Project 2

Multimodal Learning with Limited Labels and Missing Modalities

[ENN585] Advanced Machine Learning

Cameron Stroud (n11552123)

1 Introduction

This project aims to investigate the effects on multimodal learning that missing modalities and limited labeling has on performance.

1.1 Dataset

The COCO single class dataset [1] is used. This consists of a series of images and captions that result in one of 91 possible classes, corresponding to the largest integer label. Whilst there are 91 possible classes, only 64 of them appear in the validation set, and only 80 within the training set.

1.2 Training

Training is performed on either 10% or 100% of the provided training set depending on the task. There are 24186 samples generated in total, so the 10% case has been capped at 2400. Appendix **??** shows various config parameters used throughout the model creation and training process. Of note, 10 epochs are used for training purposes.

2 Unimodal Classification Performance

2.1 Image-only Classification

The provided pre-trained CLIP model uses a TinyViT model as the image encoder: `tiny_vit_21m_224.dist_in22k_ft_in1k`. A new class `UnimodalImageModel` was created that uses this image encoder, with frozen weights, and appends a trainable linear layer for making the prediction.

2.2 Text-only Classification

The text encoder applied by the pre-trained CLIP is Distillbert: `distilbert-base-uncased`. Like with the unimodal image classification, a new class `UnimodalTextModel` was created that uses a frozen Distillbert encoder, with .

2.3 Results

We use the Top-1 and Top-5 accuracy for evaluating the performance in each missing modality case. The results of which are shown in Figure 1 below.

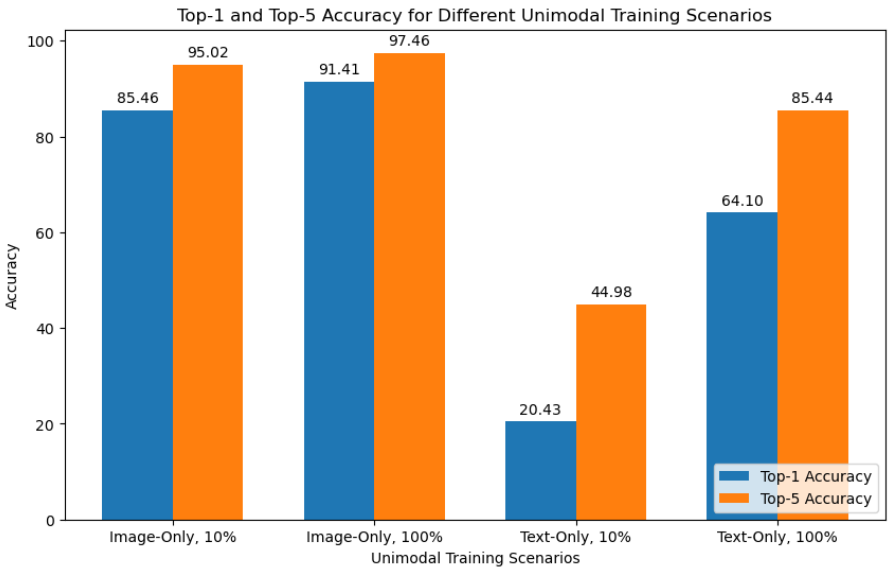

Figure 1: Top-1 and Top-5 Accuracy for Unimodal Scenarios

3 Multimodal Classification Performance

3.1 Multimodal CLIP

The multimodal CLIP applies both the pre-trained Distillbert and TinyViT encoders, and applies a mid-fusion method in order to process both image and text information. The image and text embeddings are concatenated together and passed as input into the trainable linear classification layer. The results of this are shown in Figure 2 below:

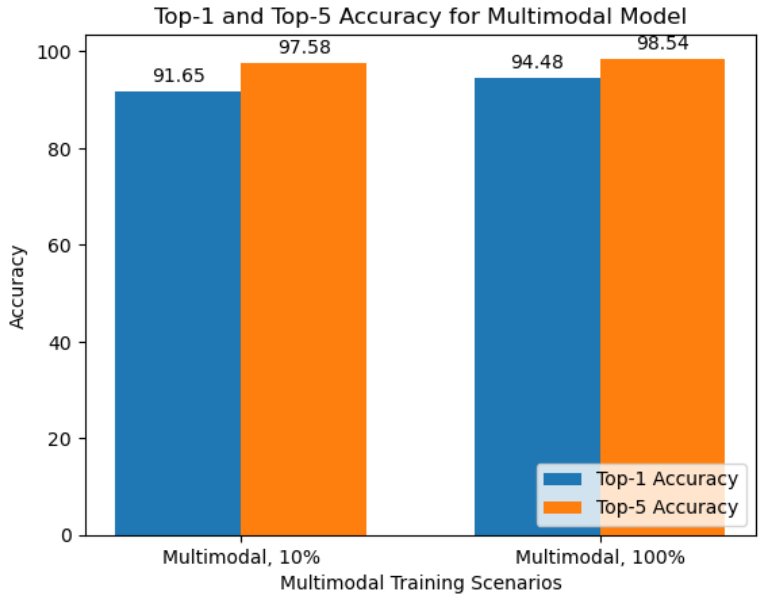

Figure 2: Top-1 and Top-5 Accuracy for Multimodal Scenarios

3.2 Modality Comparison

Comparing with the previous models, the multimodal model is a clear winner when looking at the Top-1 and Top-5 performance. The multimodal with 10% of the training labels has comparable to the fully-trained image encoder, with some marginal improvements in the 100% case. The image encoder still shows excellent results as a standalone model, while the text encoder shows significant error. This is likely a result of the lack of information provided in some captions. For example, Figure 3 shows the ambiguity of some captions with respect to the class. The strong image performance is likely a result of similar data in the ImageNet training set which the model was trained on.

Figure 3: Ambiguous Caption for a Fire Hydrant

3.3 Limited Label Improvement

As the text classifier has the biggest disparity in performance between 10% and 100% labels, it's been chosen to undergo some pre-training self-supervised learning. The encoder weights were unfrozen and the CLIP process followed in order to better align the text and image encoders, with the intent of better adapting the text encoder classifier for this dataset. CLIP is performed on the full dataset, not applying the use of any labels, and the resulting weights of the CLIP text encoder is applied to the unimodal text classifier.

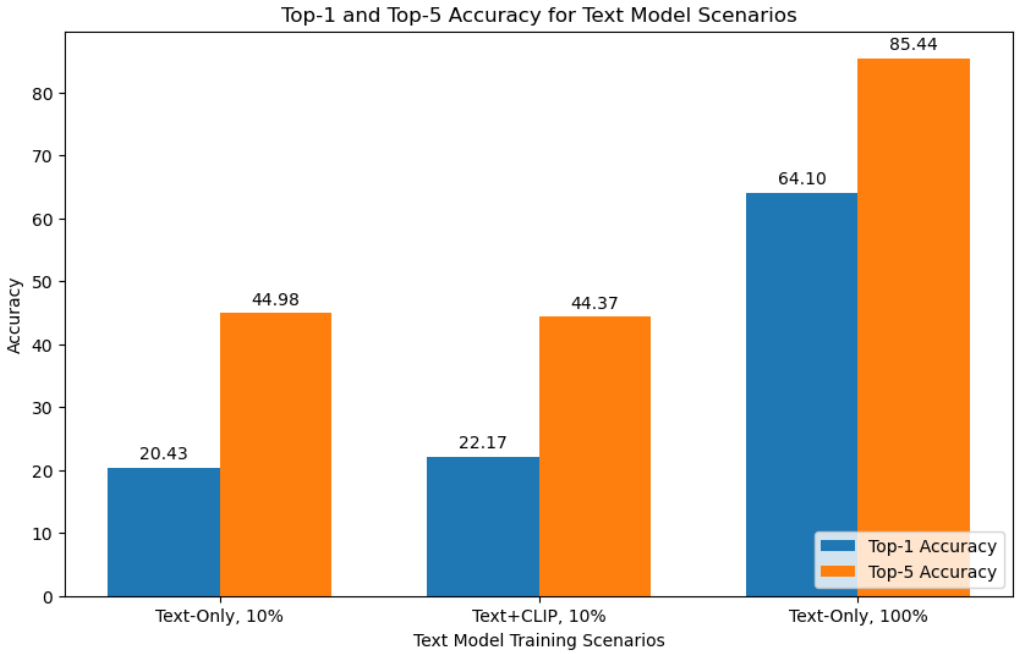

Figure 4: Top-1 and Top-5 Accuracy for Text Model Scenarios

Applying CLIP for the unlabelled training set did not alter the performance much at all. While there is some improvement in the Top-1 accuracy, the Top-5 accuracy actually shows degraded performance. While it may be possible that more epochs for the CLIP self-supervised pretraining would improve performance, at the current accuracy there should be a greater improvement over the base 10% trained model, and ultimately this process does not mitigate the impact of limited labels, given the significant difference compared to the 100% case.

The CLIP process was intended to align the text embedding space with the image embedding space, which has a much higher performance in the limited label case. The results show, however, that even with the updated encoder weights, the text classifier falls short, making CLIP inappropriate for improving the limited label performance in this context.

4 Modality Ablation

An ablasive analysis of the missing modalities was performed across different completeness of a particular modality. The previously applied multimodal model was applied (without training), and trained for a single epoch in each scenario, missing between 0% and 100% of data for each text and image. Text samples are replaced with an empty string, and image samples are replaced with a zeroed RGB image of the same size as the other images in the dataset. Figure 5 shows the result of this study.

We expect the performance to be trending downwards as data is removed and the 0% cases should be inline with the performance of the respective unimodal models, with some minor differences as a result of using 0-value modalities instead of solely the absence.

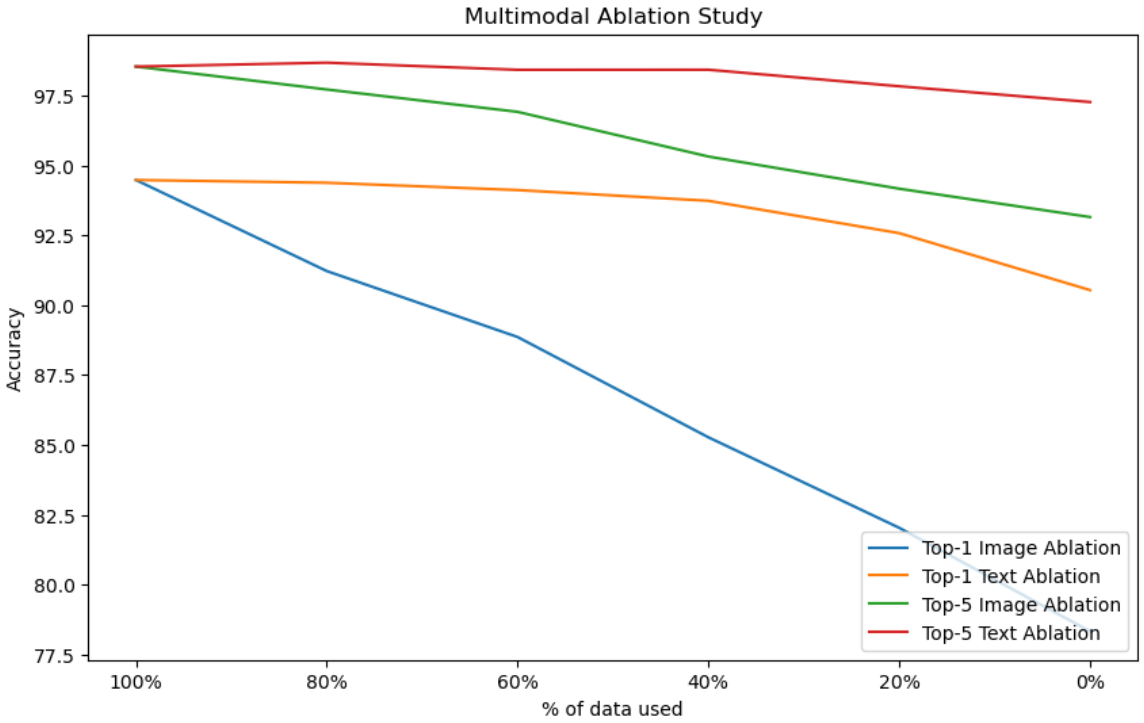

Figure 5: Ablation Study of Multimodal Performance

The results shown in Figure 5 align with our expectations, for the most part, but we see that the model actually performs better than the unimodal text, even when the image data has been zeroed.

5 Robust Multimodal Learning

5.1 Design

To improve the robustness of multimodal learning, we will enable the model to better deal with missing modalities. We will follow a missing-aware prompt approach, as described by Lee et. al. [2], specifically, input-level prompting, as it shows the strongest performance in the absence of modalities.

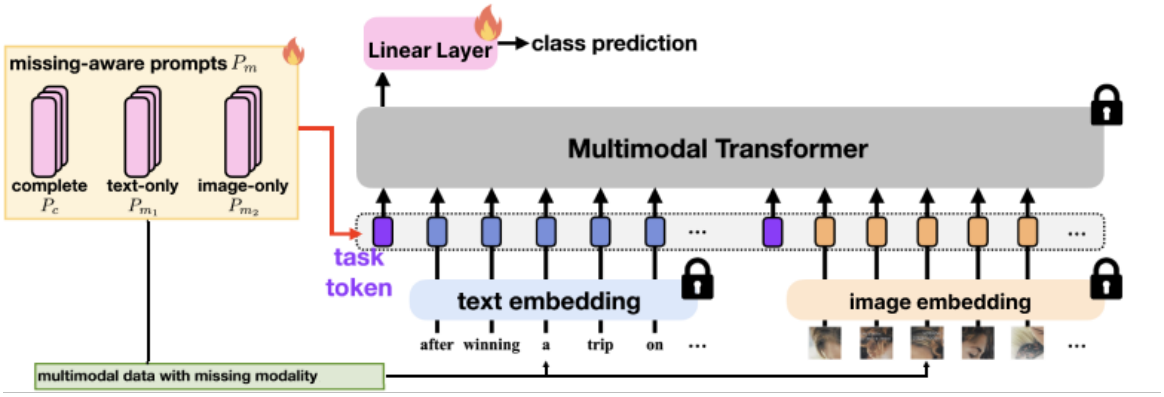

Figure 6: Design for Robust Multimodal Model, based off the work in [2]

In Figure 6 we see multimodal data passed through the respective encoders to gain the embeddings required as input, but also passing it through a missing-aware prompt generator. This creates a missing-aware prompt for each case of complete, text-only and image-only, and the aforementioned input-level method is applied wherein the output prompt is prepended to the input embeddings. The concatenated embeddings are then passed through the multimodal model we've been training thus far, with the trainable linear layer to provide the classification.

5.2 Implementation and Results

5.3 Recommendations for Future Work

References

[1] T.-Y. Lin, M. Maire, S. Belongie, L. Bourdev, R. Girshick, J. Hays, P. Perona, D. Ramanan, C. L. Zitnick, and P. Dollár, "Microsoft COCO: Common Objects in Context," Feb. 2015, arXiv:1405.0312 [cs]. [Online]. Available: http://arxiv.org/abs/1405.0312

[2] Y.-L. Lee, Y.-H. Tsai, W.-C. Chiu, and C.-Y. Lee, "Multimodal Prompting with Missing Modalities for Visual Recognition," in *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Vancouver, BC, Canada: IEEE, Jun. 2023, pp. 14 943–14 952. [Online]. Available: https://ieeexplore.ieee.org/document/10203663/

ChatGPT was used to generate utility functions used in the evaluation and plotting of material.