Learning visual features from large-scale datasets - Seminar report

Abdullah Amawi
Supervised by Dr. Timo Lüddecke
Neural Data Science Group
University of Göttingen
Göttingen, Germany
https://eckerlab.org

***Abstract*—In this report we will be exploring some of the recent advancements in the field of computer vision; Specifically dealing with learning visual features from large-scale datasets. Previously, when computer vision models were mentioned and studied, it used to be almost strictly supervised models. Presently, we have a lot of advancements in studying weakly-supervised, unsupervised models, and new techniques being used in supervised learning, such as transfer learning. Those advancements and new techniques are evident especially when studying large-scale datasets, which is the case in our report. The advancements achieved in weakly-supervised and unsupervised learning would save us a lot of time and effort in comparison to supervised models, coupled with learning from large-scale datasets and utilizing transfer learning could lead into leaps in learning visual features, which is what we will inspect in this report.**

***Index Terms*—Computer vision, Transfer learning, Learning visual features, Large-scale datasets.**

I. Introduction

In recent years, computer vision datasets are ever-increasing in size, this trend in datasets growing larger can be beneficial and may result in state-of-the-art results on many different models that deal with various computer vision problems [1]. Moreover, even though we do have major benefit of the increasing size of the datasets to be able to achieve better results, but we are faced with many challenges, such as labeling the images in the dataset when it comes to the usual approach of supervised learning, which requires a lot of time and effort made by human annotators that traditionally had to manually label the images correctly and accurately in order to create such large, high-quality datasets; An example of those datasets is what came to be one of the most known high-quality datasets, is the ImageNet dataset [2].

Even though we have a good amount of publicly available labeled datasets such as the aforementioned ImageNet dataset [2] that can allow us to avoid the daunting task of creating and labeling a new dataset to work on; We are still faced by many other problems that result from labeling itself. One of the major problems that result from manual human selection and annotating usually introduces bias, either as a side effect of the labeling task, or as Joulin et al. mentions that the dataset itself is manually selected to solve a specific task [8], [9]. On the other hand, we also have the idea that in order to get a strong performance in deep learning models, we need a large amount of data for that specific task according to the works of Kolesniko et al [3], which again makes our task seems like problematic in order to scale it up, so we can gain more benefit, which also makes it increasingly expensive computationally.

In order to try to overcome those problems, many works have been done to advance learning visual features from large-scale datasets and how to acquire those datasets, from different techniques, approaches and improvements on the existing approaches. We will review and summarize the ideas and findings in those works. When it comes to supervised learning, we will review the works of Liao et al [6] that suggests good practices on how to annotate large-scale image classifications datasets, and the works of Kolesnikov et al [3] that deals with utilizing transfer learning on pre-trained large supervised datasets. Another very promising approach deals with utilizing weakly supervised data, such as the works of Joulin et al [1], and the works of Mahajan et al [5]. Last and not least, we will focus on a great unsupervised/self-supervised approach that has the potential in totally avoiding previous problems, and that approach is contrasting learning; Both, the works of Radford et al [7] and the works of He et al [10]

II. Semi-supervised & Transfer learning

Studied approaches that deal with supervised learning take two paths, either that they offer practices to annotate large-scale datasets such as in the works of Liao et al [6], or the use of transfer learning to learn from pre-trained supervised datasets as in the works of Kolesnikov et al [3].

The idea of the work done by Liao et al [6] is not only limited to supervised or semi-supervised learning; they also do touch upon the advances in self-supervised learning in addition to the proposed notes towards the practices advice they give that can be utilized in different settings, or for what they call ”efficient human-in-the-loop multi-class labeling” [6]. On the other hand, when it comes to strictly supervised learning and improving the efficiency of human annotation, the works of Liao et al [6] mainly note the following:

- Improving upon prior works to improve the efficiency of time and effort used for labeling.
- Focusing on a fixed worker pool in contrast to using crowdsourcing to avoid additional annotation noise.

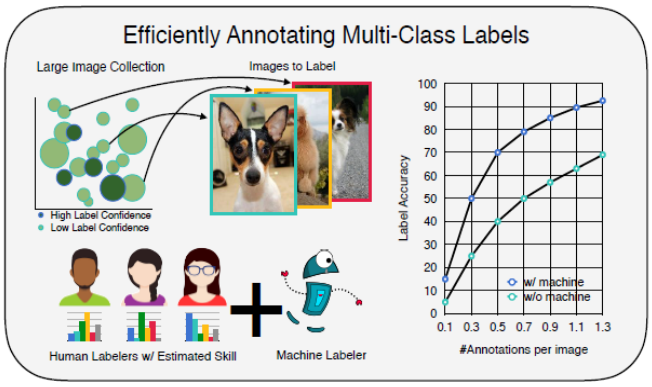

Fig. 1. Liao et al [6] Efficient Anotating w/ & w/o model assistance

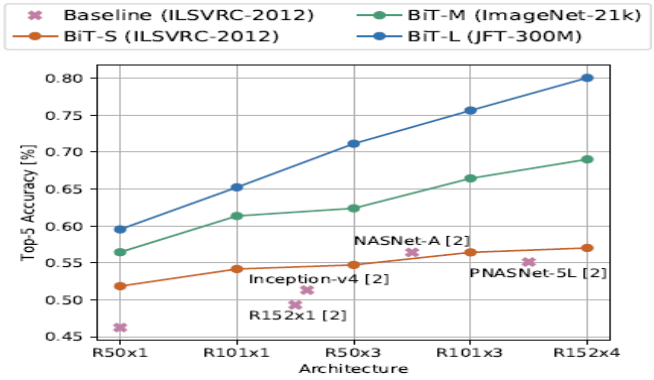

Fig. 2. BiT at different sizes Vs Baseline [3]

Summarizing the improvements that can be done, Fig. 1 demonstrates the efficiency provided by the works of Liao et al [6] including the usage of human labelers with and without the inclusion of model-assisted annotation of multi-class labels at a large scale. Moreover, one of the most important findings that i found in the works of Liao et al [6] is that the reduction of works used for annotations improves the accuracy, specifically going from 50 to 10 workers increased the accuracy to 17% in one of their simulated datasets. lastly, one of their simulations validates that their method works when applied to human workers, and improves upon previous works, requiring only 50% of the annotations with respect to previous work, while achieving 91% accuracy [6].

The second approach we investigate when it comes to supervised learning utilizes a combination of transfer learning and the use of previous supervised datasets; the works of Kolesnikov et al [3] takes the mentioned path.

As we mentioned earlier, there are many available high-quality supervised datasets, and the bigger the dataset is, we can learn more visual features from it, especially if it is a high-quality dataset such as the ImageNet dataset [2]. the works of Kolesnikov et al [3], referred to as "Big Transfer(BiT), utilizes the available supervised datasets in different sizes and not only limited to the high-quality datasets, trained on different sizes, the largest being BiT-L, which is trained on the JFT-300M dataset that has 300 million noisily labelled images [11].

The main idea of the works of kolesnikov et al [3] is that in order to achieve strong performance, we need to utilize a big amount of compute and a task-specific dataset, which makes it expensive, to solve that, they utilized transfer-learning and pre-training, the deep learning network is then trained on a large generic dataset, then the gained weights can be utilized for the new task, which requires less compute costs [3]. It is important to note that "Big Transfer"(BiT) authors claim that they do not introduce complexities but rather a methodology that is minimalist and builds upon the recent improvements in the deep learning network training [3]. Moreover, the authors note that BiT have some major advantages over previous works, which are the following:

- BiT is pre-trained only once then the fine-tuning on the new task is cheap. In contrast other methods are much more expensive due to the expensive training.
- BiT also does not require a lot of hyperparameter tuning on the new task due to the new methods that were validated on the evaluation suite.
- BiT shows strong performance on different scales and was utilized using three different dataset sizes, BiT-S, BiT-M, and BiT-L.

The promising results of BiT shows that it performs strong on many tasks, one of the large tasks that the authors utilized and can be used as an indicator for the performance of BiT is VTAB, which is a test suite that includes 19 tasks [3]. Moreover, the use of BiT-hyperRule outperforms the previous state-of-the-art and according to the authors tests BiT performs better on natural, specialized, and structured tasks [3]. BiT also outperforms the baseline ObjectNet as Fig.2 demonstrates.

Fig.3 demonstrates the performance achieved and the comparison between the current state-of-the-art and BiT-L; Not only that BiT achieves better performance, but it does so while using the single BiT-HyperRule instead of using the previous work("4 HPs"), which of course saves time and compute, which is another reason why BiT is cheap in comparison as indicated earlier.

III. Weakly-Supervised Learning

Another approach is to utilize the very large, ever-growing datasets of platforms such as Flickr and Instagram, we inspected two works that utilize this approach that is based on weakly-supervised learing, the first of which is the work of Joulin et al [1] using Flick as the source of weakly-supervised image datasets, and the second being the work of Mahajan et al [5] using Instagram as the source of the weakly-supervised image datasets. Since we demonstrated in the previous section that large supervised image datasets serve as the base for state-of-the-art models in computer vision, and we mentioned that we may need the datasets to be larger and larger for to improve learning visual features and hence performance; On top of that, that these previous systems are largely based on supervised manually labeled datasets, we can see that it can also be increasingly larger in compute and labeling, which

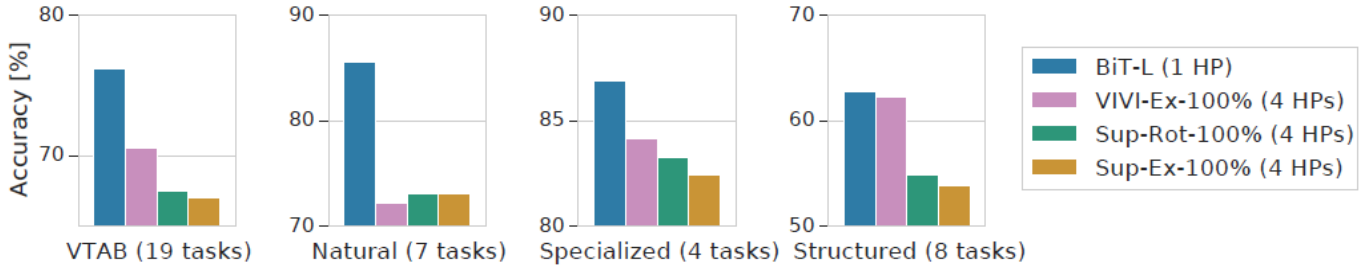

Fig. 3. BiT-L performance Vs SOTA using VTAB suite [3]

makes it expensive, time consuming, non scalable, and may introduce bias towards a specific task [1]. The works of Joulin et al [1] offers a solution by using a dataset of Flickr photos and captions, hence the term "Weakly-Supervised", the authors show that the network utilizing it demonstrates that it captures word similarity, and produces features that performs good in computer vision problems [1].

The authors ask a very important question "can we learn high-quality visual features from scratch without using any fully supervised data?" [1]. In order to answer their own questions, the authors perform their experiment by training images that are from a 100 million Flickr dataset, and those images come with their associated captions, as in Flickr platform as seen in Fig. 4. According to Joulin et al work [1], using weakly supervised data has the following advantages:

- Infinite, ever-growing amount of weakly supervised data is available.
- Training data in this case is not biased towards a specific task.
- Training this way is much more similar to how humans learn to solve computer vision problems.

The paper demonstrates that using weakly supervised image datasets is indeed a valid approach and it can be utilized from scratch, without the need to use manually labeled supervised datasets, and the models also are able to learn semantic structures from the image-word pairs [1]. In our opinion, this result also paved the path for other later works such as the work of Mahajan et al [5].

Mahajan et al [5] work utilizes the same idea of using weakly supervised datasets, but this time using a much larger dataset of 3.5 billion images from Instagram platform. The advantages here are similar to the previous reviewed paper since it is a similar approach, on the other hand, we have the similar disadvantages such that hashtags are noisy, and Mahajan et al [5] work mentions that this way may also harm transfer learning performance. But on the other hand The authors findings confirm that using a larger dataset in pretraining does indeed result in better performance. On the downside, performance seems to be bottlenecked by the model capacity, but on the upside, as model capacity increases, it retains a consistent increase with the larger datasets [5].

Mahajan et al [5] concludes the paper discussion in a couple of points, we see the following as the most important:

Fig. 4. Random images with their associated caption from Flickr [1]

- They found that selecting the label space may be as important as the benefit of increasing the pretraining dataset.
- They observe that current network architectures demonstrate underfitting when trained on such large datasets.
- They found that different models learn better features than others.
- This approach improves classification performance, but on the other hand may harm localization performance.

IV. Contrastive Learning

Last, but not least when it comes to the methods inspected in this report, there is self-supervised learning including contrastive learning. This approach has some of the most interesting ideas to try to overcome the pitfalls of older, more "conservative" approaches such as supervised learning.

Both reviewed works of He et al [10] and Radford et al [7] try to address the need of visual representation learning in large-scale datasets through the use of contrastive learning; Contrastive learning is based on the idea of contrastive loss. Contrastive loss is a method that measures how similar and dissimilar sample pairs are in the representation space; Moreover, contrastive loss measurements can change during training, so it is not a fixed target for the data representations [10]. He et al work [10] utilizes contrastive learning as a dictionary look-up.

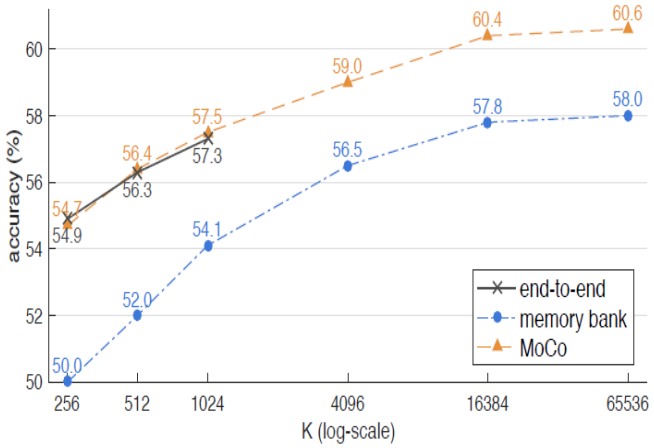

Fig. 5. Contrastive loss mechanisms comparison [10]

When it comes to He et al [10] work, in addition to contrastive learning and its use as dictionary lookup and the contrastive loss function to map the keys as positive or negative, or similar, not similar; The authors noted that they followed the footsteps of previous works [12] by choosing an image and performing a data augmentation on it to form the positive pair. But momentum contrast (MoCo) is where this work [10] shines; The paper denotes that good features can be learned by a large dictionary in contrastive learning and by keeping the dictionary as a queue of the data samples, it allows the decoupling of its size from the mini-batch size, allowing it to become much bigger and not limited by the mini-batch size that is limited by the GPU memory, but this also introduces a problem it makes it unable to manage the keys for back-propagation [10]. To address this, there are a couple of approaches to fix the issue, (1) using end-to-end update which uses the samples in the mini-batch as the dictionary, but the major downside is that it couples it to it, which limits it by the GPU memory size. (2) memory bank which consists of all samples in the dataset, so, it does support a large dictionary size, but on the expense of consistency due to the fact that the representations are updated when they were last seen. Momentum update addresses this by adopting the memory bank method but MoCo does not keep track of every sample, making it more memory-efficient and can be trained on very large datasets, which makes it manageable in the memory bank [10]. Fig. 5 shows the demonstrates MoCo improvements using ImageNet and MoCo contrastive loss mechanism [10].

The final method in contrastive learning and in this report is the work of Radford et al [7], and we will be referring to it as CLIP(Contrastive Language-Image Pre-training) as denoted in the paper. CLIP uses Contrastive pre-training as discussed earlier but with the combination of natural language supervision; CLIP uses a ResNet0-based or ViT-based image encoder, with ViT-based encoder being the larger one. On the other hand, CLIP uses a transformer-based text encoder for the texts that pairs off with the images [7]. CLIP is pre-trained on 400 million image-text pairs selected from the internet, this pre-training is done in a contrastive manner as discussed earlier in He et al [10] work but this time with the corresponding text, which is considering the image and the corresponding text as a similar pair, or positive pair, assuming they are similar since it is paired with it, and on the other hand assuming that the image and texts that not corresponding to the image are dissimilar, or negative pairs. So, CLIP works on three stages (1) the contrastive pre-training as explained (2) creating a dataset classifier from the label text (3) use a zero-shot prediction for the images, as Fig. 6 demonstrates, noting that zero-shot usually refers to images that the model has never seen before, but in the case of CLIP, it denotes it is zero-shot on the dataset level, that it is a dataset that CLIP had not trained on before, and it did not view the dataset in question beforehand, as the authors refer to their use of the term ”zero-shot” [7]. CLIP experiments focused on a couple of aspects, the top of which were zero-shot transfer, as defined by the authors [7], representation learning, and robustness to natural distribution shit. The following points will summarize the findings of the authors [7] for those experiments.

- When it comes to zero-shot transfer, CLIP approach also uses prompt engineering, which adds a prompt in combination of the label text when created in the second step, as shown in Fig. 6 which improves the efficiency gain by four times, and five points improvement in average score. CLIP also has a better score in zero-shot in most datasets when compared to ResNet50, and other similar attempts using a bag of word prediction such as Joulin et al [1].
- In representation learning, CLIP is compared to 10 state-of-the-art vision models on 27 datasets and both the smaller ResNet based CLIP-ResNet and the larger transformer based CLIP-ViT outperform the other SOTA vision models on average score, with CLIP-ViT leading over CLIP-ResNet. Moreover, CLIP was compared on representation learning against EfficientNet L2 NS [13] and won against it in most of the tested datasets by up to 23.6%.
- The final experiment we cover here is the robustness to natural distribution shift which means that deep learning vision models usually are good at finding the patterns and correlations in their training datasets, which means that switching to other distributions will harm the performance, meaning harming the performance on new datasets that we want to use as a downstream task; But CLIP shows very good robustness to the distribution shift on other datasets and performs very good when compared again to the SOTA models and maintains a lead on the dataset suit used in testing. The authors also demonstrated that when they included more datasets with different kinds, meaning that they have very different distributions, CLIP not only maintained the lead and robustness, but also the margin of lead versus other models even increased. When compared to ResNet101 model on ImageNet datasets, CLIP led in zero-shot performance.

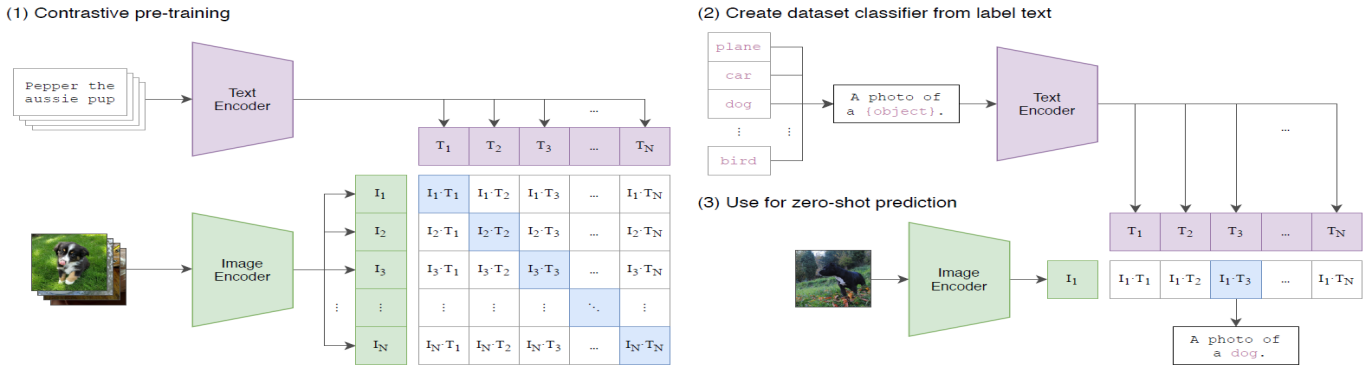

Fig. 6. Three step CLIP approach [7]

V. Discussion

When it comes to learning visual features from large-scale datasets, we believe that different methods still have their upsides and downsides, and one of the current methods still can't be the number one pick in every scenario, therefor, we will list our summary of findings of pros and cons of the studied methods and we think that different recommendations should be matched with different applications of usage. In this section we will provide a summary for each approach studied.

A. Semi-Supervised & Transfer learning

Methods studied here are based on supervised datasets, but since supervised learning on its own is highly unscalable, the authors of [6] resorted to semi-supervised annotations and in [3] the authors resorted to transfer learning, so we will summarize both together as follows:

Advantages:

- Semi-supervised learning reduces the time and effort for annotations significantly, while maintaining very good accuracy levels.
- Approaches based on high-quality supervised datasets still perform very good, especially when there is a good model fine-tuning.
- Very good choice for classification when appropriate upstream dataset is available.

Disadvantages:

- Semi-Supervised learning still deals with much smaller datasets than other approaches.
- Transfer learning can scale supervised learning better, but still not large enough in comparison with weakly-supervised and contrastive learning
- Both approaches are still prone to human mistakes in labeling or mistakes datasets used for pre-training.

B. Weakly-Supervised learning

Weakly-Supervised learning approaches tries to overcome the previous methods resulting in their own advantages and disadvantages as follows:

Advantages:

- Can be trained without any use of manual annotations.
- Highly scalable method with datasets created reaching billions of "labeled" images.
- Used data sources can provide ever-increasing datasets infinitely.

Disadvantages:

- Can be biased and harm the usage of transfer learning.
- Current network architectures may not be the best fit to this approach.
- The usage "labels" from hashtags is very noisy, such as in [1], [5].

C. Contrastive learning

Contrastive learning aims to further avoid the labeling while maintaining very high scalability and learning features which is advantageous, but also comes with its own challenges as follows:

Advantages:

- Contrastive learning totally avoids manual labeling.
- Very high scalability is demonstrated using contrastive learning methods.
- Contrastive learning can learn both visual and textual representations.
- This approach demonstrates a much higher flexibility and generalization in covered task.

Disadvantages:

- Can have weak performance on specific tasks, such as satellite imaging in [7].
- Even though good in covered task, contrastive learning can be very poor in generalization to types of images not covered in pre-training dataset.

VI. Conclusion

We conclude that the studied methods have major advantages and disadvantages and there is no magical recipe to solve every task; Therefor, we believe that we are still in a stage that different methods in learning large-scale visual features still need to be researched further in order to advance them in their respective directions, but while noting that all of the studied works [1], [3], [5]–[7], [10] are promising in their own way and demonstrate great results, but as their own authors noted, there is always room for improvement.

REFERENCES

[1] Joulin, A., van der Maaten, L., Jabri, A., Vasilache, N.: Learning visual features from large weakly supervised data. In: ECCV (2016)

[2] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In 2009 IEEE conference on computer vision and pattern recognition, pages 248–255. Ieee, 2009.

[3] Kolesnikov, A., Beyer, L., Zhai, X., Puigcerver, J., Yung, J., Gelly, S., and Houlsby, N. Large scale learning of general visual representations for transfer. arXiv preprint arXiv:1912.11370, 2019.

[4] Nenad Tomasev and Ioana Bica and Brian McWilliams and Lars Buesing and Razvan Pascanu and Charles Blundell and Jovana Mitrovic. Pushing the limits of self-supervised ResNets: Can we outperform supervised learning without labels on ImageNet. arXiv:2201.05119, 2022.

[5] Mahajan, D., Girshick, R., Ramanathan, V., He, K., Paluri, M., Li, Y., Bharambe, A., and van der Maaten, L. Exploring the limits of weakly supervised pretraining. In Proceedings of the European Conference on Computer Vision (ECCV), pp. 181–196, 2018.

[6] Yuan-Hong Liao and Amlan Kar and Sanja Fidler. Towards Good Practices for Efficiently Annotating Large-Scale Image Classification Datasets. arXiv:2104.12690, 2021.

[7] Alec Radford and Jong Wook Kim and Chris Hallacy and Aditya Ramesh and Gabriel Goh and Sandhini Agarwal and Girish Sastry and Amanda Askell and Pamela Mishkin and Jack Clark and Gretchen Krueger and Ilya Sutskever. Learning Transferable Visual Models From Natural Language Supervision. arXiv:2103.00020, 2021.

[8] J. Ponce, T.L. Berg, M. Everingham, D.A. Forsyth, M. Hebert, S. Lazebnik, M. Marszalek, C. Schmid, B.C. Russell, A. Torralba, C.K.I. Williams, J. Zhang, and A. Zisserman. Dataset issues in object recognition. In Lecture Notes in Computer Science 4170, pages 29–48, 2006.

[9] A. Torralba and A.A. Efros. Unbiased look at dataset bias. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 1521– 1528, 2011.

[10] K. He and H. Fan, and Y. Wu, and S. Xie, R. Girshick. Momentum Contrast for Unsupervised Visual Representation Learning. Facebook AI Research(FAIR).

[11] Sun, C., Shrivastava, A., Singh, S., Gupta, A. Revisiting unreasonable effectiveness of data in deep learning era. In: ICCV (2017)

[12] Zhirong Wu, Yuanjun Xiong, Stella Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In CVPR, 2018. Updated version accessed at: https://arxiv.org/abs/1805.01978v1.

[13] Tan, M. and Le, Q. V. Efficientnet: Rethinking model scaling for convolutional neural networks. arXiv preprint arXiv:1905.11946, 2019.