SEMANTIC SIMILARITY BASED LABEL AUGMENTATION FOR VISUAL CLASSIFICATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Real-world applications may present visual categories for which examples are many but a definition is elusive. When data augmentation helps little and hand-crafted heuristics fail to warrant weak supervision, similarity remains a simple but effective guide for augmenting training labels. This paper showcases, for recognizing clutters in e-commerce photography, how similarity between visual representations pre-trained to capture semantics provides quality supervision. Our approach significantly improves on the texture analysis baseline, with classification precision and recall above 0.8 in most cases.

1 INTRODUCTION

Besides their success in zero-shot transfer, language-image models with contrastive pre-training like CLIP (Radford et al., 2021) allow us to measure text-image similarity and thereby *semantic* similarity between images themselves. This similarity can prove quite useful in label augmentation for visual classification. As a case study, we consider clutter recognition in e-commerce photography, where we have only a handful of labeled photos and a myriad of unlabeled ones.

Clutters can be *anything* adding no values but distracting customers from viewing the product; see Figure 1 for examples. The concept, compared to usual object categories like dogs or cars, is vague enough to defy common image augmentation techniques, which are unlikely to deliver a diversity of examples naturally seen in reality.

Nor do we have working heuristics for weak supervision (WS). A texture descriptor like standard deviation (Gonzalez, 2018) may somewhat predict $-$CLUTTER, but its precision with $+$CLUTTER is below random guess, as it ignores higher-level semantics.

2 RELATED WORK

Data augmentation For visual classification, data augmentation synthesizes new images out of existing labeled examples and retains their labels. Shorten & Khoshgoftaar (2019) offer a thorough survey on rule-based (e.g., geometric transformations) and learned (Zhu et al., 2017; Li et al., 2023) image augmentation techniques for deep learning models.

Label augmentation Label augmentation generates noisy labels for unlabeled data, sometimes by exploiting a small number of ground truths. The idea underlies WS, where several heuristics collaborate in label generation (Ratner et al., 2016; 2017). Apart from the recent progress in data programming for WS reviewed in Solmaz et al. (2022), the latter itself proposes to extend the coverage of labeling functions based on similarities among data points.

3 APPROACH

We perform neighbor search with Faiss (Johnson et al., 2019) to iteratively grow the classes $\mathbb{P}$ and $\mathbb{N}$ of $\pm$CLUTTER images. After embedding the universe of all labeled and unlabeled images with CLIP[1] and holding out a labeled subset $\mathbb{H}$, we set $\mathbb{P}_0$ as $+$CLUTTER seeds and $\mathbb{N}_0$ as $-$CLUTTER

[1]Implemented at `https://github.com/huggingface/transformers`.

seeds. Let $n_k(\boldsymbol{x})$ be the set of k nearest neighbors of $\boldsymbol{x}$ other than $\boldsymbol{x}$ in the universe; for $i \geq 1$,

$$\begin{aligned}
\mathbb{P}'_i &= \bigcup_{\boldsymbol{x} \in \mathbb{P}_{i-1}} n_k(\boldsymbol{x}) \setminus \bigcup_{0 \leq j < i} \mathbb{P}_j \setminus \bigcup_{0 \leq j < i} \mathbb{N}_j, \\
\mathbb{N}'_i &= \bigcup_{\boldsymbol{x} \in \mathbb{N}_{i-1}} n_k(\boldsymbol{x}) \setminus \bigcup_{0 \leq j < i} \mathbb{P}_j \setminus \bigcup_{0 \leq j < i} \mathbb{N}_j, \\
\mathbb{P}_i &= \mathbb{P}'_i \setminus \mathbb{N}'_i \setminus \mathbb{H}, \\
\mathbb{N}_i &= \mathbb{N}'_i \setminus \mathbb{P}'_i \setminus \mathbb{H}.
\end{aligned}$$

Iteration stops as soon as $|\mathbb{P}_{i-1}| < |\mathbb{P}_i|$ or $|\mathbb{N}_{i-1}| < |\mathbb{N}_i|$, since the search henceforth will return increasingly more points that have appeared before. The union of all iterations $\mathbb{P}_0$, ..., $\mathbb{P}_i$ gives the augmented class $\mathbb{P}$ of $+$CLUTTER images. $\mathbb{N}$ is similarly obtained.

Here the hyperparameter k and the iteration stopping criterion give us some control over the similarity between seeds and sampled neighbors; taking the neighbors of neighbors into account contributes to the diversity; throughout the process we have kept $\mathbb{P}$ and $\mathbb{N}$ disjoint.

4 EXPERIMENTS

We validated the quality of similarity-augmented labels with three learning experiments, resp. for clutter recognition in the *Home & Garden* (e.g. sofas), *Fashion* (e.g. sneakers), and *Electronics* (e.g. laptops) departments.

Data Of each department 10K images were labeled by 3 up to 7 workers after due instruction on clutters; a total of 125,284 judgments were collected with an agreement rate above 0.8 more than 70% of time. A mean opinion score (MSO) $\in$ [0,1] arises from summing the responses *Yes* = 1, *No* = 0, *Unsure* = 0.5 with the workers' credibility as weights.

A MOS $\geq$ 0.5 reads $+$CLUTTER; otherwise $-$CLUTTER if all workers agreed the image showed exactly one product. We hold out 30% of our labeled examples for validation and make the rest the seeds of label augmentation, with $k = 10$. Table 2 summarizes the scale of our datasets.

Classification We added on top of frozen CLIP a single dense layer and trained this classifier with and without similarity-augmented labels. As Table 1 shows, the results by far outperform the baseline, optimal[2] thresholding on the mean standard deviation of the background (given by salient object detection (Qin et al., 2020)). Training with extra labels may only have a thin edge on training without, but that was expected as either way validation loss quickly matched training loss.

Table 1: Validation of classification vs. baseline per $\pm$(CLUTTER) **P**(recision) and **R**(ecall).

	Home & Garden				**Fashion**				**Electronics**			
	+P	+R	-P	-R	+P	+R	-P	-R	+P	+R	-P	-R
Augmented training	.80	**.69**	**.91**	.95	.82	**.60**	**.91**	**.97**	.79	**.90**	**.90**	.80
Seed training	**.85**	.64	.90	**.97**	**.83**	.58	.90	**.97**	**.81**	.86	.88	**.83**
Baseline	.18	.58	.61	.20	.09	.34	.48	.16	.23	.27	.27	.23

5 CONCLUSION

We have shown that semantic similarity can be directly used to generate quality labels. While our method is generally applicable to novel visual categories, future work is needed to explore how larger-scale augmentation—comparable to the scale of typical datasets on which modern computer vision networks are trained—may benefit the generalization of models deeper than ours.

[2]Maximizing the harmonic mean of the four reported metrics.

REFERENCES

Rafael C. Gonzalez. *Digital image processing*. Pearson Education, 4th edition, 2018.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, 7(3):535–547, 2019.

Bohan Li, Xinghao Wang, Xiao Xu, Yutai Hou, Yunlong Feng, Feng Wang, and Wanxiang Che. Semantic-guided image augmentation with pre-trained models. *arXiv:2302.02070*, 2023.

Xuebin Qin, Zichen Zhang, Chenyang Huang, Masood Dehghan, Osmar R Zaiane, and Martin Jagersand. U2-net: Going deeper with nested u-structure for salient object detection. *Pattern recognition*, 106:107404, 2020.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PMLR, 2021.

Alexander Ratner, Stephen H Bach, Henry Ehrenberg, Jason Fries, Sen Wu, and Christopher Ré. Snorkel: Rapid training data creation with weak supervision. In *Proceedings of the VLDB Endowment. International Conference on Very Large Data Bases*, volume 11, pp. 269. NIH Public Access, 2017.

Alexander J Ratner, Christopher M De Sa, Sen Wu, Daniel Selsam, and Christopher Ré. Data programming: Creating large training sets, quickly. *Advances in neural information processing systems*, 29, 2016.

Connor Shorten and Taghi M Khoshgoftaar. A survey on image data augmentation for deep learning. *Journal of big data*, 6(1):1–48, 2019.

Gürkan Solmaz, Flavio Cirillo, Fabio Maresca, and Anagha Gode Anil Kumar. Label augmentation with reinforced labeling for weak supervision. *arXiv preprint arXiv:2204.06436*, 2022.

Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pp. 2223–2232, 2017.

A APPENDIX

Table 2: The scale of the ±CLUTTER seed, augmented, and validation sets.

	Home & Garden		**Fashion**		**Electronics**	
	+CLUTTER	−CLUTTER	+CLUTTER	−CLUTTER	+CLUTTER	−CLUTTER
Seeds	1.3K	4.2K	1.2K	4.8K	2.8K	3.3K
Augmented	19.6K	54.3K	50.4K	170.8K	70.0K	76.5K
Validation	.5K	1.8K	.5K	2.1K	1.2K	1.4K

+CLUTTER

Clutter due to mirror reflection and human model.

+CLUTTER

The prop iPhone blocks and isn't sold with the item (purse).

+CLUTTER

Clutter due to artificial texts and artworks.

+CLUTTER

Clutter due to photo collage.

−CLUTTER

Staging the item (bed frame) in a cozy bedroom adds values.

−CLUTTER

The complex pattern belongs to the item.

−CLUTTER

The complex pattern belongs to the item's container.

−CLUTTER

Plain background.

Figure 1: Examples of cluttered (+CLUTTER) and non-cluttered (−CLUTTER) e-commerce images.