

An Evaluation of Pre-Trained Models for Feature Extraction in Image Classification

Erick da Silva Puls, Matheus V. Todescato^a and Joel L. Carbonera^b

Institute of Informatics, UFRGS, Porto Alegre, Brazil

Keywords: Image Classification, Transfer Learning, Deep Learning, Geology.

Abstract: In recent years, we have witnessed a considerable increase in performance in image classification tasks. This performance improvement is mainly due to the adoption of deep learning techniques. Generally, deep learning techniques demand a large set of annotated data, making it challenging when applied to small datasets. Transfer learning strategies have become a promising alternative to overcome these issues in this scenario. This work compares the performance of different pre-trained neural networks for feature extraction in image classification tasks. We evaluated 16 different pre-trained models in four image datasets. Our results demonstrate that the best general performance along the datasets was achieved by CLIP-ViT-B and ViT-H-14, where the CLIP-ResNet50 model had similar performance but with less variability. Therefore, our study provides evidence supporting the choice of models for feature extraction in image classification tasks.

1 INTRODUCTION

The rapid technological advancements in the last decades have pushed organizations to produce and accumulate all kinds of data. In the past, critical organizational information was primarily represented by structured data stored in databases. However, nowadays, a significant part of this information is represented unstructured, such as images (Pferd, 2010).


There is a need to develop approaches capable of recovering and evaluating images in applications of several fields (Pferd, 2010). In that sense, one of the challenges concerning image recovery is that the semantic content of images is not apparent, so this information is not easily acquired through direct queries. An alternative for recovering images is annotating them first (Hollink et al., 2003) in a way that allows us to retrieve them by querying for the annotations. However, it is necessary to bear in mind that manual annotation of large databases of images is time-consuming and impractical. In this context, we can use machine learning approaches to automatically classify these large databases of images, thus enabling retrieval through direct queries.


Image classification (IC) tasks aim to classify the image by assigning a specific label. Usually, labels in

an IC task refer to objects that appear in the image, kinds of images (photographs, drawings, etc.), feelings (sadness, happiness, etc.), etc (Lanchantin et al., 2021).

Most of the recent approaches for IC are based on deep neural network (DNN) architectures. These architectures usually demand a large set of annotated data, making it challenging to apply *deep learning* when small amounts of data are available. In this scenario, *transfer learning* strategies have become a promising alternative to overcome these issues. One of the main alternatives of *transfer learning* is through *feature extraction*, where models trained on large datasets can produce informative features that another classifier can use. Using *transfer learning*, we can leverage knowledge previously learned by neural network models on a large dataset and use this knowledge in a context where just small datasets are available.

There are currently several large datasets available, such as Imagenet (Deng et al., 2009), and a range of models that were pre-trained on these datasets¹. The literature suggests that particular tasks on distinct datasets can benefit from different pre-trained models (Mallouh et al., 2019; Arslan et al., 2021).

^a  <https://orcid.org/0000-0001-7568-8784>

^b  <https://orcid.org/0000-0002-4499-3601>

¹Some pre-trained models are found in <https://pytorch.org/vision/stable/models.html>

It is essential to notice that there are different approaches for transfer learning for image classification, such as *fine-tuning* and *feature extraction*. When adopting *fine-tuning*, a neural network pre-trained in a big dataset is retrained in a novel task, for which usually only a small dataset is available. The goal of this approach is to use the knowledge (represented by the weights of the model) acquired in the first training process as a starting point for the training in the second task, and the weights of the pre-trained model are updated during the training in the target task. In the case of *feature extraction*, the pre-trained model extracts features that represent the images and can be used as input for a classifier. Notice that in this approach, the pre-trained model is kept frozen; that is, their weights are not updated during the training of the classifier used in the target task. Some studies (Kieffer et al., 2017; Mormont et al., 2018) comparing fine-tuning and feature extraction demonstrate that fine-tuning achieves higher performance. Still, the results also suggest that feature extraction achieves a comparable performance while requiring fewer computational resources for training. In this context, the main goal of this work is to compare and evaluate the performance of *feature extraction* (FE) of various pre-trained models in the image classification task.

In this study, Geological Images (Todescato et al., 2023; Abel et al., 2019; Todescato et al., 2024), Stanford Cars (Krause et al., 2013), CIFAR-10 (Krizhevsky et al., 2009), and STL10 (Coates et al., 2011) are the datasets adopted for analyzing the performance of FE of the following pre-trained models: AlexNet, ConvNeXt Large, DenseNet-161, GoogLeNet, Inception V3, MNASNet 1.3, MobileNet V3 Large, RegNetY-3.2GF, ResNeXt101-64x4D, ShuffleNet V2 X2.0, SqueezeNet 1.1, VGG19 BN, VisionTransformer-H/14, Wide ResNet-101-2, and both CLIP-ResNet50, and CLIP-ViT-B. We evaluate the performance of the considered pre-trained models using these metrics: accuracy, macro F1-measure, and weighted F1-measure. Our results analysis involves a comparison between the models, analyzing the potential of each one in each dataset and also analyzing the correlation between each model. Furthermore, we also explore the results of each dataset to understand which is the most difficult and the easiest for the models to classify.

Our results indicate that the pre-trained models CLIP-ResNet50, CLIP-ViT-B, and VisionTransformer-H/14 had significantly better performance than the other considered pre-trained models for all datasets. It is important to notice that these are the only three models among those considered in our experiments that include transformers in

their architecture, while the others are based solely on CNN architectures. Our analysis also indicates differences regarding the pattern of performances of these three transformer-based architectures compared to those of the CNN-based architectures across the datasets in all the considered metrics. These differences become evident when we analyze the Pearson correlation in Section 3.4. Moreover, our analysis suggests that the Stanford Cars dataset is the most challenging of all datasets analyzed. We hypothesize that it is due to its large number of classes, few samples per class, and the inclusion of images with different sizes and features at different scales.

The remainder of this paper is structured as follows. Section 2 discusses the related work. In Section 3, we present our experiments and discuss our results. Finally, Section 4 presents the conclusions.

2 RELATED WORK

The TL approach based on FE has been adopted for IC in several domains, such as Biomedicine (Alzubaidi et al., 2021), and Geology (Dosovitskiy et al., 2020; Maniar et al., 2018; Karpatne et al., 2018). In this work, we reviewed the literature covering the last five years that focused on comparing the performance of FE for different pre-trained models. In our literature review, the most frequently used pre-trained models were the VGG16, the Inception V3, and AlexNet.

An extensive range of pre-trained models can be applied for transfer learning. The main expected result of FE from the pre-trained models is to improve classification quality. The size and similarity of the target dataset and the source task can be used to choose the pre-trained model (Fawaz et al., 2018).

The literature suggests that each dataset may need a different pre-trained model. For instance, for plankton classification (Lumini and Nanni, 2019), when adopting the models as a feature extractor, the best result among a wide range of pre-trained models (Inception V3, AlexNet, VGG16, VGG19, ResNet50, ResNet101, DenseNet-161, and GoogLeNet) is using DenseNet-161. On the other hand, when classifying pathological brain images, Kaur & Gandhi (2020) found that the AlexNet showed the best results among eight pre-trained models.

Finally, using the CIFAR-10 dataset and experimenting with the Inception V3, GoogLeNet, SqueezeNet 1.1, and DarkNet53, ShuffleNet models, (Kumar et al., 2022) found that, overall, the Inception V3 model achieved the highest accuracy, as well as higher values in other evaluation metrics including

precision, sensitivity, specificity, and F1-score (Kumar et al., 2022).

In summary, finding a suitable pre-trained model can be challenging for specific application needs. Different models can present better results for other datasets and different performance parameters (Abou Baker et al., 2022). Therefore, it is essential to systematically investigate the usability of several pre-trained models to find the best match for specific datasets.

Recent papers are showing the capacity of DL and TL to facilitate the analysis of uninterpreted images that have been neglected due to a limited number of experts, such as fossil images, slabbed cores, or petrographic thin sections (De Lima et al., 2019), or even for environmental images (Sun et al., 2021). The ability to create distinctive models for specific datasets allows a versatile application of those techniques.

When comparing pre-trained models, in (De Lima et al., 2019), the authors found that both MobileNet V2 and Inception V3 showed promising results on geologic data interpretation, with MobileNet V2 having slightly better results. Also, in (Sun et al., 2021), the authors compared the performance of AlexNet, VGG16, ResNet50, GLNet (AlexNet), GLNet (VGG16), and GLNet (ResNet) pre-trained models on remote sensing scene classification using FE, concluding that their proposed new model shows better results compared to other traditional DNN architectures. The proposed model GLNet, which uses VGG16 as its base, got over 95% accuracy in analyzing a clear environment and over 94% in a cloudy environment. In contrast, the traditional VGG16 got over 93% and over 78%, for clear and cloud environments, respectively (Sun et al., 2021).

It is important to notice that the literature does not provide a systematic comparison of the performance in feature extraction for image classification tasks, covering a broad range of models and datasets with different characteristics. Our work aims to use models that achieve promising performances in these related works and other more recent models that do not appear in these comparisons. We apply these selected models to image classification benchmark datasets to evaluate their performances in different datasets and provide evidence for supporting the choice of the suitable model.

3 EXPERIMENTS

In this section, we discuss the experiments to evaluate different pre-trained models in different datasets. Firstly, we present the datasets and the models used in

our experiments. After, we describe the methodology adopted in these experiments. Finally, we present and discuss the results of the experiments.

3.1 Datasets

There are several widely used image datasets in computer vision research. We adopted the following ones in our experiments: Stanford Cars (Krause et al., 2013), CIFAR-10 (Krizhevsky et al., 2009), and STL10 (Coates et al., 2011). In addition to these, we also consider the Geological Images dataset (Abel et al., 2019), a domain-specific “real-world” dataset that includes a set of annotated images that are relevant for applications in Geosciences (Todescato et al., 2023). These datasets are widely used in the literature, are colorful, and have different characteristics.

Table 1 shows the main information of all datasets used in this work.

Notice that CIFAR-10 and STL10 are *balanced*, include sets of images of *homogeneous size* (96x96 in the STL-10 and 32x32 in the CIFAR-10), and have a *small number of classes*. The Geological images and the Stanford Cars datasets are unbalanced (Stanford Cars is slightly unbalanced) and have images of *heterogeneous sizes* and a *higher number of classes* when comparing with CIFAR-10 and STL10.

3.2 Pre-Trained Models

Due to the increasing adoption of transfer learning, several pre-trained models are available in the literature nowadays. In our work, we use a wide range of pre-trained models available in repositories². The majority of these models considered in this work were pre-trained using ImageNet-1K³(Deng et al., 2009) dataset except for the CLIP(Radford et al., 2021) based models that were pre-trained in a dataset with 400 million images called WebImageText (WIT). Table 2 presents the following properties of the selected models: number of output features, number of parameters, training dataset, and architecture. Notice that clip-rn50 and clip-vit-b adopt two encoders, one for images and the other for text, and they were pre-trained using pairs of images and text. Thus, in Table 2, the notation CNN + Tr means that the image encoder is based on CNN and the text encoder is based on transformers.

²Can be accessed through <https://pytorch.org/vision/stable/models.html> and <https://github.com/openai/CLIP>

³Can be accessed through <https://image-net.org/index.php>

Table 1: Datasets Information.

Dataset	Instances	Classes	Avg Instances \pm Std per Class
Geo Images (Abel et al., 2019)	25725	45	571,67 \pm 1290,90
Stanford Cars (Krause et al., 2013)	16185	196	84 \pm 6,28
CIFAR-10 (Krizhevsky et al., 2009)	60000	10	6000 \pm 0
STL10 (Coates et al., 2011)	100000	10	10000 \pm 0

Table 2: Pre-trained Models Information. CNN indicates a convolutional neural networks architecture and Tr indicates a transformer-based architecture.

Models	Output Features	Parameters	Architecture
alexnet(Krizhevsky, 2014)	256	61,100,840	CNN
clip_rn50(He et al., 2016; Radford et al., 2021)	1024	63,000,000	CNN + Tr
clip_vit_b(Radford et al., 2021)	512	63,000,000	Tr + Tr
convnext_large(Liu et al., 2022)	1536	197,767,336	CNN
densenet161(Huang et al., 2017)	2208	28,681,000	CNN
googlenet(Szegedy et al., 2015)	1000	6,624,904	CNN
inception_v3(Szegedy et al., 2016)	1000	27,161,264	CNN
mnasnet1_3(Tan et al., 2019)	1000	6,282,256	CNN
mobilenet_v3_large(Howard et al., 2019)	960	5,483,032	CNN
regnet_y_3_2gf(Radosavovic et al., 2020)	1000	19,436,338	CNN
resnext101_64x4d(Xie et al., 2017)	1000	83,455,272	CNN
shufflenet_v2_x2_0(Ma et al., 2018)	1000	7,393,996	CNN
squeezenet1_1(Iandola et al., 2016)	512	1,235,496	CNN
vgg19_bn(Simonyan and Zisserman, 2014)	512	143,678,248	CNN
vit_h_14(Dosovitskiy et al., 2020)	1000	632,045,800	Tr
wide_resnet101_2(Zagoruyko and Komodakis, 2016)	1000	126,886,696	CNN

3.3 Methodology

We aim to evaluate the performance of different available pre-trained models as feature extractors in the image classification task in other datasets. We used the datasets and models previously detailed for conducting our experiments. Notice also that since different versions are available for each family of models, we have selected a single model for each family that presented the best overall performances according to the literature.

Since we are considering 16 models and four datasets, 64 experiments considering pairs of models and datasets were performed.

In each experiment, each pre-trained model was used as a feature extractor. Therefore, in this context, all initial layers (except the last one) of the model, responsible for extracting relevant features from the input images, were maintained, while the last layer was replaced by a new classification layer, with output size N (where N is proportional to the number of classes in the used dataset) using a linear activation function and a softmax. During the training, the weights of the initial layers (responsible for extracting features)

are kept frozen while the weights of the last layer are adjusted.

For each experiment, the datasets went through a homogeneous pre-processing. The pre-processing consisted of applying resizing, center cropping, and normalization. The resize is always done by decreasing or increasing the size of the image's smallest dimension to the size of the pre-trained model's input. Then, we perform the center crop, where the central area of the image is cut as a square that matches the size of the model's input. Finally, we ensure that all images are converted to RGB.

Each model was evaluated considering a 5-fold cross-validation process. To comprehensively assess the efficacy of our approach, we adopted three different metrics: accuracy, macro F1-score, and weighted F1-score. These metrics offer a robust evaluation of the results since they cover several evaluation aspects in a multiclass classification setting. Accuracy is a fundamental measure of overall correctness (although it can be misleading in contexts with data imbalance), while the macro F1-score offers insights into the model's ability to perform effectively across all classes, irrespective of class imbalances. Addition-

Table 3: Geological Images Dataset.

Geological Images Dataset							
Model \ Metrics	Accuracy	Macro			Weighted		
		Precision	Recall	F-Score	Precision	Recall	F-Score
alexnet	0,85	0,74	0,67	0,69	0,84	0,85	0,84
clip_rn50	0,93	0,86	0,83	0,84	0,92	0,93	0,92
clip_vit_b	0,93	0,86	0,83	0,84	0,93	0,93	0,93
convnext_large	0,91	0,84	0,80	0,82	0,91	0,91	0,91
densenet161	0,90	0,83	0,78	0,80	0,90	0,90	0,90
googlenet	0,87	0,75	0,72	0,73	0,86	0,87	0,86
inception_v3	0,83	0,70	0,65	0,67	0,82	0,83	0,83
mnasnet1_3	0,88	0,77	0,73	0,75	0,87	0,88	0,87
mobilenet_v3_large	0,90	0,82	0,77	0,79	0,90	0,90	0,90
regnet_y_3_2gf	0,89	0,79	0,76	0,77	0,89	0,89	0,89
resnext101_64x4d	0,88	0,79	0,74	0,76	0,88	0,88	0,88
shufflenet_v2_x2_0	0,89	0,80	0,76	0,78	0,89	0,89	0,89
squeezenet1_1	0,87	0,77	0,72	0,74	0,87	0,87	0,87
vgg19_bn	0,88	0,79	0,74	0,76	0,88	0,88	0,88
vit_h_14	0,91	0,82	0,79	0,80	0,90	0,91	0,90
wide_resnet101_2	0,89	0,79	0,75	0,77	0,89	0,89	0,89
Average	0,89	0,79	0,75	0,77	0,88	0,89	0,88
Standard Deviation	0,02	0,04	0,05	0,05	0,03	0,02	0,03

Table 4: Stanford Cars Dataset.

Stanford Cars Dataset							
Model \ Metrics	Accuracy	Macro			Weighted		
		Precision	Recall	F-Score	Precision	Recall	F-Score
alexnet	0,28	0,26	0,28	0,26	0,26	0,28	0,26
clip_rn50	0,82	0,82	0,82	0,82	0,82	0,82	0,82
clip_vit_b	0,83	0,83	0,83	0,83	0,83	0,83	0,83
convnext_large	0,65	0,65	0,64	0,64	0,65	0,65	0,64
densenet161	0,64	0,64	0,64	0,64	0,64	0,64	0,64
googlenet	0,41	0,41	0,41	0,41	0,40	0,41	0,40
inception_v3	0,34	0,33	0,34	0,33	0,33	0,34	0,33
mnasnet1_3	0,42	0,42	0,42	0,42	0,41	0,42	0,41
mobilenet_v3_large	0,56	0,56	0,56	0,55	0,56	0,56	0,55
regnet_y_3_2gf	0,49	0,49	0,49	0,49	0,49	0,49	0,49
resnext101_64x4d	0,35	0,35	0,35	0,34	0,34	0,35	0,34
shufflenet_v2_x2_0	0,50	0,50	0,50	0,50	0,50	0,50	0,50
squeezenet1_1	0,42	0,42	0,42	0,41	0,41	0,42	0,41
vgg19_bn	0,51	0,50	0,51	0,50	0,50	0,51	0,50
vit_h_14	0,86	0,86	0,85	0,85	0,86	0,86	0,86
wide_resnet101_2	0,44	0,44	0,44	0,44	0,44	0,44	0,44
Average	0,53	0,53	0,53	0,53	0,53	0,53	0,53
Standard Deviation	0,18	0,18	0,18	0,18	0,18	0,18	0,18

Table 5: CIFAR-10 Dataset.

CIFAR-10 Dataset							
Model \ Metrics	Accuracy	Macro			Weighted		
		Precision	Recall	F-Score	Precision	Recall	F-Score
alexnet	0,79	0,79	0,79	0,79	0,79	0,79	0,79
clip_rn50	0,88	0,88	0,88	0,88	0,88	0,88	0,88
clip_vit_b	0,95	0,95	0,95	0,95	0,95	0,95	0,95
convnext_large	0,96	0,96	0,96	0,96	0,96	0,96	0,96
densenet161	0,93	0,93	0,93	0,93	0,93	0,93	0,93
googlenet	0,87	0,87	0,87	0,87	0,87	0,87	0,87
inception_v3	0,86	0,86	0,86	0,86	0,86	0,86	0,86
mnasnet1_3	0,90	0,90	0,90	0,90	0,90	0,90	0,90
mobilenet_v3_large	0,91	0,91	0,91	0,91	0,91	0,91	0,91
regnet_y_3_2gf	0,93	0,93	0,93	0,93	0,93	0,93	0,93
resnext101_64x4d	0,95	0,95	0,95	0,95	0,95	0,95	0,95
shufflenet_v2_x2_0	0,92	0,92	0,92	0,92	0,92	0,92	0,92
squeezenet1_1	0,85	0,85	0,85	0,85	0,85	0,85	0,85
vgg19_bn	0,88	0,88	0,88	0,88	0,88	0,88	0,88
vit_h_14	0,98	0,98	0,98	0,98	0,98	0,98	0,98
wide_resnet101_2	0,95	0,95	0,95	0,95	0,95	0,95	0,95
Average	0,91	0,91	0,91	0,91	0,91	0,91	0,91
Standard Deviation	0,05	0,05	0,05	0,05	0,05	0,05	0,05

Table 6: STL10 Dataset.

STL10 Dataset							
Model \ Metrics	Accuracy	Macro			Weighted		
		Precision	Recall	F-Score	Precision	Recall	F-Score
alexnet	0,88	0,88	0,88	0,88	0,88	0,88	0,88
clip_rn50	0,97	0,97	0,97	0,97	0,97	0,97	0,97
clip_vit_b	0,99	0,99	0,99	0,99	0,99	0,99	0,99
convnext_large	0,99	0,99	0,99	0,99	0,99	0,99	0,99
densenet161	0,98	0,98	0,98	0,98	0,98	0,98	0,98
googlenet	0,96	0,96	0,96	0,96	0,96	0,96	0,96
inception_v3	0,96	0,96	0,96	0,96	0,96	0,96	0,96
mnasnet1_3	0,97	0,97	0,97	0,97	0,97	0,97	0,97
mobilenet_v3_large	0,96	0,96	0,96	0,96	0,96	0,96	0,96
regnet_y_3_2gf	0,98	0,98	0,98	0,98	0,98	0,98	0,98
resnext101_64x4d	0,99	0,99	0,99	0,99	0,99	0,99	0,99
shufflenet_v2_x2_0	0,97	0,97	0,97	0,97	0,97	0,97	0,97
squeezenet1_1	0,91	0,91	0,91	0,91	0,91	0,91	0,91
vgg19_bn	0,96	0,96	0,96	0,96	0,96	0,96	0,96
vit_h_14	1,00	1,00	1,00	1,00	1,00	1,00	1,00
wide_resnet101_2	0,99	0,99	0,99	0,99	0,99	0,99	0,99
Average	0,97	0,97	0,97	0,97	0,97	0,97	0,97
Standard Deviation	0,03	0,03	0,03	0,03	0,03	0,03	0,03

ally, the weighted F1-score considers the non-uniform distribution of classes, providing a nuanced understanding of performance weighted by class prevalence. The reported metrics are averages obtained considering the performance in each test fold of this cross-validation process.

Regarding the training hyperparameters, the *learning rate* used in this study was 0.001 with a *momentum* of 0.9. We adopted the *Adam optimizer* with default parameters, with the *Cross-Entropy* loss function. All executions were done using 100 epochs and *early stopping* with a minimal improvement of 0.001 and patience of 5.

3.4 Results

The following tables represent the model’s performance according to the selected metrics for each dataset. Table 3 represents the model’s evaluation on the Geological Images dataset. Table 4 represents the model’s evaluation according to Stanford Cars dataset. Table 5 represents the model’s evaluation considering the CIFAR-10 dataset. Table 6 represents the model’s evaluation for the STL10 dataset. In each table, we highlight the model with the best performance in green and with the lowest performance in red.

To facilitate the data analysis, we have represented the results of our experiments in the following line charts, demonstrating the performance (according to different metrics) of each pre-trained model for classifying the images in the four selected datasets. Figure 1 represents the accuracy, Figure 2 demonstrates the macro f1-score, and Figure 3 indicates the weighted f1-score of each model on each dataset.

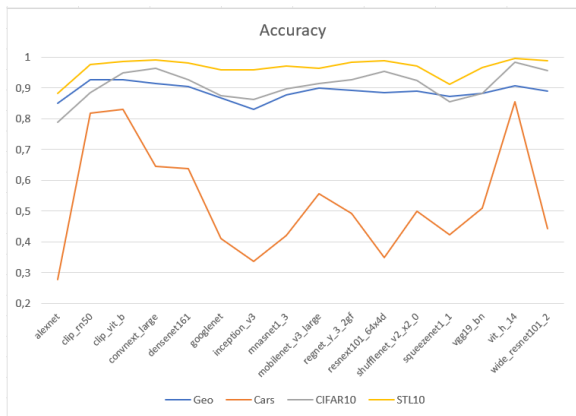


Figure 1: Line chart representing the accuracy of each model on each dataset.

The line charts in Figures 2-3 present a similar pattern of variation of the model’s performance

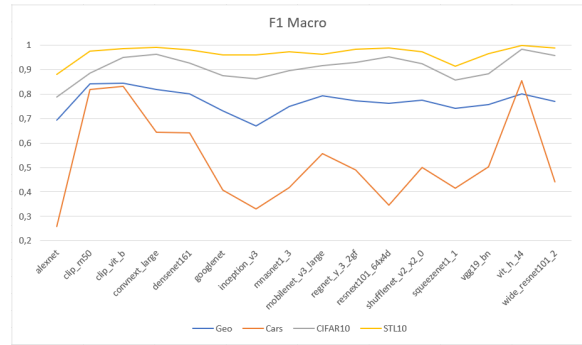


Figure 2: Line chart representing the macro f1-score of each model on each dataset.

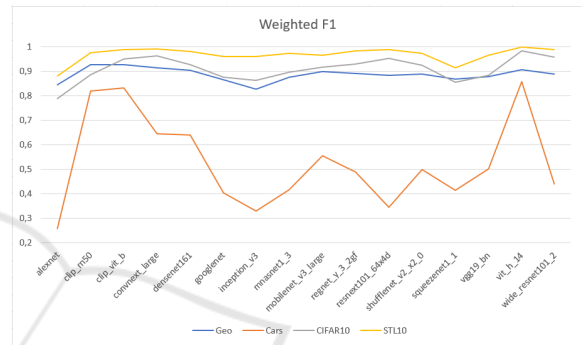


Figure 3: Line chart representing the weighted f1-score of each model on each dataset.

across all datasets. We can also notice that, in general, the model’s performance pattern increases, and the differences among patterns decrease (resulting in a smoother pattern) in the Geological Images dataset when we consider the accuracy and the weighted averages of the f1-score compared with the macro F1-score. This behavior is expected since the imbalance of this dataset is more apparent.

The CLIP-ViT-B and VisionTransformer-H/14 models generally show the best performances, considering all metrics in most datasets. The CLIP-ViT-B presents the best performance in the Geological Images dataset in all metrics. The CLIP-ResNet50 also performs well in the other datasets. However, in the case of the CIFAR-10 and Geological Images datasets, this model’s performance is reasonably lower than CLIP-ViT-B and ViT-H/14 in all metrics. In the CIFAR-10 dataset, it is also worth highlighting the good performance of ConvNeXt Large and ResNeXt101-64x4D. The ConvNeXt Large model also performs better than CLIP-ViT-B and CLIP-ResNet50 in STL10.

In our evaluation, AlexNet presents the worst performance on most datasets, considering all metrics. This is expected since it is less sophisticated than other models recently proposed. Inception V3 also

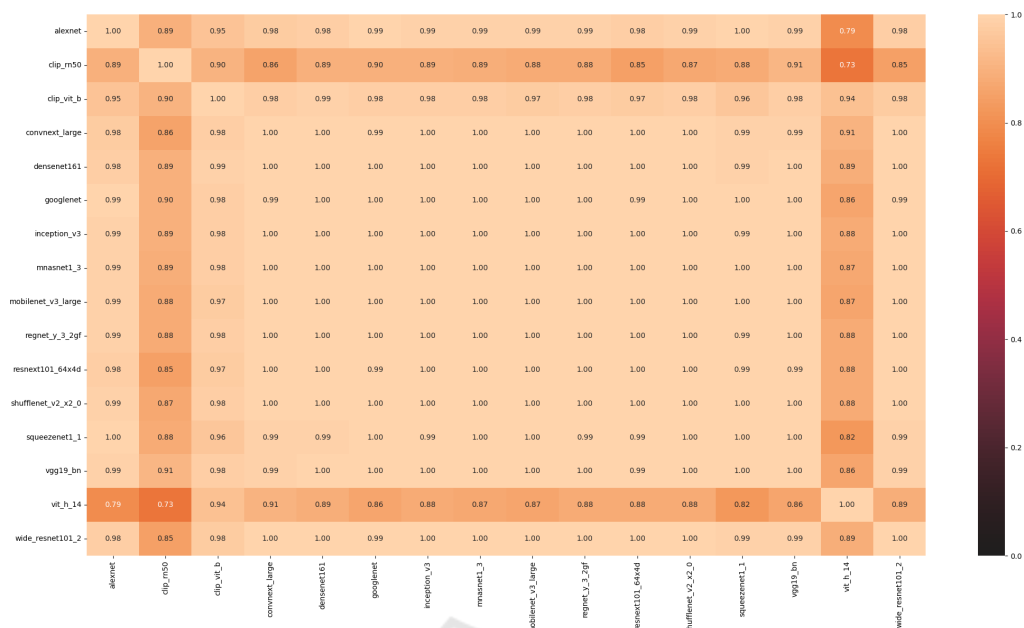


Figure 4: Heat map representing the correlation between each pair of models regarding accuracy.

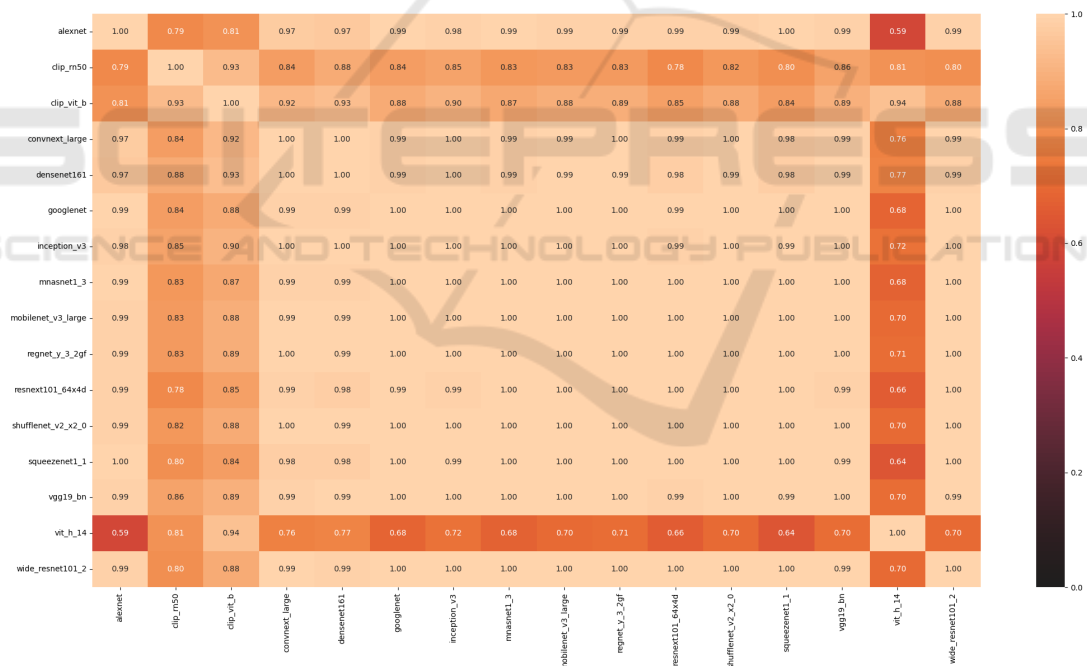


Figure 5: Heat map representing the correlation between each pair of models regarding the macro f1-score.

performed poorly on the Stanford Cars, CIFAR-10, and Geological Images datasets, where it performed worst. Another model that had reasonably low performance compared to the others was Squeezenet1-1. The poor performance of this model is more pronounced on the CIFAR-10 and STL10 datasets.

The previous analysis (Figures 1-3) suggests that some models present a very similar performance be-

havior across the datasets. In contrast, other models exhibit behaviors that do not follow the general pattern. To emphasize how similar are the model's behaviors, we analyzed the Pearson correlation (Cohen et al., 2009) of the performances of each pair of models across the datasets according to all the selected metrics. The Figures 4-6 visually represent heat maps with this information in a way that the darker a cell

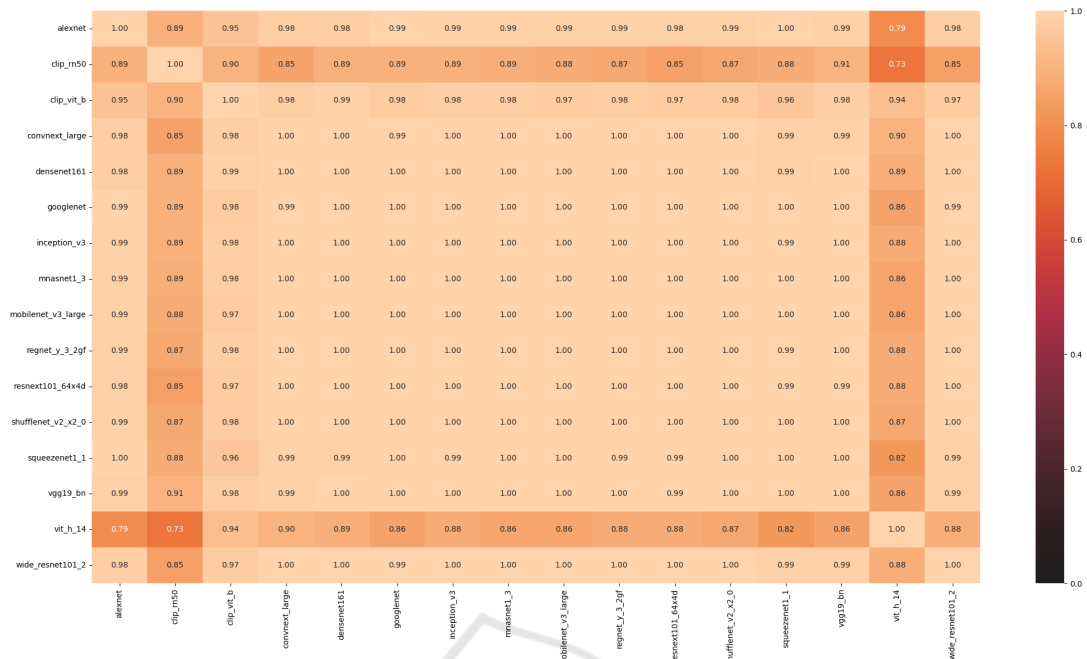


Figure 6: Heat map representing the correlation between each pair of models regarding the weighted f1-score.

gets corresponds to the lower the correlation of a given pair of models, according to a given performance metric. Figure 4 represents the pairwise Pearson correlation regarding the accuracy of each model, Figure 5 represents the pairwise Pearson correlation regarding the macro f1-score, and Figure 6 shows the pairwise Pearson correlation regarding the weighted f1-score between each pair of models.

Figures 4-6 suggest that the correlation of the performances of each pair of models presents a similar pattern in all metrics. We can also notice that, in all performance metrics, the correlation between models based solely on CNN architectures is high (generally above 0.97). However, CLIP-ResNet50, CLIP-ViT-B, and VisionTransformer-H/14 models present a lower correlation with the performances of other models solely based on CNN. In the case of CLIP-ViT-B, the correlation with the other models is subtly lower, considering accuracy and the weighted average f1-score. However, this model’s correlation is significantly lower when we consider the macro average of the f1-score. It is important to note that the CLIP-ResNet50, CLIP-ViT-B, and VisionTransformer-H/14 models include transformers in their architectures. This performance correlation analysis suggests that this difference in the basic principles of the architecture of these models is correlated with this difference in the performance pattern of these models when compared to architectures based solely on CNN. Further analysis should be done in the future to investigate this hypothesis. The heat maps also

allow us to note that the correlations among CLIP-ResNet50, CLIP-ViT-B, and VisionTransformer-H/14 models are low compared to the correlations among the performances of models based solely on CNN.

In the previous analyses, we focused on the performance of the models considered in our experiments. In the following boxplots, we focused on analyzing the datasets considered in our experiments. Figure 7 represents the variation in accuracy. Figure 8 demonstrates the variation of macro f1-score. Figure 9 shows the variation of weighted f1-score in each dataset.

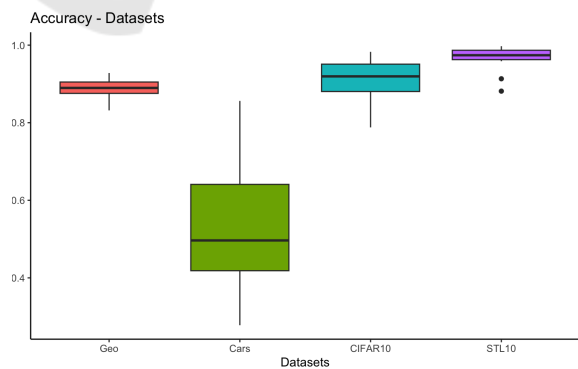


Figure 7: Boxplot of accuracy for each dataset.

The boxplots present a similar pattern seen across the different metrics. There are subtle differences when comparing the macro average of the f1-score with accuracy and weighted average. Note that, in

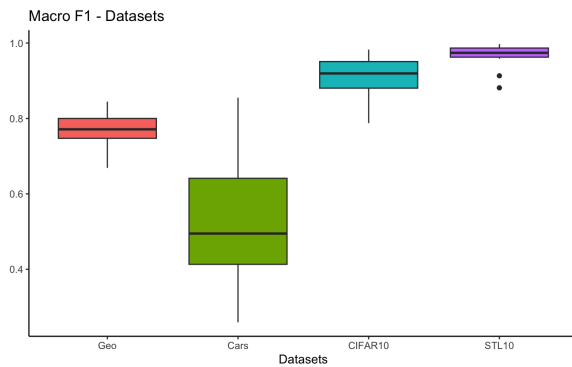


Figure 8: Boxplot of macro f1-score for each dataset.

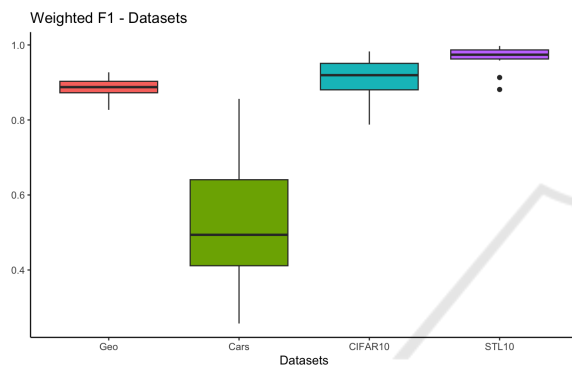


Figure 9: Boxplot of weighted f1-score for each dataset.

general, the models tend to perform better in the STL10 dataset; in second place, CIFAR-10 has the best overall results; in third place, the dataset of Geological Images and, finally, the dataset with the worst performances in general, is the Stanford Cars. The low performance in the Stanford cars is expected since this dataset has a large number of classes, few samples per class, and includes images with different sizes and features at different scales. The Geological Images dataset has similar properties but has fewer classes and more samples per class than Stanford Cars, although it presents a more significant imbalance. These charts allow us to conclude that the Stanford Cars dataset is the most challenging among those analyzed, with the worst and most considerable variability of performances in all metrics. Besides that, we can also notice that the STL10 dataset and Geological Images have a smaller variability in the performance of the different models when compared with the other two datasets.

4 CONCLUSION

In this work, our goal was to compare the performance of sixteen pre-trained neural networks for fea-

ture extraction in four different datasets. By analyzing the accuracy and macro and weighted averages of the f1-score in our experiments and considering all the datasets, our experiments suggest that CLIP-ViT-B and VisionTransformer-H/14 achieved the best performance results. Besides that, CLIP-ResNet50 achieved performance similar to the performance achieved by CLIP-ViT-B and VisionTransformer-H/14 and even lower variability. It is important to notice that CLIP-ViT-B, VisionTransformer-H/14, and CLIP-ResNet50 include transformers in their architectures. Thus, our results suggest that the principles underlying the transformers can be the reason corroborating these remarkable results, but with further studies, we can investigate this hypothesis.

Among the CNN-based architectures, ConvNeXt Large presents the best performance, in general, and lower variability when compared to other CNN-based architectures. AlexNet showed the worst performance and high variability. Besides that, ResNeXt101-64x4D, Wide ResNet 101-2, and Inception V3 also showed high variability.

Our analysis also showed that the performances of models based solely on CNN architectures present a high Pearson correlation in all performance metrics. However, the performances of CLIP-ResNet50, CLIP-ViT-B, and VisionTransformer-H/14 models show a lower correlation with other models based solely on CNN. We can hypothesize that this difference is due to differences regarding the basic principles of the architecture of these models. However, the correlations among CLIP-ResNet50, CLIP-ViT-B, and VisionTransformer-H/14 models are low compared to correlations among the performances of CNN-based models. Further studies are needed to investigate this finding better.

Our analysis also has shown that the selected models performed better on the STL10 dataset, followed by CIFAR-10, then the Geological Images dataset, and finally, the Stanford Cars dataset. Thus, the Stanford Cars dataset is the most challenging dataset evaluated in this work. The Stanford Cars present a considerable image size (compared with CIFAR-10 and STL10, for example) and many classes with just a few samples per class. These characteristics may explain this result. The Geological Images dataset shares some of the properties of the Stanford Cars dataset. Still, it presents fewer classes and has more images per class, in general.

The investigation presented in this work can provide evidence supporting the choice of transfer learning models in image classification tasks in “real-world” datasets such as the geological dataset. Since our evaluation also covered other image datasets with

different characteristics, it can suggest reasonable model choices in other domains.

In future works, it is important to expand the analysis by including other image datasets to make the analysis more comprehensive. Besides that, the investigation can also be expanded to include more pre-trained models that eventually were not considered in the scope of this work. Furthermore, future works could also investigate the relationship between the underlying principles of each architecture, the properties of the datasets used in the pre-training of these models, and the properties of the target datasets in which the pre-trained models are applied to extract features. This investigation can reveal insights into what makes the pre-trained model best suited for each task.

ACKNOWLEDGMENTS

The authors would like to thank the Brazilian National Council for Scientific and Technological Development (CNPq) and Petrobras for the financial support of this work.

REFERENCES

- Abel, M., Gastal, E. S. L., Michelin, C. R. L., Maggi, L. G., Firnkies, B. E., Pachas, F. E. H., and dos Santos Alvarenga, R. (2019). A knowledge organization system for image classification and retrieval in petroleum exploration domain. In *ONTOBRAS*.
- Abou Baker, N., Zengeler, N., and Handmann (2022). Uwe. a transfer learning evaluation of deep neural networks for image classification. *Machine Learning and Knowledge Extraction*, 4(1):22–41.
- Alzubaidi, L. et al. (2021). Review of deep learning: Concepts, cnn architectures, challenges, applications, future directions. *Journal of big Data*, 8(1):1–74.
- Arslan, Y., Allix, K., Veiber, L., Lothritz, C., Bissyandé, T. F., Klein, J., and Goujon, A. (2021). A comparison of pre-trained language models for multi-class text classification in the financial domain. In *Companion Proceedings of the Web Conference 2021*, pages 260–268.
- Coates, A., Ng, A., and Lee, H. (2011). An analysis of single-layer networks in unsupervised feature learning. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pages 215–223. JMLR Workshop and Conference Proceedings.
- Cohen, I., Huang, Y., Chen, J., Benesty, J., Benesty, J., Chen, J., Huang, Y., and Cohen, I. (2009). Pearson correlation coefficient. *Noise reduction in speech processing*, pages 1–4.
- De Lima, R. P. et al. (2019). Deep convolutional neural networks as a geological image classification tool. *The Sedimentary Record*, 17(2):4–9.
- Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee.
- Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.
- Fawaz, H. I. et al. (2018). Transfer learning for time series classification. In *2018 IEEE international conference on big data (Big Data)*, pages 1367–1376. IEEE p.
- He, K. et al. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition p*, pages 770–778.
- Hollink, L., Schreiber, G., Wielemaker, J., Wielinga, B., et al. (2003). Semantic annotation of image collections. In *Knowledge capture*, volume 2.
- Howard, A., Sandler, M., Chu, G., Chen, L.-C., Chen, B., Tan, M., Wang, W., Zhu, Y., Pang, R., Vasudevan, V., et al. (2019). Searching for mobilenetv3. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1314–1324.
- Huang, G., Liu, Z., Van Der Maaten, L., and Weinberger, K. Q. (2017). Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708.
- Iandola, F. N., Han, S., Moskewicz, M. W., Ashraf, K., Dally, W. J., and Keutzer, K. (2016). Squeezenet: Alexnet-level accuracy with 50x fewer parameters and; 0.5 mb model size. *arXiv preprint arXiv:1602.07360*.
- Karpatne, A. et al. (2018). Machine learning for the geosciences: Challenges and opportunities. *IEEE Transactions on Knowledge and Data Engineering*, 31(8):1544–1554.
- Kieffer, B., Babaie, M., Kalra, S., and Tizhoosh, H. R. (2017). Convolutional neural networks for histopathology image classification: Training vs. using pre-trained networks. In *2017 Seventh International Conference on Image Processing Theory, Tools and Applications (IPTA)*, pages 1–6. IEEE.
- Krause, J., Stark, M., Deng, J., and Fei-Fei, L. (2013). 3d object representations for fine-grained categorization. In *4th International IEEE Workshop on 3D Representation and Recognition (3dRR-13)*, Sydney, Australia.
- Krizhevsky, A. (2014). One weird trick for parallelizing convolutional neural networks. *arXiv preprint arXiv:1404.5997*.
- Krizhevsky, A., Hinton, G., et al. (2009). Learning multiple layers of features from tiny images.
- Kumar, J. S., Anuar, S., and Hassan, N. H. (2022). Transfer learning based performance comparison of the pre-trained deep neural networks. *International Journal of Advanced Computer Science and Applications*, 13:1.

- Lanchantin, J., Wang, T., Ordonez, V., and Qi, Y. (2021). General multi-label image classification with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16478–16488.
- Liu, Z., Mao, H., Wu, C.-Y., Feichtenhofer, C., Darrell, T., and Xie, S. (2022). A convnet for the 2020s. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11976–11986.
- Lumini, A. and Nanni, L. (2019). Deep learning and transfer learning features for plankton classification. *Ecological informatics*, 51:33–43.
- Ma, N., Zhang, X., Zheng, H.-T., and Sun, J. (2018). Shufflenet v2: Practical guidelines for efficient cnn architecture design. In *Proceedings of the European conference on computer vision (ECCV)*, pages 116–131.
- Mallouh, A. A., Qawaqneh, Z., and Barkana, B. D. (2019). Utilizing cnns and transfer learning of pre-trained models for age range classification from unconstrained face images. *Image and Vision Computing*, 88:41–51.
- Maniar, H. et al. (2018). Machine-learning methods in geoscience. In *2018 SEG International Exposition and Annual Meeting*. OnePetro.
- Mormont, R., Geurts, P., and Marée, R. (2018). Comparison of deep transfer learning strategies for digital pathology. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 2262–2271.
- Pferd, J. (2010). The challenges of integrating structured and unstructured data. In *14th Petroleum Network Education Conference*. s.n., S.I.
- Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. (2021). Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.
- Radosavovic, I., Kosaraju, R. P., Girshick, R., He, K., and Dollár, P. (2020). Designing network design spaces. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10428–10436.
- Simonyan, K. and Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.
- Sun, H. et al. (2021). Convolutional neural networks based remote sensing scene classification under clear and cloudy environments. In *Proceedings of the IEEE/CVF International Conference on Computer Vision p*, pages 713–720.
- Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., and Rabinovich, A. (2015). Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9.
- Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., and Wojna, Z. (2016). Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826.
- Tan, M., Chen, B., Pang, R., Vasudevan, V., Sandler, M., Howard, A., and Le, Q. V. (2019). Mnasnet: Platform-aware neural architecture search for mobile. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2820–2828.
- Todescato, M. V., Garcia, L. F., Balreira, D. G., and Carbonera, J. L. (2023). Multiscale context features for geological image classification. In Filipe, J., Smialek, M., Brodsky, A., and Hammoudi, S., editors, *Proceedings of the 25th International Conference on Enterprise Information Systems, ICEIS 2023, Volume 1, Prague, Czech Republic, April 24-26, 2023*, pages 407–418. SCITEPRESS.
- Todescato, M. V., Garcia, L. F., Balreira, D. G., and Carbonera, J. L. (2024). Multiscale patch-based feature graphs for image classification. *Expert Systems with Applications*, 235:121116.
- Xie, S., Girshick, R., Dollár, P., Tu, Z., and He, K. (2017). Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1492–1500.
- Zagoruyko, S. and Komodakis, N. (2016). Wide residual networks. *arXiv preprint arXiv:1605.07146*.