6.7900 Project Report

Ruiyi Wang
MIT
ruiwangk@mit.edu

Sarah Zhao
MIT
sarahaz@mit.edu

1. Video Link

Our video presentation is available at http://tinyurl.com/67900WangZhao.

2. Summary and contributions

Goyal et al. [Goy+22] is a paper on unsupervised training of a large, dense model on an unfiltered dataset scraped from Instagram. The authors find that an unsupervised model (SEER) with 10 billion dense parameters pre-trained on their Instagram dataset outperforms both unsupervised and supervised models pre-trained on the ImageNet dataset. The authors used the SwAV algorithm for unsupervised training and a scaled-up version of RegNet as the model architecture. The paper's key empirical finding is that the large, unsupervised model was able to capture feature properties that were much less present in the smaller, pre-trained models they used for benchmarking, leading the authors to believe that self-supervised learning on a random dataset leads to more fair, less biased, and less harmful models. Another key finding is that their model is more generalizable and performs much better on out-of-domain tasks and on out-of-distribution data. Interestingly, the authors also found that the model (SEER) was able to recover metadata from images, such as geolocation and semantic information embedded by hashtags (in multiple languages).

3. Strengths

The authors claim that their model has better results on performance, robustness, and fairness benchmarks because their dataset is much larger and more diverse than the traditional ImageNet dataset. The ImageNet dataset contains about 1.5 million images while the Instagram pre-training dataset created by the authors contains over a billion images. The authors also randomly sample 10 million images from their Instagram dataset to empirically evaluate the gender and geographical diversity of their dataset and found relatively equal gender distribution and 192 countries represented (note that EU countries were excluded in compliance with GDPR). Statistically, this is a large sample size with a low margin of error, so it is reasonable to believe that the Instagram pre-training dataset is indeed as diverse as the authors claim.

In the design of their model, the authors leverage many SOTA techniques, such as the RegNet architecture, the SwAV unsupervised training algorithm, and a dynamic programming algorithm to find optimal pre-training activation checkpoints (compute-memory tradeoff). When choosing the final model architecture, the authors experimented with scaling model width, height, resolution, and compound scaling; the variants were trained with 100 million images and validated on ImageNet-1k downstream tasks. Leveraging many SOTA techniques from previous work and experimenting along many axes to choose the final model variant are empirically sound ways to design and choose a model architecture.

In terms of experiments and results, the authors extensively evaluate their pre-trained model on over 50 benchmark tasks, including fairness indicators and several downstream computer vision tasks. The authors compare their model's fairness with two baseline models that use the same RegNet architecture: a supervised model and a self-supervised model pre-trained on the

ImageNet-1k dataset. They evaluate fairness along 3 indicators: disparities in learned representations of people's membership in social groups, harmful mislabeling of images of people, and geographical disparity in object recognition. The authors find that unsupervised pre-training on Instagram data outperforms both baselines significantly on all fairness indicators. They also perform an additional evaluation of multimodal (image and text) hate speech detection and find that their model outperforms the baselines. Taken together, the authors have comprehensively evaluated the fairness of their model and hypothesize that the model performance can be attributed to the diversity of their pre-training dataset, which is consistent with machine learning theory.

The authors are also very detailed in their approach to validating their model performance on transfer learning tasks and use relevant baseline models that allow for fair comparison on both the supervision level of model training and the pre-training dataset of the models. They use seven baseline models pre-trained on ImageNet: one is supervised and six are unsupervised SOTA models with varying architectures. To evaluate robustness, they compare model performance on several out-of-domain datasets; they find that pre-training with Instagram data and fine tuning with ImageNet achieves better performance than pre-training with ImageNet, even on ImageNet-adjacent datasets such as ImageNet-v2. To evaluate performance on fine-grained image recognition, they evaluate model performance on two well-known datasets (iNaturalist18 and iWildCam-WILDS), and find that their model outperforms all seven baselines on both datasets. They use a similar evaluation protocol for image copy detection and find that their model achieves comparable or better performance across 3 datasets. Finally, they evaluate representation learning by benchmarking their model on 25 image classification tasks and find that their model achieves comparable or better performance on all tasks. Overall, the authors are very comprehensive and detailed in their empirical methodology; the use of multiple baselines, multiple datasets, and extensive benchmarking across many relevant indicators and tasks makes their results compelling and their claims believable.

The significance and novelty of the paper is not necessarily the main results, since it is somewhat expected that a larger and more diverse dataset will improve performance on fairness and transfer learning benchmarks; however, the authors claim that both unsupervised pre-trained models and more human-centric pre-training datasets, rather than object-centric pre-training datasets such as ImageNet, will lead to more robust and fair models in the future, which is a pretty novel and bold claim. In the broader ML world, this paper could influence researchers to pursue creating more human-centric pre-training datasets and use unsupervised approaches to pre-training large models. This is especially significant since unsupervised learning did not achieve comparable performance to supervised learning until recently, and this paper claims that unsupervised learning actually outperforms supervised learning on some performance benchmarks. There is also some novelty in how their model is able to capture metadata information, such as geolocation and multilingual word embeddings, from images.

4. Weaknesses

Regarding the claim that unsupervised pre-trained models and more human-centric pre-training datasets lead to more robust and fair models, it is reasonable to assume this follows from a lack of variety within the human-centric subset of the object-centric dataset, due to limited set size of the human-centric subset. After training, biases within this subset become amplified, whereas a full-scale human-centric dataset has a higher likelihood of containing enough diversity to avoid these biases. It would be interesting to see if a smaller (perhaps on a similar scale to ImageNet), but similarly human-centric, pre-training dataset could provide the same benefits as proposed by the authors.

The authors mention adapting their checkpoint positions for trade-offs not accounted for by their dynamic programming checkpoint algorithm. Further details on these trade-offs would be helpful.

Further baseline tests against models of similar parameter size are recommended. Com-

parisons using the dSprites dataset were done with multiple versions of SEER , ranging from 40M to 10B parameters. In particular, the authors provide results from the 693M and 10B parameter models. The benchmark models range from DINO with 85M to BYOL with 250M, while SimCLR-v2 and the various SwAV modes have 585-794M parameters. Furthermore, in Section 4.2.5, the authors describe testing for learning task-agnostic high-quality visual features, and compare SEER to several models with under 100M parameters. In particular, the authors assert that SEER performs competitively with/outperforms SOTA self-supervised models, such as SwAV (RG-128Gf) (Table 8), but given the performance of the 693M parameter SEER model, it appears to be a consequence of model size, rather than the proposed architecture or training dataset.

5. Correctness

The empirical methodology is correct. For each of their evaluation criteria, they use several benchmarks, many baselines, and multiple datasets when relevant, which is a fair and balanced way to evaluate model performance. The claims made by the authors are also correct. For example, claims made by the authors regarding the diversity of their dataset and how this diversity led to less harmful and biased predictions are logical and consistent with machine learning theory. Further claims on the "human-centric" pre-training dataset leading to better performance on "human-centric" downstream tasks and comparable performance on "object-centric" tasks are also logical and consistent with theory.

6. Clarity

Overall, the paper is well-written. In Sections 2 and 3, the authors establish a coherent account of how their motivations interplay with other papers, and there is a clear narrative surrounding how this paper addresses questions raised by the related work and uses previously established methods. It is not immediately obvious that SEER is the model that the authors are proposing, as all references to SEER in Section 3 refer to it as "the model"; a brief exposition on this is recommended.

Some figures could benefit from further explanation and organization, such as Figures 9, 10 and 12. In Figure 9 and 10, we are provided a set of images where the SEER model "demonstrates better performance" than ImageNet-1K pre-trained models, but it is not clear how these images correspond to the challenges described, nor how SEER performs better. A structure similar to Figures 6 and 8 is recommended. Figure 12 is acknowledged as hard to read by the authors, and could benefit from a single close-up figure of a specific embedding grouping, such as the "wedding" hashtag described at the end of Section 5.

7. Relation to Prior Work

The two main contributions of this paper to the literature are its use of a large "in the wild" (randomly selected and unprocessed) images in pre-training and its extensive benchmarking of model performance. Previous works, such as Mathilde et al. [Car+19] and Doersch, Gupta, and Efros [DGE15], explored the use of uncurated and unlabeled datasets in unsupervised learning at a smaller scale and found mixed results; however, this paper utilizes a much larger uncurated, unlabeled dataset in unsupervised pre-training and finds greater success in its results. In terms of performance benchmarking, the most popular evaluation method for visual representation from Zhang, Isola, and Efros [ZIE16] is to train linear classifiers on top of frozen features of ImageNet, although this is seen as somewhat artificial; the authors choose to not use this form of evaluation. The authors instead choose to follow approaches proposed by other papers, such as Radford et al. [Rad+21], that have suggested benchmarking on various datasets to measure model generalization. The authors expand upon this by benchmarking on 50 computer vision tasks and using multiple datasets for each task (30 datasets used in total).

8. Reproducibility

The model documentation is detailed enough that it is feasible for the model to be repro-

duced, although this has not been attempted by the reviewers. The authors describe their algorithm for determining checkpoints in the forward and backward pass, as well as optimizing for training speed. Finally, the authors describe the parameters used in their base RegNet-Y architecture and SwAV method. They also provide details on their compute resources.

Code for the geolocalization test and fairness benchmarks have been released on Github. Regarding the size and scope of the training dataset, this could potentially be difficult to reproduce if one were to adhere to stricter ethics on data collection. In addition, details on data processing methods, such as randomization, tagging, handling duplicates, etc., are unclear.

9. Feedback, Suggestions, and Questions

A suggestion for improvement would be to include a section on ethical and societal implications of SEER (the model created by the authors) as well as more details on the data collection process. Although the authors extensively evaluate the fairness of their model within the body of the paper, it would be a valuable addition to discuss potential positive and negative outcomes of their work in the conclusion of the paper. Some questions we have on the data collection process are: If the Instagram dataset is unfiltered and uncurated, does that mean no forms of content moderation were used? Were images that violate Instagram's Terms of Service excluded from collection?

10. Confidence Score

We rate our confidence in our assessment at level 3. We are not very familiar with the related work, and we did not reproduce the model from scratch.

11. Broader Impact

There was no discussion of broader impact besides statements on how harm reduction and fairness of models are of great societal interest, which is inadequate. The authors might argue that because their model is shown to be more fair than the baseline supervised and unsupervised models, there are no negative potential outcomes from use of their pre-trained models; however, the authors give very little detail on their data collection methods. Their pre-training dataset consisted of one billion public, non-EU Instagram images; although this data collection was probably covered by Instagram's Terms of Service, it is difficult to say that all public and non-EU Instagram users are fully aware and/or fully consent to their images being used for model training. There are societal implications on data scraping methods and potential negative outcomes if consumers are not made aware that their data is being used for training, and it would have improved the quality of the paper if the authors had discussed any of these potential broader impacts of their work in their paper.

12. Potential Ethical Concerns

The paper does not propose any methods, application or data that have a primary purpose of harm or injury. It attempts to address unfair bias posed by previous models and datasets. The reviewers have some concerns regarding data collection, outlined in Section 11. In addition, due to the nature of the dataset source (Instagram), there is many channels for potential hidden biases within the dataset that the authors did not discuss within their paper. These channels include bans or limited access to Instagram in some countries, private versus public accounts, personal versus corporate accounts, etc.

13. Additional Information

Neither reviewer has had prior exposure to this paper, and the reviewers found this paper by Googling and browsing recent work in unsupervised learning vision models. This report will not be used in any other class for either reviewer. Regarding existing reviews, there exists a Medium article consisting of mostly figures taken from the paper, as well as some paper summaries, but no substantial reviews with any technical depth or insight. Ruiyi Wang

wrote sections 1, 2, 4, 6, 8, and 10 and Sarah Zhao wrote sections 3, 5, 7, 9, and 11.

References

[DGE15] Carl Doersch, Abhinav Gupta, and Alexei A. Efros. "Unsupervised Visual Representation Learning by Context Prediction". In: *CoRR* abs/1505.05192 (2015). arXiv: 1505.05192. URL: http://arxiv.org/abs/1505.05192.

[ZIE16] Richard Zhang, Phillip Isola, and Alexei A. Efros. "Colorful Image Colorization". In: *CoRR* abs/1603.08511 (2016). arXiv: 1603.08511. URL: http://arxiv.org/abs/1603.08511.

[Car+19] Mathilde Caron et al. "Leveraging Large-Scale Uncurated Data for Unsupervised Pre-training of Visual Features". In: *CoRR* abs/1905.01278 (2019). arXiv: 1905.01278. URL: http://arxiv.org/abs/1905.01278.

[Rad+21] Alec Radford et al. "Learning Transferable Visual Models From Natural Language Supervision". In: *CoRR* abs/2103.00020 (2021). arXiv: 2103.00020. URL: https://arxiv.org/abs/2103.00020.

[Goy+22] Priya Goyal et al. "Vision Models Are More Robust And Fair When Pretrained On Uncurated Images Without Supervision". In: *ArXiv* abs/2202.08360 (2022). URL: https://api.semanticscholar.org/CorpusID:246904713.