Using Contrastive Learning and Pseudolabels to Learn Representations for Retail Product Image Classification

Muktabh Mayank Srivastava[a]
ParallelDots Inc, Gurugram, India

Keywords: Transfer Learning, Semi Supervised Learning, Few Shot Classification, Retail Product Classification.

Abstract: Retail product Image classification problems are often few shot classification problems, given retail product classes cannot have the type of variations across images like a cat or dog or tree could have. Previous works have shown different methods to finetune Convolutional Neural Networks to achieve better classification accuracy on such datasets. In this work, we try to address the problem statement : Can we pretrain a Convolutional Neural Network backbone which yields good enough representations for retail product images, so that training a simple logistic regression on these representations gives us good classifiers ? We use contrastive learning and pseudolabel based noisy student training to learn representations that get accuracy in order of the effort of finetuning the entire Convnet backbone for retail product image classification.

1 INTRODUCTION

Retail product image classification is a computer vision problem frequently encountered in applications like self checkout stores, retail execution measurement, inventory management and manufacturing. A retail product, for example Nutella jar, will hardly have variations among individuals unlike say the category cat, where each individual looks different, so the expectation in most such problems is to be able to train on a minimal number of images. Common real world retail product recognition datasets are often one shot or few shot classification datasets.

In our previous work, we had proposed methods to finetune Convolutional Neural Network backbones to classify retail product images. However, given retail products have the property of all individuals of a class looking the same and most of the task of Convnets in such classification problems is to remove real world distortions and noise, one might wonder if a Convnet can be trained to create noise invariant image representations that can just be passed through a Logistic Regression or any other simple Machine Learning algorithm to learn recognizing the product. In our work we show that contrastive feature training on a large dataset of image pairs of different retail products [not containing and unrelated to the products we need to train the final classifier on] followed by a noisy pre-

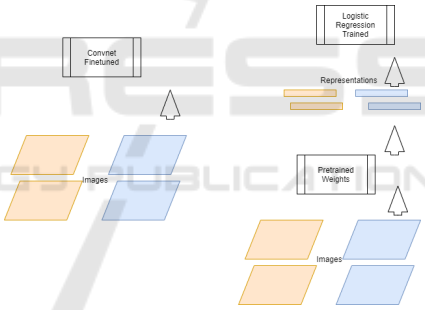

Figure 1: Previous works need to finetune the entire backbone for training model on a retail image classification dataset. In our work using representations of images from a pretrained model we get equivalent or better accuracy by training just a simple Machine Learning classifier.

training on a large dataset of unannotated retail products, we get a Convolutional backbone whose representations can be passed through a simple Logistic Regression model for classification accuracy almost as good as finetuning a Convnet on images of products we need to classify. Figure 1 shows difference between training of between previous works and current method.

2 RELATED WORK

In our previous work, we have proposed different tricks to better the accuracy while finetuning Con-

[a] https://orcid.org/0000-0002-1448-1437

Srivastava, M.
Using Contrastive Learning and Pseudolabels to Learn Representations for Retail Product Image Classification.
DOI: 10.5220/0010911000003124
In *Proceedings of the 17th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2022) - Volume 5: VISAPP*, pages 659-663
ISBN: 978-989-758-555-5; ISSN: 2184-4321

volutional Neural Networks on Retail Product Image Classification. (Srivastava, 2020) We proposed a new layer Local Concepts Accumulation [LCA] layer applied on the output feature map of the Convolutional backbone, which represents an image as a combination of local concepts. There are also published works which finetune GAN-like backbones to recognize Retail Product images using Information Retrieval techniques((Tonioni and Stefano, 2019)). Previously, keypoint matching methods like SIFT ((Lowe, 2004) and (Leutenegger et al., 2011)) have also been used to recognize retail products.

ResNext Convolutional Neural Network backbones ((Xie et al., 2017)) pretrained weakly on instagram hashtags and then finetuned on Imagenet [also called ResNext-WSL] ((Mahajan et al., 2018)) have been shown to get better results on Imagenet and on Retail Product images ((Srivastava, 2020)).

In more recent times, Contrastive Learning learned representations have shown to perform well for Image classification ((Zbontar et al., 2021), (Chen et al., 2020), (Khosla et al., 2020), (Chen and He, 2021)). Even better, these visual representation learners don't require an annotated dataset and can learn by using an image and its augmentation as training pairs for contrastive learning. However, these algorithms require very large unannotated datasets and need to load a lot of images in GPU memory in a single batch to be able to work. SimSiam which tries to optimize these contrastive learning models to bring down the batchsize can make work at batchsize of 256 as opposed to over 4096 of SIMCLR.

Noisy student training where a teacher algorithm is used to generate pseudolabels and a student is trained on these pseudolabels has also been used with great results in Computer Vision problems both image classification and object detection ((Xie et al., 2020) and (Zoph et al., 2020)).

We take the best performing architecture from our experiments in finetuning convnets for retail product image classification which is a ResNext-WSL (Mahajan et al., 2018) with a LCA layer (Srivastava, 2020) and Maximum Entropy loss and try to create a backbone using it which can be used to learn retail product image representations. Because, it is not possible for us to load large batchsizes of even 256 and train for long periods of time, we use supervised contrative learning with hard example mining on a dataset of annotated image pairs to learn features in the first step as a teacher model. This teacher model is used to produce pseudolabels on a large dataset of unannotated retail product images. In the second step of learning representations, we train a student model as a multitask learning model to learn representations.

Figure 2: Samples from TEACHER-PAIRS dataset. This dataset contains product image pairs crawled from internet and annotated by inhouse annotation team.

The two losses in the multitask learning of the student are supervised contrastive loss on an annotated dataset with hard example mining like its teacher and the pseudolabels the teacher algorithm produces on a large unannotated dataset. The representations learnt by both teacher and student are independently analyzed for their performance as input to a Logistic Regression classifier on standard datasets.

3 DATASETS

We first give a description of various datasets used in our work. The first dataset we call TEACHER-PAIRS is an annotated dataset of 250,000 retail product image pairs. This dataset is mined from many other proprietary datasets and crawled from various ecommerce websites. Figure 2 shows some samples from this dataset.

There are no negative annotations, so to train for negative samples, we take random images from outside the pair as a negative sample. A teacher model is trained to learn representations using contrastive loss combined with hard example mining on the TEACHER-PAIRS dataset. The teacher model is then run over 2 Million unannotated retail product images to generate representations of these images which are treated as pseudolabels. The dataset of unannotated images and their corresponding labels is called STUDENT-PSEUDO. Student model is then trained on TEACHER-PAIRS with contrastive

Figure 3: Unannotated samples from STUDENT-PSEUDO dataset.

Figure 4: Train test sample pair from Grozi-120 dataset.

loss and STUDENT-PSEUDO with Smooth L1 loss as Multi-Task Learning. Figure 3 shows some samples from dataset.

The representations learnt by both teacher and student are tested by creating representations of images in classification subsets of Grozi-120 ((Merler et al., 2007) and CAPG-GP (Geng et al., 2018)) datasets and training logistic regression classifier on the representations generated. Both Grozi-120 and CAPG-GP are one-shot datasets. Figures 4 and 5 show sample train-test pairs from Grozi-120 and CAPG-GP dataset respectively.

4 MODELS

As noted earlier there are two models we train to learn visual representation. Both have the same Convolutional architecture, which is a Resnext-101_32X8 architecture. The feature maps of the output of this architecture are passed through a Local Concepts Accumulation layer. Local concepts accumulation (LCA) layer average pools its input feature maps on all rectangular and square sizes larger than 1X1 and creates representations for different local concepts which are then averaged to the representation of the image. The final 2048 dimensional embedding is treated as the representation for the image. LCA layer is same as proposed in our previous work (Srivastava, 2020) and

Figure 5: Train test sample pair from CAPG-GP dataset.

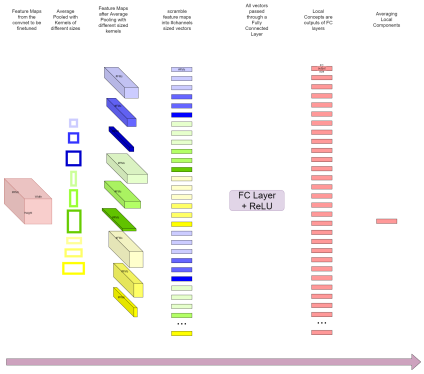

Figure 6: LCA layer is placed on a a ResNext architecture output feature map to create the representation learning backbone.

is shown in Figure 6.

In the first step of training, the model is trained on the TEACHER-PAIRS dataset using a contrastive loss function. We use hard example mining to make sure the features learnt are not too simple. The representations this model produces are called Teacher_Representations. Figure 7 shows training of teacher model.

In the second step of training, the model is trained as a multitask learner on both TEACHER-PAIRS and STUDENT-PSEUDO models. That is, while training, a part of the batch has image pairs from TEACHER-PAIRS and the other part of the batch has images and their representations from STUDENT-PSEUDO. The loss is a weighted average of the contrastive loss on pairs from TEACHER-PAIRS and Smooth L1 loss on STUDENT-PSEUDO representations. The representation from this model is called Student_Representations. Figure 8 shows training of student model.

Now for training classifiers for Grozi-120 and CAPG-GP datasets, we first get representations of dataset images from teacher and student models and then pass these representations through a Logistic Regression model to train for classification. We reemphasize that the teacher and student modules are not finetuned, just used to extract representations here.

5 RESULTS

We compare the accuracy of the simple Logistic regression model trained on both the teacher and the student representations with our best results on finetuning Convnets for retail product image recognition.

From our previous work (Srivastava, 2020), we take the accuracy of finetuning ResNext-WSL (Mahajan et al., 2018), finetuning ResNext-WSL with

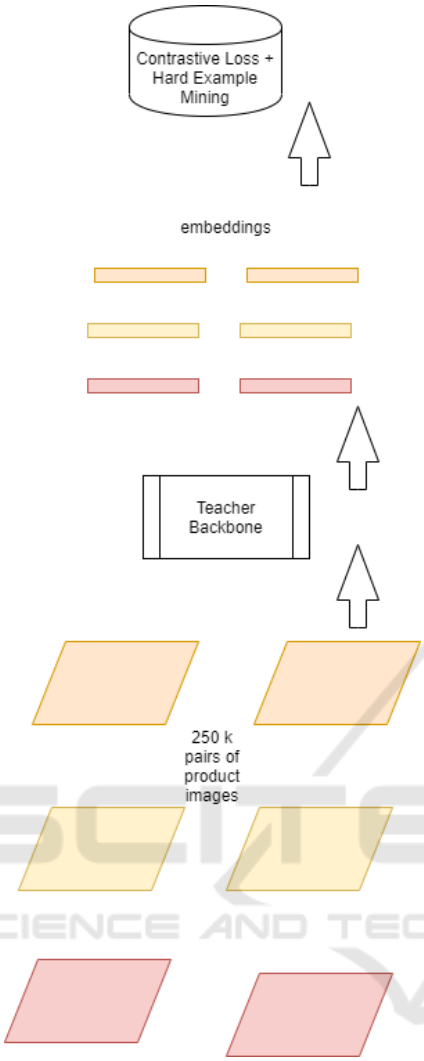

Figure 7: Teacher model is trained on the annotated pairs of TEACHER_PAIRS dataset using contrastive loss and hard example mining. Negative sample for an image is sampled randomly from images outside its pair.

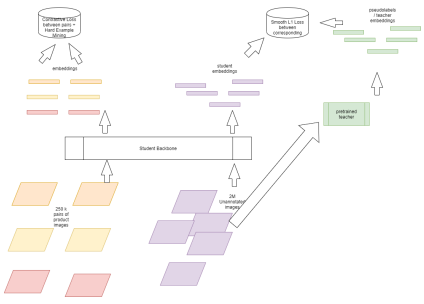

Figure 8: Student model is trained on 250 k pairs using contrastive loss and hard example mining and on pseudolabels generated by teacher models on over 2M images. A batch of student model while training half contains supervised image pairs and other half contains unannotated images and their pseudolabels.

a LCA layer and finetuning a ResNext-WSL with a LCA layer and Maximum Entropy (MaxEnt) loss as an additional loss component as baselines.

From our experiments, we conclude that Logistic Regression (LR) classifiers trained on the representations derived from the features we learn from TEACHER_PAIRS and STUDENT_PSEUDO datasets work quite competitively as compared to finetuning entire Convolutional backbone [Tables 1 and 2]. For Grozi-120 dataset, using pretrained features works much better than finetuning [Table 2].

Table 1: Results of various Models on CAPG-GP Dataset. The first 3 are results when full backbone is finetuned.The 4th and 5th results are results on training a Logisitic Regression (LR) model on the representations yielded by backbones pretrained on TEACHER_PAIRS and STUDENT_PSUEDO dataset respectively.

Model Name	Accuracy [CAPG-GP]
ResNext-WSL	84.1%
ResNext-WSL+LCA layer	90.4%
ResNext-WSL+LCA layer+MaxEnt Loss	92.2%
Teacher_Representations + LR	87.0%
Student_Representations + LR	87.6%

Table 2: Results of various Models on Grozi-120 Dataset. The first 3 are results when full backbone is finetuned.The 4th and 5th results are results on training a Logisitic Regression (LR) model on the representations yielded by backbones pretrained on TEACHER_PAIRS and STUDENT_PSUEDO dataset respectively.

Model Name	Accuracy[Grozi-120]
ResNext-WSL	60.4%
ResNext-WSL + LCA layer	70.8%
ResNext-WSL + LCA layer + MaxEnt Loss	72.3%
Teacher_Representations + LR	75.05%
Student_Representations + LR	76.19%

6 CONCLUSION

We show that a visual representation learner which learns on data annotated on any different datasets or crawled from e-commerce websites, modelled as image pairs and combined with unannotated data can be used to learn image representations which can help

train very simple and yet accurate classifiers. Retail products keep changing in appearance with new packaging and offers. Finetuning a classifier everytime with addition of new products is costly process. A image representations that allows us to just train logistic regression classifier makes accommodating new product additions very simple.

REFERENCES

Chen, T., Kornblith, S., Norouzi, M., and Hinton, G. (2020). A simple framework for contrastive learning of visual representations. *arXiv preprint arXiv:2002.05709*.

Chen, X. and He, K. (2021). Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15750–15758.

Geng, W., Han, F., Lin, J., Zhu, L., Bai, J., Wang, S., He, L., Xiao, Q., and Lai, Z. (2018). Fine-grained grocery product recognition by one-shot learning. pages 1706–1714.

Khosla, P., Teterwak, P., Wang, C., Sarna, A., Tian, Y., Isola, P., Maschinot, A., Liu, C., and Krishnan, D. (2020). Supervised contrastive learning. *arXiv preprint arXiv:2004.11362*.

Leutenegger, S., Chli, M., and Siegwart, R. Y. (2011). Brisk: Binary robust invariant scalable keypoints. In *2011 International Conference on Computer Vision*, pages 2548–2555.

Lowe, D. G. (2004). Distinctive image features from scale-invariant keypoints. volume 60, pages 91–110.

Mahajan, D., Girshick, R., Ramanathan, V., He, K., Paluri, M., Li, Y., Bharambe, A., and van der Maaten, L. (2018). Exploring the limits of weakly supervised pretraining. In Ferrari, V., Hebert, M., Sminchisescu, C., and Weiss, Y., editors, *Computer Vision – ECCV 2018*, pages 185–201, Cham. Springer International Publishing.

Merler, M., Galleguillos, C., and Belongie, S. (2007). Recognizing groceries in situ using in vitro training data. In *2007 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8.

Srivastava, M. M. (2020). Bag of tricks for retail product image classification. In Campilho, A., Karray, F., and Wang, Z., editors, *Image Analysis and Recognition*, pages 71–82, Cham. Springer International Publishing.

Tonioni, A. and Stefano, L. D. (2019). Domain invariant hierarchical embedding for grocery products recognition. *Computer Vision and Image Understanding*, 182:81–92.

Xie, Q., Luong, M.-T., Hovy, E., and Le, Q. V. (2020). Self-training with noisy student improves imagenet classification.

Xie, S., Girshick, R., Dollar, P., Tu, Z., and He, K. (2017). Aggregated residual transformations for deep neural networks. pages 5987–5995.

Zbontar, J., Jing, L., Misra, I., Lecun, Y., and Deny, S. (2021). Barlow twins: Self-supervised learning via redundancy reduction. In Meila, M. and Zhang, T., editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 12310–12320. PMLR.

Zoph, B., Ghiasi, G., Lin, T.-Y., Cui, Y., Liu, H., Cubuk, E. D., and Le, Q. V. (2020). Rethinking pre-training and self-training.