features/elements of the photographs in both of the “Peneda-Gerês” and “Sierra Nevada” datasets, which are very distinct. For instance, in “Sierra Nevada”, cold and neutral colours, such as white, grey and blue, predominate, while in “Peneda-Gerês”, warm and cold colours, like green, blue and brown, are the most common.

Also, as mentioned above, when considering the weights trained only with our dataset in the “Nature” vs. “Human” classification, the results were worse (compared to transfer learning) than those obtained for the “Peneda-Gerês” dataset. This means that the use of transfer learning allows for better generalization from one dataset to the other and that the overfitting incurred when training over only our dataset has an even more significant impact when trying to apply the model to new dataset, from a different region.

7.3. Limitations and future work

Interpreting and understanding the information contained in social media photographs can be quite challenging in the context of cultural ecosystem services (CES). Choosing whether to take/share or not a photo is subjective to the user’s interest, which makes the motive of taking the photograph unclear when unprovided of contextual information. People can take photographs in the environment to document positive and appealing features, as well as negative and unattractive ones. Moreover, people can take a photograph to just eternalize a moment or memory.

All of these questions make the assessment of CES trough social media photographs very complex, especially when trying to assess a specific CES value (e.g., a picture can represent aesthetic or other value, without absolute certainty of the correct value) (Dorwart *et al*., 2009). Thus, additional information on the context of the photograph (e.g. user description, tags or title of the photo) may be required to improve the results from image classification in the scope of CES (Moreno-Llorca *et al*. 2020a). This information can be collected, for instance, through natural language processing and sentiment analysis techniques (Gosal *et al*., 2019; Do, 2019).

Also, the classes selected in this study to classify each picture may not be mutually exclusive, since the same picture can express multiple cultural benefits from nature. For example, in the “Posing” class, people can take pictures because they find beauty in a landscape or species, they want to do an artistic picture, or simply because they want to take photos of themselves for future memory. Similarly, social media photographs are not always representative data for assessing CES, since a fraction of the people who visit a given area may not take pictures of that area.

Nevertheless, our approach and results are a first step to show that deep learning methods can offer significant contributions to assist in CES evaluation. Future work needs to encompass the use of more representative training dataset, with the possibility of generating synthetically new training images, as well as focus on the improvement of the robustness of these models against scarcely labelled data and overfitting. This could be achieved via the use of regularization methods and semi-supervised approaches by leveraging autoencoder architectures as well as generative adversarial networks (GANs).

Bibliography

Aloufi, S., Zhu, S., & El Saddik, A. (2017). On the prediction of flickr image popularity by analyzing heterogeneous social sensory data. Sensors, 17(3), 631.

Arts, K., van der Wal, R., & Adams, W. M. (2015). Digital technology and the conservation of nature. *Ambio*, *44*(4), 661-673.

Assessment, M. E. (2005). Ecosystems and human well-being (Vol. 5). Washington, DC: Island press.

Batrinca, B., & Treleaven, P. C. (2015). Social media analytics: a survey of techniques, tools and platforms. *Ai & Society*, *30*(1), 89-116.

Blicharska, M., Smithers, R. J., Hedblom, M., Hedenås, H., Mikusiński, G., Pedersen, E., ... & Svensson, J. (2017). Shades of grey challenge practical application of the cultural ecosystem services concept. *Ecosystem services*, *23*, 55-70.

Brown, G., Pullar, D., & Hausner, V. H. (2016). An empirical evaluation of spatial value transfer methods for identifying cultural ecosystem services. *Ecological indicators*, *69*, 1-11.

Brownlee, J. (2016). *Deep learning with Python: develop deep learning models on Theano and TensorFlow using Keras*. Machine Learning Mastery.

Bryce, R., Irvine, K. N., Church, A., Fish, R., Ranger, S., & Kenter, J. O. (2016). Subjective well-being indicators for large-scale assessment of cultural ecosystem services. *Ecosystem Services*, *21*, 258-269.

Castillo, A., Tabik, S., Pérez, F., Olmos, R., & Herrera, F. (2019). Brightness guided preprocessing for automatic cold steel weapon detection in surveillance videos with deep learning. *Neurocomputing*, *330*, 151-161.

Cheng, X., Van Damme, S., Li, L., & Uyttenhove, P. (2019). Evaluation of cultural ecosystem services: A review of methods. *Ecosystem services*, *37*, 100925.

Chum, L., Subramanian, A., Balasubramanian, V. N., & Jawahar, C. V. (2019). Beyond supervised learning: A computer vision perspective. *Journal of the Indian Institute of Science*, 1-23.

Di Minin, E., Fink, C., Tenkanen, H., & Hiippala, T. (2018). Machine learning for tracking illegal wildlife trade on social media. *Nature ecology & evolution*, *2*(3), 406-407.

Di Minin, E., Tenkanen, H., & Toivonen, T. (2015). Prospects and challenges for social media data in conservation science. Frontiers in Environmental Science, 3, 63.

Do, Y. (2019). Valuating aesthetic benefits of cultural ecosystem services using conservation culturomics. *Ecosystem Services*, *36*, 100894.

Dorwart, C. E., Moore, R. L., & Leung, Y. F. (2009). Visitors' perceptions of a trail environment and effects on experiences: A model for nature-based recreation experiences. *Leisure Sciences*, *32*(1), 33-54.

Dou, Y., Zhen, L., De Groot, R., Du, B., & Yu, X. (2017). Assessing the importance of cultural ecosystem services in urban areas of Beijing municipality. *Ecosystem Services*, *24*, 79-90.

Dylewski, Ł., Mikula, P., Tryjanowski, P., Morelli, F., & Yosef, R. (2017). Social media and scientific research are complementary—YouTube and shrikes as a case study. *The Science of Nature*, *104*(5-6), 48.

Eid, E., & Handal, R. (2018). Illegal hunting in Jordan: using social media to assess impacts on wildlife. *Oryx*, *52*(4), 730-735.

Ferreira, A. C., Silva, L. R., Renna, F., Brandl, H. B., Renoult, J. P., Farine, D. R., ... & Doutrelant, C. (2019). Deep learning-based methods for individual recognition in small birds. *bioRxiv*, 862557.

Fish, R., Church, A., & Winter, M. (2016a). Conceptualising cultural ecosystem services: a novel framework for research and critical engagement. *Ecosystem Services*, *21*, 208-217.

Fish, R., Church, A., Willis, C., Winter, M., Tratalos, J. A., Haines-Young, R., & Potschin, M. (2016b). Making space for cultural ecosystem services: Insights from a study of the UK nature improvement initiative. *Ecosystem Services*, *21*, 329-343.

Fisher, D. M., Wood, S. A., White, E. M., Blahna, D. J., Lange, S., Weinberg, A., ... & Lia, E. (2018). Recreational use in dispersed public lands measured using social media data and on-site counts. *Journal of environmental management*, *222*, 465-474.

Frank, M. R., Williams, J. R., Mitchell, L., Bagrow, J. P., Dodds, P. S., & Danforth, C. M. (2014). Constructing a taxonomy of fine-grained human movement and activity motifs through social media. *arXiv preprint arXiv:1410.1393*.

Fu, J., & Rui, Y. (2017). Advances in deep learning approaches for image tagging. *APSIPA Transactions on Signal and Information Processing*, *6*.

Gliozzo, G., Pettorelli, N., & Haklay, M. (2016). Using crowdsourced imagery to detect cultural ecosystem services: a case study in South Wales, UK. *Ecology and Society*, *21*(3).

Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep learning. MIT press.

Goodness, J., Andersson, E., Anderson, P. M., & Elmqvist, T. (2016). Exploring the links between functional traits and cultural ecosystem services to enhance urban ecosystem management. *Ecological Indicators*, *70*, 597-605.

Gosal, A. S., & Ziv, G. (2020). Landscape aesthetics: Spatial modelling and mapping using social media images and machine learning. *Ecological Indicators*, *117*, 106638.

Gosal, A. S., Geijzendorffer, I. R., Václavík, T., Poulin, B., & Ziv, G. (2019). Using social media, machine learning and natural language processing to map multiple recreational beneficiaries. Ecosystem Services, 38, 100958.

Hafemann, L. G., Oliveira, L. S., & Cavalin, P. (2014, August). Forest species recognition using deep convolutional neural networks. In *2014 22nd International Conference on Pattern Recognition* (pp. 1103-1107). IEEE.

Hausmann, A., Toivonen, T., Fink, C., *et al.* (2020). Understanding sentiment of national park visitors from social media data. *People Nat*, 00: 1– 11.

Hausmann, A., Toivonen, T., Slotow, R., Tenkanen, H., Moilanen, A., Heikinheimo, V., & Di Minin, E. (2018). Social media data can be used to understand tourists' preferences for nature-based experiences in protected areas. *Conservation Letters*, *11*(1), e12343.

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

Heikinheimo, V., Tenkanen, H., Hiippala, T., & Toivonen, T. (2018, October). Digital imaginations of national parks in different social media: a data exploration. In *On the Way to Platial Analysis: Can Geosocial Media Provide the Necessary Impetus? Proceedings of the First Workshop on Platial Analysis*. Heidelberg university.

Hirons, M., Comberti, C., & Dunford, R. (2016). Valuing cultural ecosystem services. *Annual Review of Environment and Resources*, *41*, 545-574.

Jabbar, H., & Khan, R. Z. (2015). Methods to avoid over-fitting and under-fitting in supervised machine learning (comparative study). *Computer Science, Communication and Instrumentation Devices*, 163-172.

Jarić, I., Correia, R. A., Brook, B. W., Buettel, J. C., Courchamp, F., Di Minin, E., ... & Ladle, R. (2020). iEcology: Harnessing Large Online Resources to Generate Ecological Insights. *Trends in Ecology & Evolution*.

Jeong, W. K., Pfister, H., & Fatica, M. (2011). Medical image processing using GPU-accelerated ITK image filters. In *GPU Computing Gems Emerald Edition* (pp. 737-749). Morgan Kaufmann.

Kanai, S., Fujiwara, Y., Yamanaka, Y., & Adachi, S. (2018). Sigsoftmax: Reanalysis of the softmax bottleneck. In *Advances in Neural Information Processing Systems* (pp. 286-296).

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

Klette, R. (2014). *Concise computer vision*. Springer, London.

Koylu, C., Zhao, C., & Shao, W. (2019). Deep neural networks and kernel density estimation for detecting human activity patterns from Geo-tagged images: A case study of birdwatching on Flickr. ISPRS International Journal of Geo-Information, 8(1), 45.

Ladle, R. J., Correia, R. A., Do, Y., Joo, G. J., Malhado, A. C., Proulx, R., ... & Jepson, P. (2016). Conservation culturomics. *Frontiers in Ecology and the Environment*, *14*(5), 269-275.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *nature*, *521*(7553), 436-444.

Lee, H., Seo, B., Koellner, T., & Lautenbach, S. (2019). Mapping cultural ecosystem services 2.0–Potential and shortcomings from unlabeled crowd sourced images. *Ecological Indicators*, *96*, 505-515.

Liu, W., Wen, Y., Yu, Z., & Yang, M. (2016, June). Large-margin softmax loss for convolutional neural networks. In *ICML* (Vol. 2, No. 3, p. 7).

Lunstrum, E. (2017). Feed them to the lions: Conservation violence goes online. *Geoforum*, *79*, 134-143.

Lusch, B., Kutz, J. N., & Brunton, S. L. (2018). Deep learning for universal linear embeddings of nonlinear dynamics. Nature communications, 9(1), 4950.

Martínez-López, J., Bagstad, K. J., Balbi, S., Magrach, A., Voigt, B., Athanasiadis, I., ... & Villa, F. (2019). Towards globally customizable ecosystem service models. *Science of the Total Environment*, *650*, 2325-2336.

Maxwell, S. L., Fuller, R. A., Brooks, T. M., & Watson, J. E. (2017). Biodiversity: The ravages of guns, nets and bulldozers (Comment).

McCay-Peet, L., & Quan-Haase, A. (2017). What is social media and what questions can social media research help us answer. *The SAGE handbook of social media research methods*, 13-26.

Mettes, P., Koelma, D. C., & Snoek, C. G. (2016, June). The imagenet shuffle: Reorganized pre-training for video event detection. In *Proceedings of the 2016 ACM on International Conference on Multimedia Retrieval* (pp. 175-182).

Moreno-Llorca R., Vaz A. S., Herrero J., Millares A., Bonet-García F.J., Alcaraz-Segura D. (2020a). Multi-scale evolution of ecosystem service provision in Sierra Nevada (Spain): an assessment over the last half-century. Ecosystem Services.

Moreno-Llorca, R., Méndez, P. F., Ros-Candeira, A., Alcaraz-Segura, D., Santamaría, L., Ramos-Ridao, Á. F., ... & Vaz, A. S. (2020b). Evaluating tourist profiles and nature-based experiences in Biosphere Reserves using Flickr: Matches and mismatches between online social surveys and photo content analysis. *Science of The Total Environment*, 140067.

Mulfari, D., Celesti, A., Fazio, M., Villari, M., & Puliafito, A. (2016, June). Using Google Cloud Vision in assistive technology scenarios. In *2016 IEEE Symposium on Computers and Communication (ISCC)* (pp. 214-219). IEEE.

Najafabadi, M. M., Villanustre, F., Khoshgoftaar, T. M., Seliya, N., Wald, R., & Muharemagic, E. (2015). Deep learning applications and challenges in big data analytics. *Journal of Big Data*, *2*(1), 1.

Nash, W., Drummond, T., & Birbilis, N. (2018). A review of deep learning in the study of materials degradation. *npj Materials Degradation*, *2*(1), 1-12.

Nielsen, M. A. (2015). *Neural networks and deep learning* (Vol. 2018). San Francisco, CA: Determination press.

Nielsen, M., Bengio, Y., & Couville, A. (2017). Deep learning. Retrieved from http://neuralnetworksanddeeplearning.

Pathak, N., Henry, M. J., & Volkova, S. (2017, March). Understanding Social Media's Take on Climate Change through Large-Scale Analysis of Targeted Opinions and Emotions. In *2017 AAAI Spring Symposium Series*.

Plieninger, T., Bieling, C., Fagerholm, N., Byg, A., Hartel, T., Hurley, P., ... & van der Horst, D. (2015). The role of cultural ecosystem services in landscape management and planning. *Current Opinion in Environmental Sustainability*, *14*, 28-33.

Powers, D. M. (2011). Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation.

Rawat, W., & Wang, Z. (2017). Deep convolutional neural networks for image classification: A comprehensive review. *Neural computation*, *29*(9), 2352-2449.

Renna, F., Oliveira, J. H., & Coimbra, M. T. (2019). Deep Convolutional Neural Networks for Heart Sound Segmentation. *IEEE journal of biomedical and health informatics*.

Richards, D. R., & Friess, D. A. (2015). A rapid indicator of cultural ecosystem service usage at a fine spatial scale: content analysis of social media photographs. *Ecological Indicators*, *53*, 187-195.

Richards, D. R., & Tunçer, B. (2018). Using image recognition to automate assessment of cultural ecosystem services from social media photographs. Ecosystem services, 31, 318-325.

Riechers, M., Barkmann, J., & Tscharntke, T. (2016). Perceptions of cultural ecosystem services from urban green. *Ecosystem Services*, *17*, 33-39.

Ros-Candeira, A., Moreno Llorca, R., Alcaraz Segura, D., Bonet García, F. J., & Vaz, A. S. (2020). Social media photo content for Sierra Nevada: a dataset to support the assessment of cultural ecosystem services in protected areas.

Rosenblatt, F. (1958). The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological review*, *65*(6), 386.

Ruder, S. (2016). An overview of gradient descent optimization algorithms. *arXiv preprint arXiv:1609.04747*.

Rumelhart, D. E., Hinton, G. E., and Williams, R. J. (1986). Learning internal representations by error propagation. In Parallel Distributed Processing: Explorations in the Microstructure of Cognition, Volume 1: Foundations. MIT Press.

Salman, A., Jalal, A., Shafait, F., Mian, A., Shortis, M., Seager, J., & Harvey, E. (2016). Fish species classification in unconstrained underwater environments based on deep learning. *Limnology and Oceanography: Methods*, *14*(9), 570-585.

Santarem, F., Silva, R., & Santos, P. (2015). Assessing ecotourism potential of hiking trails: A framework to incorporate ecological and cultural features and seasonality. Tourism Management Perspectives, 16, 190-206.

Sathya, R., & Abraham, A. (2013). Comparison of supervised and unsupervised learning algorithms for pattern classification. *International Journal of Advanced Research in Artificial Intelligence*, *2*(2), 34-38.

Schmidt, K., Walz, A., Jones, I., & Metzger, M. J. (2016). The sociocultural value of upland regions in the vicinity of cities in comparison with urban green spaces. *Mountain Research and Development*, *36*(4), 465-474.

Seresinhe, C. I., Preis, T., & Moat, H. S. (2017). Using deep learning to quantify the beauty of outdoor places. *Royal Society open science*, *4*(7), 170170.

Sherren, K., Smit, M., Holmlund, M., Parkins, J. R., & Chen, Y. (2017). Conservation culturomics should include images and a wider range of scholars. *Frontiers in Ecology and the Environment*, *15*(6), 289-290.

Silva, B. P. C., Silva, M. L. N., Avalos, F. A. P., de Menezes, M. D., & Curi, N. (2019). Digital soil mapping including additional point sampling in Posses ecosystem services pilot watershed, southeastern Brazil. *Scientific reports*, *9*(1), 1-12.

Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Sladojevic, S., Arsenovic, M., Anderla, A., Culibrk, D., & Stefanovic, D. (2016). Deep neural networks based recognition of plant diseases by leaf image classification. Computational intelligence and neuroscience, 2016.

Szeliski, R. (2010). *Computer vision: algorithms and applications*. Springer Science & Business Media.

Tharwat, A. (2020). Classification assessment methods. *Applied Computing and Informatics*.

Tenkanen, H., Di Minin, E., Heikinheimo, V., Hausmann, A., Herbst, M., Kajala, L., & Toivonen, T. (2017). Instagram, Flickr, or Twitter: Assessing the usability of social media data for visitor monitoring in protected areas. *Scientific reports*, *7*(1), 1-11.

Toivonen, T., Heikinheimo, V., Fink, C., Hausmann, A., Hiippala, T., Järv, O., ... & Di Minin, E. (2019). Social media data for conservation science: a methodological overview. *Biological Conservation*, *233*, 298-315.

van der Wal, R., & Arts, K. (2015). Digital conservation: An introduction. Ambio, 44(4), 517-521.

Vaz, A. S., Gonçalves, J. F., Pereira, P., Santarém, F., Vicente, J. R., & Honrado, J. P. (2019). Earth observation and social media: Evaluating the spatiotemporal contribution of non-native trees to cultural ecosystem services.*Remote Sensing of Environment*, *230*, 111193.

Vaz, A.S., Moreno-Llorca, R. Gonçalves, J.F., Vicente, J.R., Méndez, P.F., Revilla, E., Santamaria, L. Bonet-García, F.J., Honrado, J.H., Alcaraz-Segura, D. (2020) Digital conservation in biosphere reserves: Earth observations, social media, and nature's cultural contributions to people. Conservation Letters; e12704. https://doi.org/10.1111/conl.12704 (in press).

Villa, F. (2009). Semantically driven meta-modelling: automating model construction in an environmental decision support system for the assessment of ecosystem services flows. In *Information Technologies in Environmental Engineering* (pp. 23-36). Springer, Berlin, Heidelberg.

Vinuesa, R., Azizpour, H., Leite, I., Balaam, M., Dignum, V., Domisch, S., ... & Nerini, F. F. (2020). The role of artificial intelligence in achieving the Sustainable Development Goals. *Nature Communications*, *11*(1), 1-10.

Wang, J., Korayem, M., Blanco, S., & Crandall, D. J. (2016, October). Tracking natural events through social media and computer vision. In *Proceedings of the 24th ACM international conference on Multimedia* (pp. 1097-1101). ACM.

Weiss, K., Khoshgoftaar, T. M., & Wang, D. (2016). A survey of transfer learning. *Journal of Big data*, *3*(1), 9.

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

Willcock, S., Martínez-López, J., Hooftman, D. A., Bagstad, K. J., Balbi, S., Marzo, A., ... & Villa, F. (2018). Machine learning for ecosystem services. *Ecosystem services*, *33*, 165- 174.

Witten, I. H., Frank, E., Hall, M., & Pal, C. J. (2016). Bias. *Data Mining: Practical Machine Learning Tools and Techniques*, 33f.

Wu, Y., Xie, L., Huang, S. L., Li, P., Yuan, Z., & Liu, W. (2018). Using social media to strengthen public awareness of wildlife conservation. *Ocean & Coastal Management*, *153*, 76-83.

Yoshimura, N., & Hiura, T. (2017). Demand and supply of cultural ecosystem services: Use of geotagged photos to map the aesthetic value of landscapes in Hokkaido. Ecosystem services, 24, 68-78.

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

Zhou, B., Lapedriza, A., Khosla, A., Oliva, A., & Torralba, A. (2017). Places: A 10 million image database for scene recognition. *IEEE transactions on pattern analysis and machine intelligence*, *40*(6), 1452-1464.

Zhu, Y., Zhou, Y., Xu, H., Ye, Q., Doermann, D., & Jiao, J. (2019). Learning instance activation maps for weakly supervised instance segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 3116-3125).

Zou, J., Han, Y., & So, S. S. (2008). Overview of artificial neural networks. In *Artificial Neural Networks* (pp. 14-22). Humana Press.

Appendix

In this appendix I want to remark all of the extra work I did during the development of this thesis: oral communication “Computer Vision of Cultural Ecosystem Services” at the online postgraduate course “Linking biodiversity to ecosystem functioning and services under global change” organized by the Centre of Molecular and the Institute of Science and Innovation for Bio-Sustainability (IB-S*)* at the University of Minho, proposal for the Con X Tech Prize of the Conservation X Labs, application for a PhD grant of the MIT 2020 call for PhD grants and Poster and Paper presentation at the 26th Portuguese Conference on Pattern Recognition (RECPAD 2020).

All of the code used to develop the algorithms of this thesis is available at https://github.com/anasccardoso/Deep-learning-for-cultural-ecosystem-services.git.

Deep learning to automate the assessment of cultural ecosystem services from social media data

Ana Sofia Cardoso[1]
up201804335@fc.up.pt
Ana Sofia Vaz[2]
sofia.linovaz@gmail.com
Francesco Renna[1]
frarenna@dcc.fc.up.pt

[1]Instituto de Telecomunicações, Faculdade de Ciências da Universidade do Porto
[2]Andalusian Inter-University Institute for Earth System Research (IISTA-CEAMA), University of Granada & Research Centre in Biodiversity and Genetic Resources (CIBIO-InBIO), University of Porto

Abstract

Cultural ecosystem services (CES) result from the interactions between humans and nature, contributing to people's physical and mental well-being. Most social media content analyses considered in the context of CES are based on the manual classification of photos or texts shared by social media users. Inevitably, the manual classification of big photographic data is too time consuming and costly, particularly when it comes to large study areas and audiences. In this work we studied automated image classification techniques using deep learning approaches to address CES.

1 Introduction

Nowadays, computer science and related fields have been highly invested in the use and combination of methods that incorporate social media analytics [1]. Social media platforms represent a very significant fraction of all the available digital data, constituting an efficient method to collect big data that provide information on people's interactions with each other and with their environment [2]. Fast improvements in computational power and data storage capacity during the last years have motivated the emergent fields of Digital Conservation, iEcology and conservation culturomics [3]. These disciplinary fields refer to the use of digital (big) data and technology to understand human-nature interactions and to provide evidence in favour of nature conservation and of the sustainable management of ecosystems [4]. Among these human-nature interactions are cultural ecosystem services (CES), which constitute the non-material benefits that people can experience from nature, such as recreation and ecotourism, as well as those pertaining to spiritual, religious, aesthetic or heritage values, among others [5].

An approach that combines different data from social media with advanced analytics, besides spatial analysis, remains underexplored in the context of CES assessment. Thus, the investment in methods that can identify features of ecosystems and nature through the content analysis of shared photos (or text), can constitute an asset to support the evaluation of CES, particularly, related to aesthetics and recreation or ecotourism [6]. Lee *et al*., for example, proposed a method for analysing large amounts of social media photographs, as well as to derive indicators of socio-cultural usage of landscapes, through cluster detection with Convolutional Neural Networks (CNNs) [7]. This project aims to develop an automated classification of social media photographs that can be useful for CES evaluation and for providing innovative solutions to the scientific community. Specifically, this study aims to answer the following questions: (1) can deep learning algorithms be developed to support an automated classification of social media photographs in the context of CES? and (2) how can those algorithms and models be improved so as to promote statistically reliable image classifications? To achieve this, deep learning algorithms are developed and tested, more specifically CNNs and transfer learning strategies are applied to the classification of digital photographs of the "Peneda-Gerês" protected area (Northern Portugal) obtained from the social media platforms Flickr and Wikiloc.

2 Methods

2.1Image classification methodology

We performed a classification of the content of photographs from the protected area "Peneda-Gerês" (Northern Portugal), that were withdrawn from the Flickr and Wikiloc social media platforms, specifying a time window of 2003-2017 (1778 images in total). This classification was based on "Nature" and "Human" labels (Figure 1). To achieve that, two different CNNs architectures were implemented, the VGG16 and the ResNet152, in order to verify the most appropriate and suitable for our study.

The proposed image classification methods were evaluated over the dataset using a 5-fold-cross validation method, following the literature and taking into account the computational resources and the running time. The considered performance metrics (accuracy, sensitivity, specificity, and F1-score) were computed as the mean of the performance metrics obtained over the 5 different folds. During training, in each of the 5 folds, 10% of the training data was retained to perform model validation, in order to determine the training parameters that guaranteed the highest accuracy over the validation set.

Since we are coping with a small dataset, in order to improve the generalization of the model and avoid the overfitting, transfer learning and data augmentation schemes were considered.

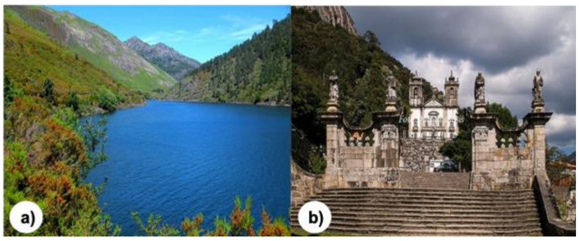


Figure 1: Examples of images belonging to the Nature and Human labels. a) Nature, b) Human.

2.2CNN architectures and transfer learning

The VGG16 and ResNet152 were the chosen CNNs architectures. For both CNN architectures, three different sets of weights were considered: (1) weights obtained by training over the dataset "Places365", (2) weights obtained by training over the database "ImageNet" and (3) weights obtained by training the networks from scratch.

The Places365 dataset is the latest subset of the database Places, comprising around 1.8 million scene photographs of different places, labelled with 365 scene semantic categories, including photographs with similar elements to the ones under study. The ImageNet database constitutes a large-scale hierarchical image database, that has several applications in the broadest areas, comprising more than 14 million cleanly annotated images spread over around 21,000 categories. Both databases were selected due to their freely available online resources (weights and models).

Regarding the details of the transfer learning strategy implemented, all the convolutional layers were kept frozen when training over our dataset, while the remaining 3 (for VGG16) and 1 (for ResNet152) fully connected layers were trained with our dataset. Moreover, for both architectures, an additional dense layer with 128 units and a rectifier linear unit activation function was also included (to allow better fit of the model/network to the classification task) before the output layer, which was modified in order to have 2 units.

Regarding the training details, both networks were trained using the Adam optimizer. For VGG16, the best performance was verified when considering a learning rate of 0.000001 while, for ResNet152, it was 0.0001 the most accurate learning rate. Also, it was observed that, for VGG16, the model accuracy and loss had fully converged after 50 epochs, having been decided, because of that, to use only 50 epochs to build the VGG16 model, as well as the ResNet152 model, due to computing resource management.

2.3Data augmentation

Regarding data augmentation, 5 transformations (including horizontal flip, width shift, height shift and zoom) were implemented individually for each of the images in the training set. The images in the validation set were not included in this process, in order to avoid biased results. The total number of transformations applied to each photograph (5 per image) was selected taking into account the overall running time of the algorithm, as well as the available computational memory.

3 Results

3.1Nature vs. Human classification

When comparing the two transfer learning scenarios and the weights obtained by training only over our dataset (Figure 1), it was observed that, ImageNet had, overall, a higher accuracy for the two architectures under study (86.11 vs 87.18), followed by Places365 and weights trained only with our dataset, with the exception of Places365 in VGG16, that resulted in an equally high accuracy (87.01). Also, it was verified that, for Places365, VGG16 had a better performance when compared to ResNet152 (87.01 vs 86.00), while for the remaining scenarios, ResNet152 model was more accurate than the one for VGG16.

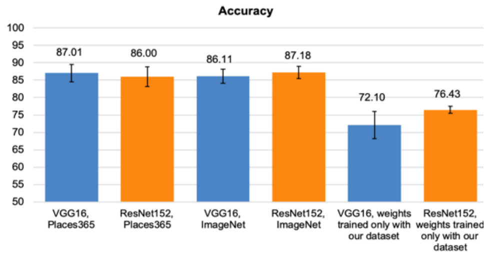


Figure 1: Accuracy of the VGG16 and ResNet152 model performance for the two transfer learning scenarios and the weights from scratch.

Considering sensitivity (Figure 2), it was verified that ImageNet had, overall, better results for the two architectures under study (86.71 and 86.78), followed by Places365 and weights trained only with our dataset, with the exception of Places365 in VGG16, that resulted in a higher sensitivity value (88.48). Likewise, it was observed that ResNet152 had slightly finer sensitivity results when compared to VGG16, except for Places365, where VGG16 showed the best result (88.48 vs 83.40).

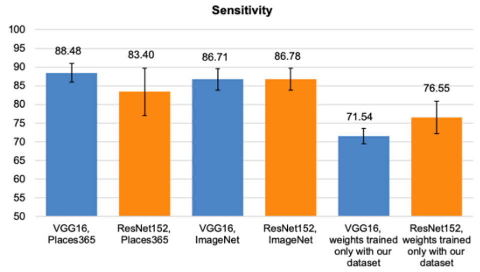


Figure 2: Sensitivity of the VGG16 and ResNet152 model performance for the two transfer learning scenarios and the weights from scratch.

For specificity (Figure 3), it was observed that Places365 had finer specificity results for the two architectures under study (85.54 and 88.46), followed by ImageNet and weights trained only with our dataset. Similarly, it was verified that ResNet152 had better specificity results when compared to VGG16, for all the scenarios under study.

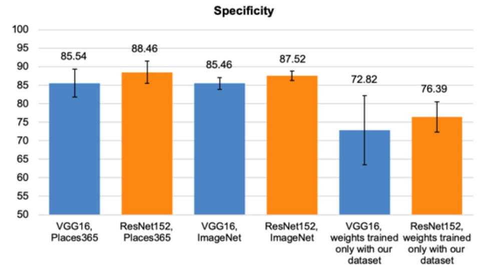


Figure 3: Specificity of the VGG16 and ResNet152 model performance for the two transfer learning scenarios and the weights from scratch.

Considering the F1-score (Figure 4), it was verified that ImageNet had slightly better F1-score results for the two architectures under study (86.53 and 87.44), followed by Places365 and weights trained only with our dataset. Also, it was observed that ResNet152 had finer F1-score results when compared to VGG16, except for Places365, where VGG16 showed the best result (87.53 vs 85.89).

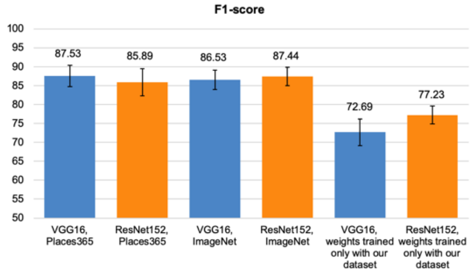


Figure 4: F1-score of the VGG16 and ResNet152 model performance for the two transfer learning scenarios and the weights from scratch.

4 Discussion and Conclusions

When comparing the two considered transfer learning scenarios and the weights obtained by training only over our dataset, it was expected that the model implemented with the Places365 weights would have a finer performance than the other two (with ImageNet weights and weights trained only with our dataset), since all the photographs contained in this dataset are exclusively related with landscapes and places in general, constituting the database that most resembles our dataset. Perhaps surprisingly, this was not the case for both VGG16 and ResNet152, as ImageNet was undoubtedly the database where the two transfer learning scenarios achieved better results. A possible explanation for this behavior can reside in the observation that deep learning models achieve more accurate results when trained in the presence of large datasets. In fact, ImageNet, by containing a larger number of photographs (more than 14 million) than Places365 (around 1.8 million), has led to a better performance of the model. Also, ImageNet contains a greater diversity of images that seems to contribute to a better generalization of the model.

The results showed that deep learning methods can offer significant contributions to assist in CES evaluation. Future work will focus on the improvement of the robustness of these models against scarcely labeled data via the use of semi-supervised approaches by leveraging autoencoder architectures and generative adversarial networks.

Acknowledgements

This work is funded by FCT/MCTES through national funds and when applicable co-funded EU funds under the project UIDB/50008/2020.

References

[1] Sherren Kate et al. Conservation culturomics should include images and a wider range of scholars. *Frontiers in Ecology and the Environment*, 2017, 15.6: 289-290. doi: 10.1002/fee.1507.

[2] Di Minin et al. Prospects and challenges for social media data in conservation science. *Frontiers in Environmental Science*, 2015, 3: 63. doi: 10.3389/fenvs.2015.00063.

[3] Jarić Ivan et al. iEcology: Harnessing Large Online Resources to Generate Ecological Insights. *Trends in Ecology & Evolution*, 2020. doi: 10.1016/j.tree.2020.03.003.

[4] Toivonen Tuuli et al. Social media data for conservation science: a methodological overview. *Biological Conservation*, 2019, 233: 298-315. doi: 10.1016/j.biocon.2019.01.023.

[5] Assessment Millennium Ecosystem et al. *Ecosystems and human well-being* (Vol. 5). United States of America: Island press, 2005. Doi:

[6] Richards Daniel R.; Tunçer, Bige. Using image recognition to automate assessment of cultural ecosystem services from social media photographs. *Ecosystem services*, 2018, 31: 318- 325. doi: 10.1016/j.ecoser.2017.09.004.

[7] Lee Heera et al. Mapping cultural ecosystem services 2.0–Potential and shortcomings from unlabeled crowd sourced images. *Ecological Indicators*, 2019, 96: 505-515. doi: 10.1016/j.ecolind.2018.08.035.

Universidade do Minho

cbma centre of **molecular** and environmental **biology**

IB-S
INSTITUTO DE CIÊNCIA E INOVAÇÃO
PARA A BIO-SUSTENTABILIDADE

TO WHOM IT MAY CONCERN

We confirm that Ana Sofia Cardoso, Master student at the Faculty of Sciences of the University of Porto, presented the results of her Master thesis on *Computer Vision of Cultural Ecosystem Services* during the online postgraduate course *Linking biodiversity to ecosystem functioning and services under global change* organized by the Centre of Molecular and the Institute of Science and Innovation for Bio-Sustainability (IB-S) at the University of Minho, Braga, Portugal, from 4 to 15 May 2020.

University of Minho, 19 May 2020

On behalf on the coordinators of the Course

F. Cássio

Fernanda Cássio

(Full Professor)

CBMA director

Cláudia Pascoal

(Associate Professor)

IB-S vice-director