SEEING IN WORDS: LEARNING TO CLASSIFY THROUGH LANGUAGE BOTTLENECK

Anonymous authors
Paper under double-blind review

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., 2022; Wang et al., 2022). These models leverage both visual and textual signals to reason about their contents and generate meaningful outputs.

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.

In this paper, we take the first stab at exploring the idea of implanting a language bottleneck in traditional image classification pipelines. By converting image features into words and using the words to classify the image, we aim to investigate the performance of such a model. This approach is analogous to giving humans rich object descriptions and asking them to identify the object without looking, thereby exploring the role of language in visual reasoning.

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. It can also help us better understand how these models learn and reason about the content of images.

In summary, our work aims to explore the potential of language as a bridge between computer vision and natural language processing and how it can be used to improve the interpretability and performance of image classification models.

2 METHOD

We use BLIP (Li et al., 2022), 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. Subsequently, a language model (Devlin et al., 2018) generates hard tokens by incorporating the signals from image embeddings with the help of the cross-attention layer.

To ensure that the pipeline remains end-to-end and differentiable, we feed n trainable soft prompts as gold input to the text encoder instead of sampling to generate hard tokens. After that, the decoder produces n logits, we normalize them with softmax and perform a matrix-matrix multiplication with the word embedding matrix. This 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.

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 repitition sampling	62.050	59.935	58.39	59.952	59.721

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

This is a very difficult optimization problem that arises the issue of the model mostly generating non-human-readable texts, and repeating words. Therefore, based on the pipe, we also test out three variants to get more diverse image descriptions and make the text more aligned with the classifier's preference:

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 texts, 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 language text should produce low likelihood).

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

3 RESULTS

We have experimented with the "seeing in words" concept on normal ImageNet (Deng et al., 2009) and with the one with corruptions (Hendrycks & Dietterich, 2019). We train the soft prompt and linear head for 5 epochs with learning rate of $1e^{-1}$ and $5e^{-3}$ respectively. A natural candidate for our baseline is a system that uses the caption from the finetuned 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 would. But as the results show in Table 1, we don't get the optimal classification results from that. Our method can substantially improve 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 some 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 see images through words alone. This kind of pipeline may provide insights into the interpretability and performance of vision-language models. Future work could 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.

URM STATEMENT

The authors acknowledge that at least one key author of this work meets the URM criteria of ICLR 2023 Tiny Papers Track.

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.

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.

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, 2022.

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.

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.

A APPENDIX

B APPENDIX

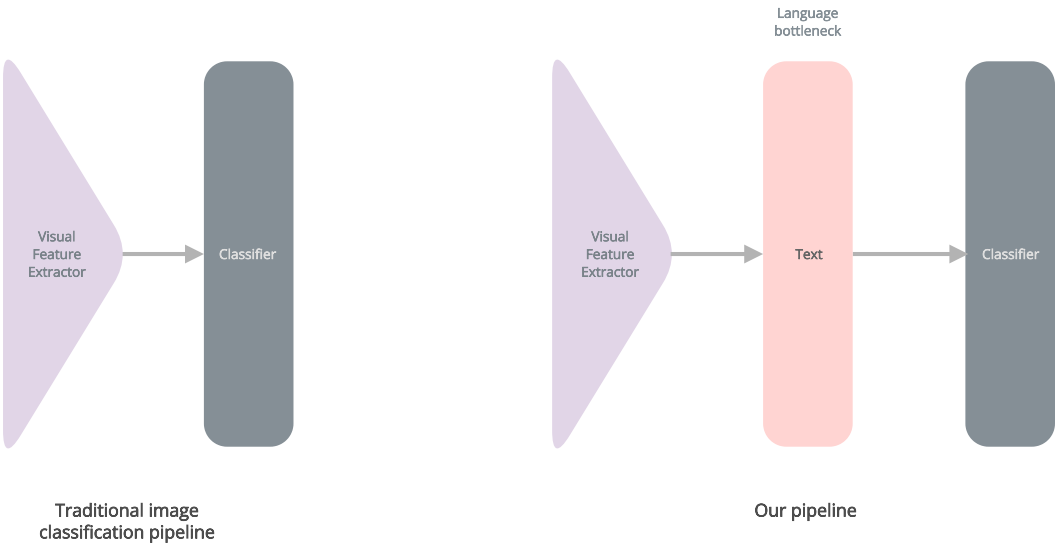

Figure 1: Architecture

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