SEEING IN WORDS: LEARNING TO CLASSIFY THROUGH LANGUAGE BOTTLENECKS

Khalid Saifullah[1], Yuxin Wen[1], Jonas Geiping[1], Micah Goldblum[2], and Tom Goldstein[1]
[1]University of Maryland, [2]New York University
{khalids, ywen, jgeiping, tomg}@umd.edu, goldblum@nyu.edu

ABSTRACT

Neural networks for computer vision extract uninterpretable features despite achieving high accuracy on benchmarks. In contrast, humans can explain their predictions using succinct and intuitive descriptions. To incorporate explainability into neural networks, we train a vision model whose feature representations are text. We show that such a model can effectively classify ImageNet images, and we discuss the challenges we encountered when training it.

1 INTRODUCTION

In recent years, there has been a surge of interest in vision-language models (VLMs) that combine the power of computer vision and natural language processing to perform tasks such as image captioning, visual question answering, and image retrieval (Alayrac et al., 2022; Radford et al., 2021; Li et al., 2022b; Wang et al., 2022; Zeng et al., 2021; Singh et al., 2022). These models leverage both visual and textual signals to reason about their inputs and generate meaningful outputs (Li et al., 2022a; Xu et al., 2015; Anderson et al., 2018; Li et al., 2019; Zhou et al., 2020; Li et al., 2020).

One popular approach to building VLMs is through self-supervised learning (SSL), which involves training a model to make predictions about a given input without any human-labeled annotations. SSL has shown great promise in achieving state-of-the-art performance on various tasks in computer vision and natural language processing (Balestriero et al., 2023; Devlin et al., 2018).

Prior work has explored generating text descriptions from images using a variety of approaches. (Liu et al., 2023) encode images into text tokens using a pretrained codebook, but their generated text may lack semantic meaning. (Wickramanayake et al., 2021) use CNNs to associate visual features with human-annotated word phrases but require a manual definition of those phrases. Our work differs by employing an image-grounded language model decoder, eliminating the need for a codebook. This allows us to generate text descriptions that are semantically meaningful without relying on predefined word phrases.

In this paper, we take a stab at implanting a language bottleneck in traditional image classification pipelines (see Figure 1). By converting image features into words and using the words to classify the image, our proposed method can provide insights into the interpretability of classification models, as the language bottleneck serves as a "universal interface" between the visual and textual modalities. Extracting human-readable language features can also help us better understand how these models learn and reason about the content of images.

2 METHOD

Our goal is to create an image classifier that uses text (rather than continuous-valued "features") as an intermediately representation. Ideally, this representation should be a detailed description of the image contents. These descriptive tokens are generated by a feature extractor, and then fed into a classification layer that outputs a final image label. By training this system in an end-to-end way, we learn a useful text-based description of each image using only class-level (rather than caption-level) supervision.

Table 1: Validation Results on ImageNet.

Method	ImageNet	+Gaussian	+Impulse	+Shot	+Defocus
BLIP Caption	42.819	40.185	38.874	39.866	39.147
Ours	67.117	64.799	63.201	64.784	64.287
+ Token similarity	68.894	66.444	64.961	66.466	65.774
+ LLM loss	64.035	62.162	60.89	62.195	61.602
+ No repetition sampling	62.050	59.935	58.39	59.952	59.721

We build this system off BLIP (Li et al., 2022b), a fine-tuned image-to-text caption model as the basis of our pipeline. The model has two modules, a visual transformer (Dosovitskiy et al., 2020), which transforms input images into embedding vectors, and a language model (Devlin et al., 2018) that generates hard tokens by incorporating the signals from image embeddings with the help of the cross-attention layer.

The feature extractor of our system generates a sequence of tokens, but the process of multi-token sampling is not directly differentiable. To ensure that the pipeline remains end-to-end differentiable, we feed n trainable soft prompts into the text encoder instead of sampling to generate hard tokens. After that, the decoder produces n logits, each one representing the next-token prediction for a word in the prompt. We softmax the predictions and perform a matrix-matrix multiplication with the word embedding matrix. This multiplication results in n word embeddings for the image. To obtain a single vector, we mean pool the n word embedding vectors. Finally, we pass the pooled vector through a linear classification head to predict the class.

Note that we only train the soft prompt and the linear head parameters. Also, during validation, we use argmax on the logits to retrieve the word embeddings as hard tokens. As a result, the linear head only "sees the words" to make its predictions.

Training such a model yields a challenging optimization problem and often leads to a model that mostly outputs non-human-readable text and repeated words. Therefore, we also design three variants to produce more diverse and human-readable image descriptions:

Token similarity loss: We compute the cosine similarity between all token pairs in the sequence, then calculate the average to get the sequence word similarity. This number tells us how much the generated tokens are similar to each other, and we minimize this loss to get more diverse tokens in the sequence.

LLM loss: In order to get human-readable text, we feed the generated tokens that we get from our language bottleneck through the language model again to get the likelihood of those tokens (bad text should produce low likelihood).

No repetition sampling: This is a sampling procedure we use during inference, where we perform auto-regressive sampling but skip the tokens that were already generated in the sequence.

3 RESULTS

We test our pipeline on ImageNet (Deng et al., 2009) as well as ImageNet test sets with common corruptions (Hendrycks & Dietterich, 2019). We train the soft prompt and linear head for 5 epochs with learning rate $1e^{-1}$ and $5e^{-3}$ respectively. A natural candidate for a baseline is a system that uses the caption from the fine-tuned BLIP model, the motivation behind this is that BLIP has learned through the supervision of human captions on images, so it describes images the way a human might. But as the results show in Table 1, we do not get the optimal classification results with this baseline. Our method substantially improves the baseline accuracy. Meanwhile, adding token similarity loss produces the best validation accuracy. From our manual evaluation, training with token similarity loss drives the model to produce text tokens that have the best balance of being human-readable and helpful for the classifier. We provide sample generations in Appendix Figure 2.

4 CONCLUSIONS

In this paper, we proposed a method for implanting a language bottleneck in traditional image classification pipelines and investigated the performance of such a model. We found that incorporating language into the pipeline produces non-trivial classification accuracy on the ImageNet dataset. Our results suggest that language can serve as a universal interface and models can learn to express visual features through words alone. This kind of pipeline may provide insights into the interpretability and performance of vision-language models. Future work might explore better ways to make the language bottleneck produce more salient and human-readable words, and use this kind of pipeline to experiment on out-of-domain image samples. It would also be valuable to investigate how this approach can handle more complex images with multiple objects or scenes, and to consider its implications for model transparency and accountability.

5 ACKNOWLEDGEMENTS

This work was made possible by the ONR MURI program, DARPA GARD (HR00112020007), the Office of Naval Research (N000142112557), and the AFOSR MURI program. Commercial support was provided by Capital One Bank, the Amazon Research Award program, and Open Philanthropy. Further support was provided by the National Science Foundation (IIS-2212182), and by the NSF TRAILS Institute (2229885).

REFERENCES

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *arXiv preprint arXiv:2204.14198*, 2022.

Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 6077–6086, 2018.

Randall Balestriero, Mark Ibrahim, Vlad Sobal, Ari Morcos, Shashank Shekhar, Tom Goldstein, Florian Bordes, Adrien Bardes, Gregoire Mialon, Yuandong Tian, et al. A cookbook of self-supervised learning. *arXiv preprint arXiv:2304.12210*, 2023.

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*, pp. 248–255, 2009. doi: 10.1109/CVPR.2009.5206848.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. *arXiv preprint arXiv:1903.12261*, 2019.

Boyi Li, Kilian Q Weinberger, Serge Belongie, Vladlen Koltun, and Rene Ranftl. Language-driven semantic segmentation. In *International Conference on Learning Representations*, 2022a. URL `https://openreview.net/forum?id=RriDjddCLN`.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, pp. 12888–12900. PMLR, 2022b.

Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*, 2019.

Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXX 16*, pp. 121–137. Springer, 2020.

Hao Liu, Wilson Yan, and Pieter Abbeel. Language quantized autoencoders: Towards unsupervised text-image alignment. *arXiv preprint arXiv:2302.00902*, 2023.

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.

Amanpreet Singh, Ronghang Hu, Vedanuj Goswami, Guillaume Couairon, Wojciech Galuba, Marcus Rohrbach, and Douwe Kiela. Flava: A foundational language and vision alignment model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15638–15650, 2022.

Jianfeng Wang, Zhengyuan Yang, Xiaowei Hu, Linjie Li, Kevin Lin, Zhe Gan, Zicheng Liu, Ce Liu, and Lijuan Wang. Git: A generative image-to-text transformer for vision and language. *arXiv preprint arXiv:2205.14100*, 2022.

Sandareka Wickramanayake, Wynne Hsu, and Mong Li Lee. Comprehensible convolutional neural networks via guided concept learning. In *2021 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8. IEEE, 2021.

Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, pp. 2048–2057. PMLR, 2015.

Yan Zeng, Xinsong Zhang, and Hang Li. Multi-grained vision language pre-training: Aligning texts with visual concepts. *arXiv preprint arXiv:2111.08276*, 2021.

Luowei Zhou, Hamid Palangi, Lei Zhang, Houdong Hu, Jason Corso, and Jianfeng Gao. Unified vision-language pre-training for image captioning and vqa. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pp. 13041–13049, 2020.

A Appendix

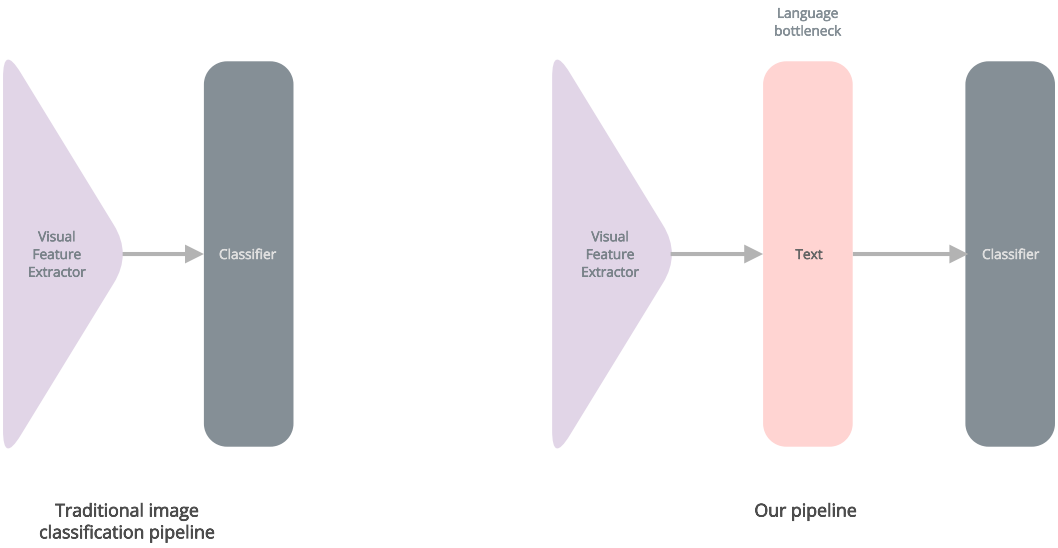

Figure 1: Architecture

Image	BLIP Caption	Ours	+With No Repetition	Label
	a picture of there are three people standing together in a lab	three fish oyster aquarium aquarium oyster oyster oyster	three a seafood oyster shell table aba tray bowl	lab coat
	a picture of a bunch of kids sitting down in front of a bass	there musician trombone trombone trombone	there a music trombone trumpets chair brass trumpet	trombone
	a picture of a dog is sitting in the grass with a frisbee	there dog span span bern bern bern bern bern	there a black english span dog border hound puppy	English springer
	a picture of a man in military gear using a machine gun	rifle gun rifle rifle rifle rifle rifle rifle rifle	rifle a military army ak gun ar machine man	rifle
	a picture of a totemologist is sitting on the back of a	there bird to to statue to to to pole	there a native to hai pole thunder statue hook	totem pole

Figure 2: Sample Generations