Automated Categorization of Wikipedia Images – Image Classification

Matheus V. Bernat
SCIPER: 347152
matheus.vieirabernat@epfl.ch

***Abstract*—Wikipedia is full of articles... and images! Having over 53 million articles in 299 languages containing 11.5 million unique images, there is a great need for automated organization of all this data. Inspired by ORES, an ensemble of machine learning systems in Wikipedia that provides among others automated labeling of articles, this project aims at automated *topic* labeling of images in Wikipedia. In this report, experiments are made using images labeled with the ORES labels of the articles where they are present, and with the custom labels that were generated with a heuristic in the taxonomy part of this semester project. Two different models (EfficientNetB0 and EfficientNetB2) are trained on this data using 10 or 20 labels. As the main insights we understood that: the custom labels were inferior to ORES labels according to our metrics; the network with more parameters, EfficientNetB2, yielded higher prediction values having greater average recall but does not outperform EfficientNEtB0 with regards to the ROC curves; the labels with better performance are those that are most present in the dataset used in pre-training.**

***Index Terms*—Multi-label topic-based image classification, Transfer learning**

I. Introduction

Wikipedia is the largest encyclopedia in history, containing over 53 million articles and having around 1 billion page views per day. Besides text, images play an important role in readers' interaction with Wikipedia articles, as shown in a recent study by Rama et al. [1]. With the number of unique images on Wikipedia surpassing 11 million, labeling these images into broader *topics* (rather than into the specific objects in the image) is becoming increasingly important to tackle tasks such as visual vandalism detection (is the topic of the image related to the topic of the article?), finding visual knowledge gaps (what topics of images is Wikipedia missing the most?), and explanation of reader pattern (do readers interact with images differently depending on the topic of the image?).

Problem formulation. The lack of standard metadata describing the broader picture of images in Wikipedia poses a hinder to the exploration of the full potential of this visual data. There is today no way to perform a *topic*-based search of images, rather than an object-based, where the latter (to our best knowledge) does not exist in Wikipedia but that could easily be implemented using networks pre-trained on standard image datasets.

Prior solutions. To address this problem of automated topic-labeling of Wikipedia images, off-the-shelf networks trained on e.g. ImageNet do not yield satisfactory results due to the variety and uniqueness of images in Wikipedia, as mentioned by Redi in [2]. In the same article, Redi develops a taxonomy of labels by pairing the 6.7 million Commons categories to the 160 COCO [3] categories of visible images and then uses fine-tuning of a deep learning model pre-trained on ImageNet to classify images. This solution, though, still falls short in classifying images in terms of the image topic. Moreover, in [4], Huang sets out to classify chart images of Wikipedia Commons, obtaining the best overall accuracy when fine-tuning an already pre-trained model.

Proposed solution. Our solution to the problem of automated classification of Wikimedia images is to develop a customized taxonomy of topic labels based on the Commons categories (in the work done by Salvi in [5]) and then to fine-tune a pre-trained deep learning model with the Wikipedia image data labeled with the customized topic labels. To be more specific, the deep learning model is given an image and a set of predefined labels and displays a subset of labels that are relevant to describe the image. The assignment of each label is done independently of the others, so the network can be seen as an ensemble of several binary classifiers. Note that the terms *class* and *label* are used as interchangeably as synonyms.

II. Related work

ORES. ORES [6] is an ensemble of machine learning techniques in Wikipedia whose goal is to help editors and content moderators to deal with the immense work of administrating this gigantic encyclopedia. Functions offered by ORES are e.g. vandalism detection, judging article quality, and predicting the topics of an article [7].

WIT dataset. The Wikipedia-based Image Text (WIT) dataset [8] is a large multimodal and multilingual dataset containing 37.6 million image-text entries, with 11.5 million unique images across 108 Wikipedia languages. From the English Wikipedia, 3.9 unique images were gathered by us by reading the segments and removing duplicate images. Each entry contains an image and the textual context in the article where that image is present, and the metadata of the image itself, e.g. caption and image name. In the time scope of this project, only the image data was used.

ImageNet. ImageNet [9] is a dataset of 1.4 million images, each classified with *one single* label out of 1000 possible labels. For over a decade, it has been the benchmark dataset for the training of image classification models.

EfficientNet. EfficientNet [10] is a family of deep learning networks that have been shown to achieve better accuracy

while requiring fewer parameters on ImageNet compared to other convolutional networks. It utilizes a rule for scaling the width, depth, and resolution of the network for better performance.

Transfer learning. Transfer learning is a machine learning method that aims to reuse the knowledge learned in a problem in another similar problem. In the field of image classification, the transferred knowledge is image features such as corners, shapes, and backgrounds. Networks pre-trained on ImageNet are widely used with great success for different reasons, as studied by Huh et al. in [11].

Importance of classifying images in Wikipedia. Images play an important role in understanding and engaging readers, as highlighted in a large body of literature from educational psychology as in [12]. It is not different in Wikipedia; as shown in [13], images coming from Commons have a high monetary and societal value. So, having in mind the value of these images, and also the over 11 million unique images in Wikipedia, it is clear that an automated classification of these according to a standard set of labels is vital for unleashing the potential of the visual content.

III. Method and Data

Method. The method used in this *classification* part of the semester project is to fine-tune a deep learning model pre-trained on ImageNet, and then generate different metrics to assess the quality of the model. By fine-tuning a network, it is meant that the base model's last layer is replaced by a dense layer followed by an output layer of size equal to the number of classes. The weights of the last two layers of this network are then trained, while the other weights are kept unchanged. To assess the quality of the model, the chosen metrics are precision, recall, and receiver operating characteristic area under the curve (ROC AUC).

Data. When it comes to the data, the images coming from the WIT dataset [8] were used, where each image was assigned with a subset of labels starting from the Commons categories of the articles in which the image was present. The finite set of 42 labels was generated by Salvi in a tree-search manner in the first part of this semester's project [5]. See Figure 1 for the number of images per label.

In the scope of this semester's project, the goal is to develop prototypes of the described classifier rather than a fully-packaged solution. Thus, a rather strict pre-processing of the data was made to reduce the training time and the source of possible errors. First, the image data was taken only from the English articles on Wikipedia, which left us with 3.9 million out of the 11.5 unique million images in the WIT dataset. Next, only the 2 million images with a non-empty label set were kept. Finally, only the 1.6 million images of the .jpg and .jpeg formats were kept to avoid problems with the conversion of .png files. In this process, also some other couples of thousands of images were removed, images that were not found for not existing among the downloaded images from the WIT dataset, or images whose names had an encoding unreadable to the operating system.

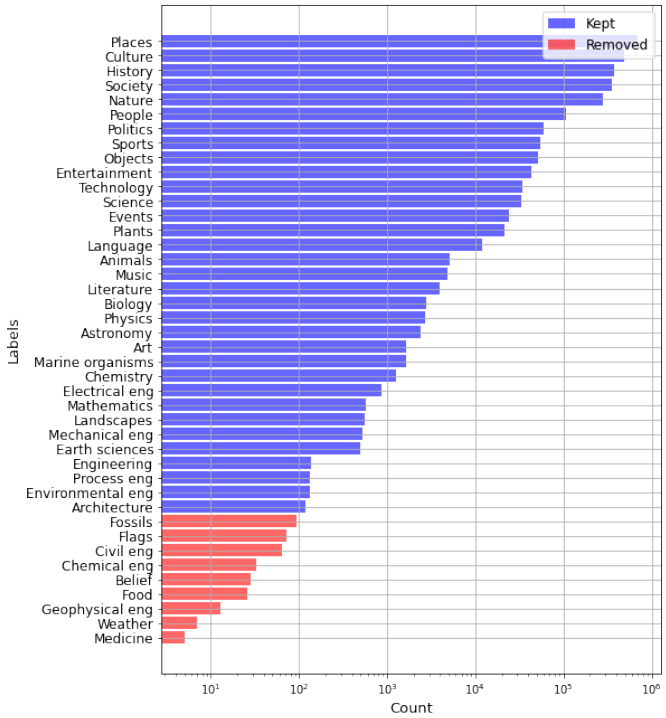

Figure 1. Label distribution through the 1.6M images. Only the labels with more than 100 instances are kept, that is, 33 of them.

IV. Implementation

In this section, more details on the different facets of the implementation itself are covered.

Fine-tuning. The final network used in the experiments had an EfficientNet-based network pre-trained on ImageNet. EfficientNet is the base model, where the last layer is replaced by a dense layer of 128 layers and an output layer with the same number of neurons as the total number of labels. During the fine-tuning, the weights of the base model are left unchanged, so only the weights of the two added layers are updated. See Figure 2 for a scheme of the assembled network.

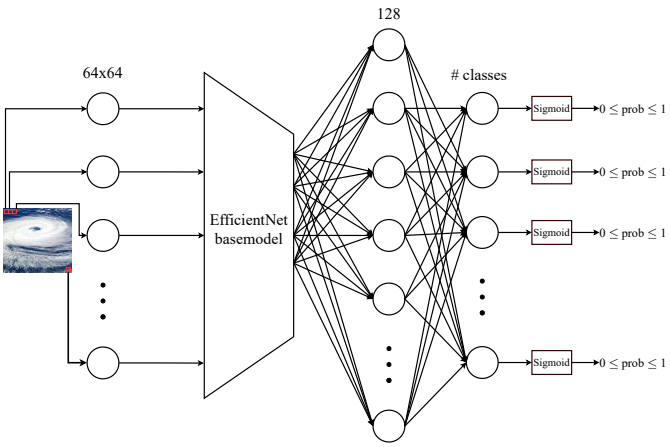

Figure 2. Schema of the used neural network. The EfficientNet based model is either of type B0 (5.3 million parameters) or B2 (9.2 million parameters).

Loss function. The loss function chosen for this *multi-class* image classification problem – where each image can be assigned several labels – was set as the *binary cross-entropy*. The idea is that each label shall be judged as a binary classifier independent of the other labels probabilities. The formula of the loss function is:

$$L(\mathbf{p}) = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{M} y_{ij}\log(p_{ij}),$$

where N is the number of images, M is the number of labels, $y_{ij} \in \{0, 1\}$ is the ground-truth on whether the i^{th} image is labeled with the j^{th} label, and $p_{ij} \in [0, 1]$ is the probability given by the model that the i^{th} image is labeled with the j^{th} label.

Training. The training was performed during 15 epochs, where the class weights were used to compensate for the unbalanced class distribution. Moreover, a decreasing learning rate was tested at an early stage of the project but without any improvement, therefore the learning rate is set to be constant throughout. The image data was all the same with 570 thousand images and evaluating it at 30 thousand images. Each epoch took 30 minutes on average on a machine with 48 cores and 250GB of RAM.

Experiments. To experiment on the performance of the network for the given data and labels, different setups of the network and the labels tested:

- The number of total classes was set to 10, 20;
- Base models EfficientNetB0, EfficientNetB2;
- ORES labels and the labels generated by Salvi.

V. Evaluation

ORES vs. Custom labels. In this first experiment, we want to compare the separability of the ORES labels contra the separability of our custom labels. To do that, the EfficientNetB0-based network was fine-tuned with data labeled with 10 ORES labels, and then with 10 of our custom labels.

For the custom labels, the 10 labels with most images were taken, while for ORES, 10 hand-picked labels out of the top 20 top labels were picked. Note that in this *hand-picking* of 10 the labels from the top 20 classes, we left out the Geography labels specific to a region (e.g. *Geography.Regions.NorthernEurope* and *Geography.Regions.Asia*) and kept the more general *Geography.Geographical*.

See in Table I the evaluation metrics after training the EfficientNetB0-based network where the data all labeled with ORES labels. Then, see in Table II the same metrics for data labeled with our custom labels.

EfficientNetB0 vs. EfficientNetB2. In this second set of experiments, a comparison between the performances of fine-tuning EfficientNetB0 and EfficientB2 is made, when having data labeled with the top 20 custom labels. See Table III for the evaluation metrics of the EfficientNetB0-based network, and Table IV for the EfficientNetB2-based one.

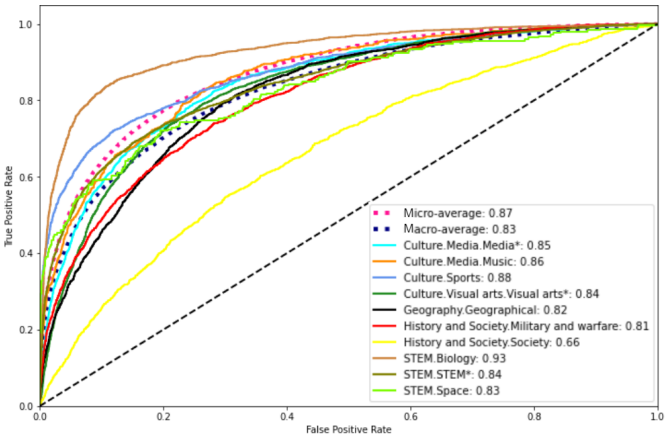

Figure 3. ROC curves for 10 **ORES labels**.

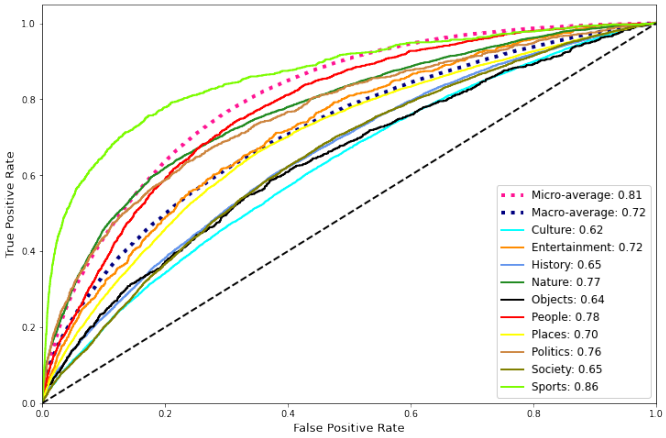

Figure 4. ROC curves for top 10 **custom labels**.

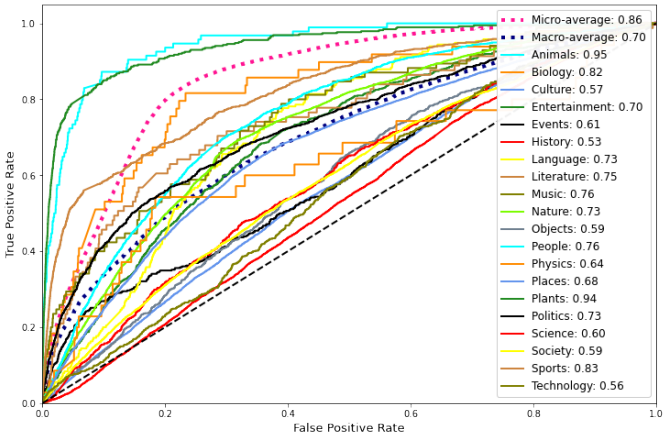

Figure 5. ROC curves for top 20 custom labels, **EfficentNetB0**-based model.

Table I
EVALUATION METRICS WHEN USING ORES LABELS.

	Precision	Recall	ROC AUC
Media	0.58	$\frac{429}{1360} = 0.32$	0.85
Music	0.64	$\frac{124}{614} = 0.20$	0.86
Sports	**0.87**	$\frac{700}{1790} = 0.39$	0.88
Visual arts	0.68	$\frac{1204}{3289} = 0.37$	0.84
Geographical	0.66	$\frac{509}{2267} = 0.23$	0.82
Military and warfare	0.64	$\frac{481}{1924} = 0.25$	0.81
Society	0.18	$\frac{7}{877} = 0.01$	0.66
Biology	0.80	$\frac{1138}{1939} =$ **0.59**	**0.93**
S.T.E.M.	0.81	$\frac{2126}{4203} = 0.51$	0.84
Space	0.85	$\frac{52}{254} = 0.21$	0.83
Micro average	0.74	0.37	0.87
Macro average	0.67	0.31	0.83

Table II
EVALUATION METRICS WHEN USING OUR CUSTOM LABELS.

	Precision	Recall	ROC AUC
Culture	0.64	$\frac{263}{9355} = 0.03$	0.62
Entertainment	0.21	$\frac{11}{795} = 0.01$	0.72
History	0.54	$\frac{511}{7216} = 0.07$	0.65
Nature	0.53	$\frac{1937}{5166} = 0.38$	0.77
Objects	0.16	$\frac{34}{937} = 0.04$	0.64
People	0.60	$\frac{35}{2042} = 0.02$	0.78
Places	**0.66**	$\frac{5558}{13288} =$ **0.42**	0.70
Politics	0.29	$\frac{158}{1074} = 0.15$	0.76
Society	0.52	$\frac{71}{6555} = 0.01$	0.65
Sports	0.45	$\frac{353}{1023} = 0.35$	**0.86**
Micro average	0.59	0.19	0.81
Macro average	0.46	0.15	0.72

VI. DISCUSSION

ORES vs. Custom labels. As can be seen from the comparison between the metrics in Table II and Table I (see Figure 4 and 3 for the ROC curves), the network performs substantially better with ORES-labeled data. The difference is the most remarkable when comparing the average recall: the network trained and evaluated on the custom labeled data yields lower prediction values and is thus more unsure. The reason for this is believed to be the quality of our method to assign the custom labels to the images.

EfficientNetB0 vs EfficientNetB2. Comparing the average recalls in Table III and Table IV, we see that the EfficientNetB2-based model is greater by a factor of 6 (0.19 vs 0.03). This means that the EfficientNetB2-based model yields greater valued predictions and thus surpassing the threshold of 0.5 more times. This phenomenon is also observed by the greater mean number of predicted labels per image (0.26 vs 0.11). Notice though that the average precisions have closer values, which is also confirmed by the very close values of ROC AUCs. This means that the EfficientNetB0-based model needs only a smaller threshold to achieve an average recall similar to the EfficientNetB2-based one.

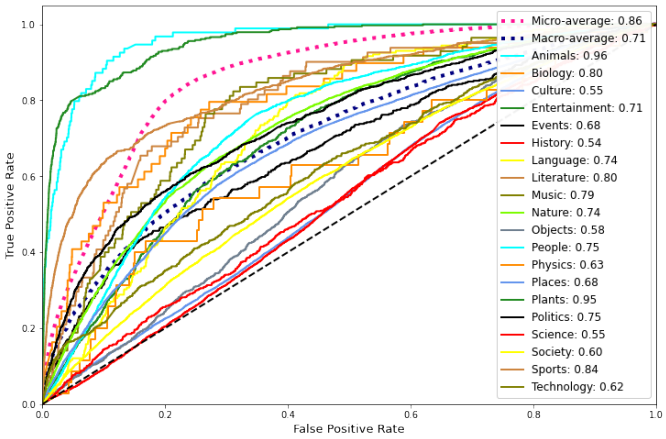

Figure 6. ROC curves for top 20 custom labels, **EfficentNetB2**-based model.

Table III
EVALUATION METRICS FOR CUSTOM LABELS, 20 LABELS, EFFICIENTNETB0. 4.7M TOTAL PARAMETERS, 658K TRAINABLE PARAMETERS. MEAN NUMBER OF PREDICTED LABELS PER IMAGE: 0.11.

	Precision	Recall	ROC AUC
Animals	0.08	$\frac{42}{94} = 0.45$	**0.95**
Biology	0.03	$\frac{3}{49} = 0.06$	0.82
Culture	0.50	$\frac{1}{9355} = 0.00$	0.57
Entertainment	0.00	$\frac{0}{795} = 0.38$	0.70
Events	0.10	$\frac{5}{458} = 0.01$	0.61
History	1.00	$\frac{1}{7216} = 0.00$	0.53
Language	0.00	$\frac{0}{215} = 0.00$	0.73
Literature	0.00	$\frac{0}{81} = 0.00$	0.75
Music	0.07	$\frac{6}{85} = 0.07$	0.76
Nature	0.54	$\frac{105}{5166} = 0.02$	0.73
Objects	**1.00**	$\frac{1}{937} = 0.00$	0.59
People	0.44	$\frac{8}{2042} = 0.00$	0.76
Physics	0.00	$\frac{0}{35} = 0.00$	0.64
Places	0.71	$\frac{1134}{13288} =$**0.9**	0.68
Plants	0.40	$\frac{177}{387} = 0.46$	**0.94**
Politics	0.37	$\frac{38}{1074} = 0.04$	0.73
Science	0.00	$\frac{0}{622} = 0.00$	0.60
Society	0.33	$\frac{1}{6555} = 0.00$	0.59
Sports	0.48	$\frac{131}{1023} = 0.13$	0.83
Technology	0.00	$\frac{1}{675} = 0.00$	0.56
Micro average	0.49	0.03	0.86
Macro average	0.30	0.04	0.70

About the labels. The labels with the greater number of image assignments were expected to have the best performance metrics given the heavily unbalanced dataset, and thus the fact that the network has learned more varied features from these labels. Notice, though, that the labels with the best performance metrics in the top 20 case are Plants and Animals.

Table IV
EVALUATION METRICS FOR CUSTOM LABELS, 20 LABELS, EFFICIENTNETB2. 8.5M TOTAL PARAMETERS, 723K TRAINABLE PARAMETERS. MEAN NUMBER OF PREDICTED LABELS PER IMAGE: 0.26.

	Precision	Recall	ROC AUC
Animals	0.09	$\frac{49}{94} = 0.52$	**0.96**
Biology	0.29	$\frac{2}{49} = 0.04$	0.80
Culture	0.00	$\frac{0}{9355} = 0.00$	0.55
Entertainment	0.00	$\frac{0}{795} = 0.38$	0.71
Events	0.06	$\frac{4}{458} = 0.01$	0.68
History	0.00	$\frac{0}{7216} = 0.00$	0.54
Language	0.00	$\frac{0}{215} = 0.00$	0.74
Literature	0.00	$\frac{0}{81} = 0.00$	0.80
Music	0.00	$\frac{0}{85} = 0.00$	0.79
Nature	0.49	$\frac{210}{5166} = 0.04$	0.74
Objects	0.00	$\frac{0}{937} = 0.00$	0.58
People	0.10	$\frac{4}{2042} = 0.00$	0.75
Physics	0.00	$\frac{0}{35} = 0.00$	0.63
Places	**0.67**	$\frac{3911}{13288} = 0.29$	0.68
Plants	0.40	$\frac{215}{387} = \mathbf{0.56}$	**0.95**
Politics	0.00	$\frac{0}{1074} = 0.00$	0.75
Science	0.00	$\frac{0}{622} = 0.00$	0.55
Society	0.00	$\frac{0}{6555} = 0.00$	0.60
Sports	0.53	$\frac{143}{1023} = 0.14$	0.85
Technology	0.13	$\frac{1}{675} = 0.00$	0.62
Micro average	0.59	0.19	0.86
Macro average	0.14	0.15	0.71

This is believed to be caused by the pre-training: ImageNet has a substantial number of plant and animal classes (reference?). This trend is also observed in the network that uses ORES labels: the Biology label has the greatest ROC AUC.

VII. FUTURE WORK

There are several facets of the image classification part of the project to be further explored.

To begin with, extending the model to be multi-modal – image and text – to also use the text related to the image as model input. An example of text input that can be used is the image name or the image caption. The C-Tran [14] is a model that can be tried. A further study of which kind the textual data in the WIT dataset related to an image (e.g. image name, caption, attribute name, etc) yields the best results would be insightful.

Next, studying the impact of training all network parameters, including those of the base model, would be interesting to discover how this impacts the evaluation metrics. As mentioned before, only the added final two layers' parameters were updated.

Furthermore, being less restrictive with the image filtering to be able to handle also .png files is a necessary extension to be able to have as much data as possible.

REFERENCES

[1] D. Rama, T. Piccardi, M. Redi, and R. Schifanella, "A large scale study of reader interactions with images on wikipedia," *CoRR*, vol. abs/2112.01868, 2021. [Online]. Available: https://arxiv.org/abs/2112.01868

[2] M. Redi, "Prototypes of image classifiers trained on commons categories," https://meta.wikimedia.org/wiki/Research:Prototypes_of_Image_Classifiers_Trained_on_Commons_Categories, 2020.

[3] H. Caesar, J. R. R. Uijlings, and V. Ferrari, "Coco-stuff: Thing and stuff classes in context," *CoRR*, vol. abs/1612.03716, 2016. [Online]. Available: http://arxiv.org/abs/1612.03716

[4] S. Huang, "An image classification tool of wikimedia commons," Master's thesis, Humboldt-Universität zu Berlin, Wirtschaftswissenschaftliche Fakultät, 2020.

[5] F. Salvi, "Taxonomy for automated categorization of wikipedia images," 2022.

[6] A. Halfaker and R. S. Geiger, "ORES: lowering barriers with participatory machine learning in wikipedia," *CoRR*, vol. abs/1909.05189, 2019. [Online]. Available: http://arxiv.org/abs/1909.05189

[7] "Ores."

[8] K. Srinivasan, K. Raman, J. Chen, M. Bendersky, and M. Najork, "WIT: wikipedia-based image text dataset for multimodal multilingual machine learning," *CoRR*, vol. abs/2103.01913, 2021. [Online]. Available: https://arxiv.org/abs/2103.01913

[9] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 248–255.

[10] M. Tan and Q. V. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," 2019. [Online]. Available: https://arxiv.org/abs/1905.11946

[11] M. Huh, P. Agrawal, and A. A. Efros, "What makes imagenet good for transfer learning?" *CoRR*, vol. abs/1608.08614, 2016. [Online]. Available: http://arxiv.org/abs/1608.08614

[12] D. Guo, S. Zhang, K. L. Wright, and E. M. McTigue, "Do you get the picture? a meta-analysis of the effect of graphics on reading comprehension," *AERA Open*, vol. 6, no. 1, p. 2332858420901696, 2020. [Online]. Available: https://doi.org/10.1177/2332858420901696

[13] P. Heald, K. Erickson, and M. Kretschmer, "The valuation of unprotected works: A case study of public domain photographs on wikipedia," *SSRN Electronic Journal*, 01 2015.

[14] J. Lanchantin, T. Wang, V. Ordonez, and Y. Qi, "General multi-label image classification with transformers," *CoRR*, vol. abs/2011.14027, 2020. [Online]. Available: https://arxiv.org/abs/2011.14027