Object Recognition & CV: Assignment 3 report

Carlos Cuevas Villarmin
ENS Paris-Saclay
4 avenue des Sciences 91 190 Gif-sur-Yvette
carlos.cuevas_villarminr@ens-paris-saclay.fr

Abstract

The purpose of this assignment is to participate in a Kaggle competition on the ImageNet-Sketch dataset. The objective of the competition is to produce a model that gives the highest possible accuracy on a test dataset containing the same categories. To improve the results of the DINOv2 model [5], the chosen approach combines visual embeddings with text embeddings derived from image captions.

1. Introduction

Image classification is a widely studied task, with new state-of-the-art models published annually. Currently, DINOv2 [5] is among the top models for this task. While comparing DINOv2 with other models like ViT [2] or SAM [3] could optimize performance on the ImageNet-Sketch dataset, this approach has been excluded for academic purposes in order not to follow a pipeline that does not require making decissions. Instead, it is hypothesized that *mixing DINOv2's embeddings with generated captions embeddings will improve the performance*, considering DINOv2 as the baseline. Code[1] and dataset[2] are publicly available.

2. Material and Methods

The ImageNet-Sketch dataset consists of a predefined data split with 500 classes to identify. To add text information, the following models have been chosen: BLIP-2 [4] for caption generation and BERT [1] to get the embeddings. The pipeline can be seen in Figure 1. The models have not been fine-tuned, which is left as further work.

Different strategies have been tested to combine both embeddings. In this report, we will focus on two of them. Firstly, concatenating the embeddings and passing them through an MLP. Secondly, a combination of two classifiers: one for each embedding. After this, a linear layer is added to combine both logits, and the final output, Eq. 1, is a weighted combination of the logits, where f and g are the classifiers for visual and text embeddings, respectively.

$$z = \alpha f(v_{emb}) + \beta g(t_{emb}) \tag{1}$$

See Figure 2 for a general overview. Regularization has been added to α and β to make both sum to 1 and remain positive. Making this weights learnable will allow confirmation or rejection of the hypothesis.

3. Results

In this section, the baseline (DINOv2), the naive combination (concatenation) of embeddings, and two scenarios of the proposed model with two weights for the regularization term are summarized in Table 1. The evolution of α and β can be seen in Appendix B.

Experiment	Train Acc.	Val. Acc.	Test Acc.
DINOv2-base/-giant	**98.19/97.79**	**89.44/90.32**	**89.78/91.2**
Concat.	95.98	86.48	86.72
Proposal-v1 ($\lambda = 0.1$)	96.59	87.72	87.95
Proposal-v2 ($\lambda = 0.3$)	96.13	86.84	87.34

Table 1. Accuracy obtained for all the models considered in train, validation and test splits. Test results are the public scores obtained in Kaggle.

4. Discussion and conclusions

This small project has served as an introduction to state-of-the-art models for image classification. Furthermore, based on the results, adding text embeddings does not always improve the model's performance, and the best performance is still achieved with DINOv2. As the reported accuracy is not far from that of DINOv2, perhaps with a more detailed analysis, the proposed alternative could surpass the defined baseline. However, the results indicate that, at least without fine-tuning BLIP-2 and BERT for text embedding generation, the information provided by those embeddings does not impact the prediction. This is left as further work, alongside hyperparameter optimization.

[1] https://github.com/cuevascarlos/recvis24_a3
[2] https://huggingface.co/datasets/cuevascarlos/ImageNet-Sketch-Embed

References

[1] 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. 1

[2] Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 1

[3] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026, 2023. 1

[4] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023. 1

[5] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 1

A. Material and Methods

Illustrative diagrams are presented to show how the dataset with text embeddings is created (Figure 1) and the proposed model in this work (Figure 2).

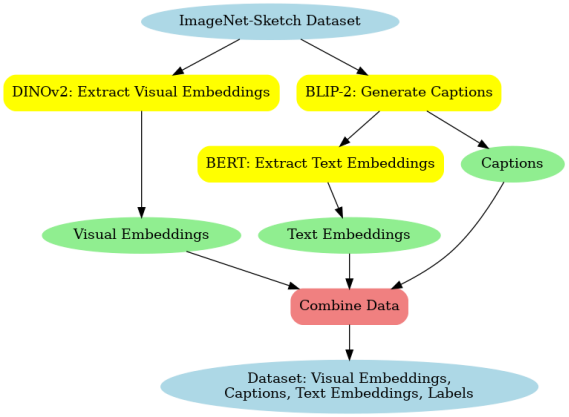

Figure 1. Pipeline for generating an additional dataset which also contains textual information of the images.

B. Results

In this appendix the evolution of the parameters α, β through the training process are analyzed. The progression over the epochs are seen in Figure 3 for α and in Figure 4 for β where the tendecy is clear: while α tends to 1 β oscillates around 0. The experimental results allow to conclude that text embeddings do not have an impact on the prediction.

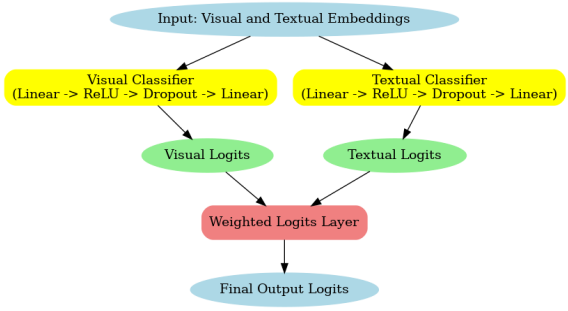

Figure 2. Illustrative diagrame of the proposed model considering two classifiers, one for each type of embedding.

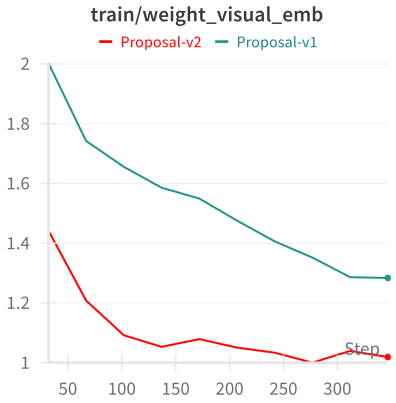

Figure 3. Evolution of the learnable parameter α in both proposed models: Proposal-v1 $\lambda = 0.1$ and Proposal-v2 $\lambda = 0.3$

Figure 4. Evolution of the learnable parameter β in both proposed models: Proposal-v1 $\lambda = 0.1$ and Proposal-v2 $\lambda = 0.3$