Few-Shot Adaptation of Contrastive Captioners (CoCa) using Parameter-Efficient Fine-Tuning

Progress Report

Narasinghe N.K.B.M.P.K.B 210407R

August 24, 2025

1. Literature review

There are new advances in deep learning using foundation models pre-trained on web-scale data to perform most tasks. In the vision-and-language domain, there are three major pretraining paradigms. There is, first, single-encoder classification, usually where visual encoders are pretrained through image classification on several large annotated datasets using a cross-entropy loss. The models of this approach learn generic visual representations that transfer well to downstream vision tasks but lack natural language understanding. Second, the dual-encoder contrastive approach, made popular by the models like CLIP [1] and ALIGN [2], employs separate encoders for images and text, training them jointly with contrastive loss to project images and the associated textual descriptions into a shared embedding space. This architecture showcases powerful zero-shot learning capabilities for tasks such as image classification and cross-modal retrieval through learning aligned unimodal representations. The third paradigm is that of the generative encoder-decoder approach: models like SimVLM [3] manifest in this paradigm. These models utilize a vision encoder, coupled with a text decoder, to be trained via captioning or language modeling objectives where the model learns generating textual descriptions given images. Such architecture naturally excels at multimodal understanding and generation tasks requiring fused image-text representations for applications such as visual question answering and image captioning.

The Contrastive Captioners (CoCa) model was introduced [4] to bring together these diverging paradigms. CoCa is an image-text foundation model that jointly pretrained from end to end using contrastive loss and captioning loss [5, 6]. It is believed that this dual objective approach enables the model to learn a high-quality aligned representation for the task of retrieval and classification, while at the same time also learning powerful

multimodal representations for reasoning and generation all before a single pretrained checkpoint.

1.1. CoCa Model Architecture

The CoCa architecture is a novel encoder-decoder design that efficiently computes both training objectives using little extra overhead [4]. Its main components are:

1. **Image Encoder:** A standard Vision Transformer (ViT) processes an image into a sequence of patch embeddings.
2. **Decoupled Text Decoder:** The main innovation of CoCa, consists of separating the text decoder into two sequential parts discussed below.
   - **2.**1. *Unimodal Decoder Layers:* The first half of the decoder layers process input text autoregressively but exclude cross-attention to the image encoder's output. These layers are essentially responsible for learning pure unimodal text features. The final hidden state corresponding to a special `[CLS]` token from this section is used as the text embedding for the contrastive loss.
   - **2.**2. *Multimodal Decoder Layers:* The second half of the decoder layers function as a standard multimodal decoder. They cascade from the unimodal layers and perform cross-attention to the image encoder's output. These layers learn to fuse visual and textual information to generate multimodal representations, which are then used to predict the next text token for the captioning loss.
3. **Attentional Poolers:** CoCa uses task specific attentional pooling to create customized visual representations from the image encoder's output tokens. These poolers are single multi-head attention layers with a set of learnable queries. Two different types of poolers are used during pre training:
   - **Contrastive Pooler:** Single query that produces one global image embedding against the contrastive loss.
   - **Generative Pooler:** Uses multi queries for generating a sequence of fine-grained visual tokens to be used against the captioning loss.

Although CoCa has shown state-of-the-art results for generalization settings both in zero shot transfer and full fine-tuning [4], it presents a unique challenge of adapting such a large model for new applications any time data would be limited. Full fine-tuning is also very expensive in computation and results in "catastrophic forgetting" of the powerful pre-trained knowledge. Thus, more efficient methods of adaptation are needed. Few-Shot Learning (FSL) [7] addresses precisely this problem by enabling models to learn new concepts from very few examples. Parameter-Efficient Fine-Tuning (PEFT) methods, including Low-Rank Adaptation (LoRA) [8], recently emerge as a promising solution. LoRA freezes the entire pretrained model weights and injects small, trainable rank-decomposition matrices into the Transformer layers, reducing trainable parameters considerably, and leading to much easier fine-tuning.

This research aims to fill the gap by applying PEFT techniques to the CoCa model and designing a computationally efficient yet highly pragmatic framework for few-shot adaptation. This study is novel in being the first to comprehensively assess the best methodology for adapting CoCa's unique dual-objective architecture to few-shot image classification.

2. Methodology Outline

This work studies few-shot learning characteristics of CoCa and develops and assesses parameter-efficient model adaptation strategies from pretrained versions of CoCa. The unified architecture of CoCa, joining contrastive and generative objectives offers opportunities and challenges in few-shot adaptation that have yet to be tackled in previous studies.

2.1. Experimental Datasets

The evaluation will be conducted on benchmark few-shot image classification datasets that are representative across multiple visual domains and complexity levels:

- **miniImageNet** : A subset of ImageNet containing 100 classes with 600 images each, split into 64/16/20 classes for training/validation/testing respectively.
- **TieredImageNet** : A larger-scale benchmark with 608 classes organized hierarchically, providing more diverse semantic relationships for robust evaluation.
- **CIFAR-FS** : Derived from CIFAR-100, offering 100 classes with lower resolution images to test adaptation under different visual characteristics.

These benchmarks enable systematic evaluation across varying numbers of support examples (1-shot, 5-shot, 10-shot) while setting standardized protocols for fair comparison with existing techniques.

2.2. Baseline Comparisons

To establish comprehensive performance analysis, the proposed approach will be evaluated against several key baselines that represent different adaptation paradigms:

1. **Zero-Shot CoCa** Direct application of pre-trained model CoCa using its intrinsic zero-shot classification capability to perform image-text similarity matching, establishing the lower bound without any task adaptation.

2. **Linear Probing**: The whole pre-trained model is kept frozen, the only thing trained is a linear head on the features it extracts. It tests how good the learned representations are without changing the architecture.

3. **Component-Specific Fine-Tuning**:Selective adaptation of specific CoCa components, such as the visual encoder, the text decoder, or even only the attentional poolers, in order to understand which architectural component contributes the most in an important way to performance in few-shot tasks.

2.3. Parameter Efficient Fine Tuning Strategies

The core methodology explores applying parameter-efficient fine-tuning techniques particularly meant for the dual-pathway architecture of CoCa. In light of constraints concerning computations and the time limitations, this study would systematically assess the following techniques:

- . **Low-Rank Adaptation (LoRA)**: Adding trainable low-rank matrices to inject into the attention layers of both visual and text components for effective adapting without loss of pre-trained knowledge.
- . **Adapter Modules**: Adding lightweight bottleneck configurations in transformer blocks for costless task-specific adaptation.
- . **Prompt Tuning**: Learning continuous prompt embeddings that direct the reaction of the model concerning the classification task by taking advantage of CoCa's understanding of text.
- . **Multi-Modal Adaptation**: New adaptation strategies that jointly incorporate contrastive and generative pathways in CoCa into a single optimizing framework, leveraging the model's unified training objectives.

LoRA will be prioritized as the primary technique due to its proven effectiveness and computational efficiency, while other approaches can be treated as secondary investigations, depending on available resources and preliminary results.

The methodology is designed to evaluate whether parameter efficient adaptation could improve the few-shot capacity of CoCa and, possibly, how effectively multimodal pretrained architectures could be transferred to new tasks with limited data.

3. Timeline

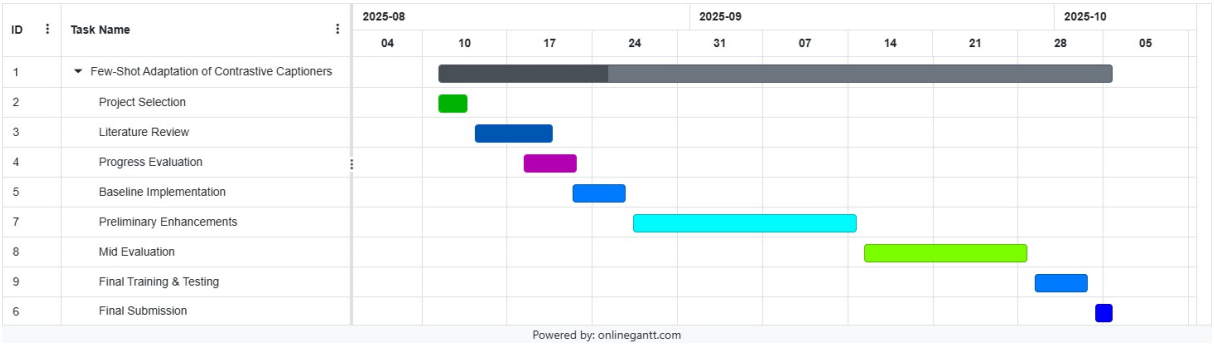

Figure 1: Proposed Timeline for the Project

References

[1] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," 2021.

[2] C. Jia, Y. Yang, Y. Xia, Y.-T. Chen, Z. Parekh, H. Pham, Q. V. Le, Y. Sung, Z. Li, and T. Duerig, "Scaling up visual and vision-language representation learning with noisy text supervision," 2021.

[3] Z. Wang, J. Yu, A. W. Yu, Z. Dai, Y. Tsvetkov, and Y. Cao, "Simvlm: Simple visual language model pretraining with weak supervision," 2022.

[4] J. Yu, Z. Wang, V. Vasudevan, L. Yeung, M. Seyedhosseini, and Y. Wu, "Coca: Contrastive captioners are image-text foundation models," 2022.

[5] P. Khosla, P. Teterwak, C. Wang, A. Sarna, Y. Tian, P. Isola, A. Maschinot, C. Liu, and D. Krishnan, "Supervised contrastive learning," 2021.

[6] R. J. Williams and D. Zipser, "A learning algorithm for continually running fully recurrent neural networks," *Neural Computation*, vol. 1, no. 2, pp. 270–280, 1989.

[7] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei, "Language models are few-shot learners," 2020.

[8] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "Lora: Low-rank adaptation of large language models," 2021.