Adapting Multimodal Foundation Models for Few-Shot Learning: A Comprehensive Study on Contrastive Captioners

Narasinghe N.K.B.M.P.K.B

210407R

October 5, 2025

Abstract

Large-scale multimodal models like Contrastive Captioners (CoCa), pre-trained on web-scale datasets, have demonstrated remarkable zero-shot capabilities. Nevertheless, they are not very adaptable to downstream tasks with sparse labeled data (few-shot learning). Full fine-tuning is computationally costly, and can be subject to overfitting. This paper presents a comprehensive empirical study on few-shot adaptation of the CoCa model for image classification. We systematically evaluate a hierarchy of methods, from a parameter-free hybrid prototype approach to parameter-efficient fine-tuning (PEFT) via linear probing. Findings indicate that the hybrid prototype approach is highly effective in utilizing the multimodality of CoCa, presenting high performance in very low shot scenarios and linear probing using hyperparameters and augmentation carefully-tuned also delivers additional accuracy improvements at a small number of trainable parameters. Preliminary experiments on LoRA fine-tuning are also conducted, although a comprehensive assessment is yet to be performed. Overall, the findings highlight that the optimal adaptation strategy is highly dependent on the number of available shots and computational constraints.

1 Introduction

The introduction of pre-trained foundation models on massive, diverse datasets has changed the face of deep learning. In vision-and-language, models like CLIP [1] and ALIGN [2] have established new benchmarks of zero-shot transfer learning by matching images and text in a common embedding space through contrastive learning. The Contrastive Captioners (CoCa) model [3] combines the contrastive and generative pre-training paradigms into one architecture and demonstrates state-of-the-art performance on a wide variety of multimodal tasks.

Despite their powerful zero-shot capabilities, effectively adapting these large models to specific downstream tasks with very limited labeled data setting known as few-shot learning (FSL) is a non-trivial problem. Full fine-tuning can frequently be computationally infeasible, data-inefficient, and can cause risks catastrophic forgetting of the useful pre-trained knowledge. Parameter-efficient fine-tuning (PEFT) methods, such as Low-Rank Adaptation (LoRA) [4], have become a viable option as they update a small fraction of parameters. Although much has been done to train large language models with these techniques, limited research has been done on multimodal complex models such as CoCa.

This paper systematically studies few-shot adaptations based on the CoCa model. From a method-centered study, we explore more generally a hierarchy of progressively complex and more parametric methods:

(i) **Parameter-free hybrid prototype** method that fuses visual and textual embeddings without any training.

(ii) **Linear probing** with a frozen visual encoder, enhanced by strong data augmentation.

(iii) **LoRA fine-tuning** of the visual encoder with an adaptive configuration and a hybrid loss function.

This work studies the performance of multimodal foundation models under data scarcity in the Mini-ImageNet benchmark. Our findings indicate that CoCa's pre-trained representations form a solid basis for FS learning, where a very simple prototype-based technique achieves remarkable performance. The other key finding was that different adaptation methods have suitability that changes depending on the data regime. In the particular case of the LoRA method, we observed a data-dependent trade-off in the choice of loss function: losses that promote learning to rank (prototypical, contrastive) were more performance on high effectiveness with very few shots, while cross-entropy became more useful as more samples arise.

2 Related Work

2.1 Multimodal Foundation Models

Vision-and-language models have evolved through several pre-training paradigms. **Single-encoder models** are typically trained for image classification, learning strong visual representations but lacking linguistic understanding. **Dual-encoder models** like CLIP [1] and ALIGN [2] use contrastive learning to bring images and text close together in a shared space, which allows for very strong zero-shot retrieval and classification capability. **Encoder-decoder models** are trained with generative objectives (e.g., captioning) and perform well on tasks requiring fused multimodal reasoning.

Robustly merging these types of approaches is CoCa [3] wherein an image encoder, unimodal text decoder operating under a contrastive learning framework producing a text embedding, and a multimodal decoder for captioning can be found. This unique, dual-objective pre-training equips CoCa with strong aligned representation and generative ability, rendering it a sound base model for multimodal tasks.

2.2 Few-Shot Learning and Parameter-Efficient Fine-Tuning

Few-Shot Learning aims to generalize from a very limited number of examples. Classic approaches to Few-shot learning include Meta-learning [5] and metric-learning [6]. With the rise of large pre-trained models, the paradigm has shifted towards adapting these models using their rich, general-purpose features.

PEFT methods are essential for this adaptation. Techniques like Adapter Modules [7], Prompt Tuning [8], and LoRA [4] allow custom modification of the model while revising only a small percentage of the parameters. LoRA, in particular, has gained traction for its performance and efficiency; it injects trainable low-rank matrices into transformer layers to approximate the weight updates. While PEFT has established grounds within NLP, its consideration for visio-linguistic models such as CoCa for few-shot learning is an exciting yet underinvestigated territory, which this work aims to address.

3 Methodology

3.1 Data Preparation

We use the Mini-ImageNet dataset, a standard FSL benchmark, having 100 classes with 600 images each. Using official splits, we construct a few-shot balanced dataset. The first 20 from original train split is used for training, the first 20 from validation split for validation and the next 20 images are used for testing in each class. Thus, there will be 2,000 images per split (6,000 in total), formatted in an Image Folder structure. The pipeline was implemented in the Hugging Face `datasets` library.

3.2 Model and Adaptation Strategies

We use a pre-trained CoCa model (ViT-L/14) with weights from `mscoco_finetuned_laion2B-s13B-b90k`. The model is kept frozen in evaluation mode for all experiments unless otherwise specified. Three adaptation strategies are explored.

3.2.1 Strategy 1: Hybrid Prototype Classification

This is a training-free method that leverages CoCa's multimodal nature.

- **Visual Prototype:** For a class c with N support images $\{I_1, ..., I_N\}$, the visual prototype is the normalized mean of their image embeddings.

$$P_{image}(c) = \text{normalize}(\frac{1}{N}\sum_{i=1}^{N} f_{img}(I_i)).$$

- **Textual Prototype:** Using prompt ensembling (e.g., "an image of a class"), the textual prototype is the normalized mean of the text embeddings.

- **Hybrid Fusion:** The final prototype is a weighted combination, where α is a fusion hyperparameter.

$$P_{hybrid}(c) = \text{normalize}((1-\alpha)P_{image}(c) + \alpha P_{text}(c))$$

A query image is classified by finding the class c with the highest cosine similarity between its embedding E_{query} and $P_{hybrid}(c)$.

3.2.2 Strategy 2: Linear Probing

We attach a new linear classification head to the frozen CoCa image encoder. Only the parameters of this head are trained. Since the training set is small, it has been augmented with strong augmentation via the following strategies: Random Resized Crop, Horizontal Flip, Color Jitter, Random Grayscale, etc. The model is trained under AdamW, cross-entropy loss with label smoothing, and a cosine learning rate scheduler with warmup.

3.2.3 Strategy 3: LoRA Fine-Tuning

We employ LoRA to adapt the internal weights of the frozen ViT image encoder. For a pre-trained weight matrix $W_0 \in \mathbb{R}^{d\times k}$, LoRA constrains the update with a low-rank decomposition: $h = W_0x + \Delta Wx = W_0x + BAx$, where $A \in \mathbb{R}^{r\times d}, B \in \mathbb{R}^{k\times r}$, and the rank $r \ll min(d, k)$. Only A and B are trained.

We use an **adaptive LoRA configuration** based on the number of shots N:

- $N \in \{1, 2\}$: Rank $r = 4$, applied only to `attn.out_proj`.
- $N \in \{3, ..., 10\}$: Rank $r = 8$, applied to `attn.out_proj`.
- $N > 10$: Rank $r = 16$, applied to `attn.out_proj`, `mlp.c_fc`, `mlp.c_proj`.

We investigate three loss functions in a **hybrid training** setup, where the total loss is $L_{total} = L_{metric} + L_{CE}$:

- **Cross-Entropy (CE) Loss**: Standard classification loss.
- **Prototypical Loss** [9]: A metric learning loss that minimizes the Euclidean distance between queries and class prototypes.
- **Supervised Contrastive (SupCon) Loss** [10]: A metric learning loss that structures the embedding space by pulling together samples from the same class.

The LoRA parameters and linear head are trained simultaneously with AdamW, using a higher learning rate for the head.

4 Experiments and Results

4.1 Experimental Setup

We do all the experiments on the MiniImageNet few-shot splits; We evaluate across different numbers of shots $N \in \{1, 3, 5, 10, 20\}$. For the hybrid prototype, we search for the optimal text weight α. For linear probing and LoRA, we report the results with augmentation and without augmentation (when applicable). The hyperparameters associated with each method are also reported. The metric is Overall Accuracy.

4.2 Results and Analysis

4.2.1 Hybrid Prototype Classification

Table 1: Accuracy (%) of Hybrid Prototype Method across different text weights (α)

N-Shot	$\alpha = 0$	$\alpha = 0.2$	$\alpha = 0.5$	$\alpha = 0.7$
1	71.75	75.00	83.60	85.60
3	86.10	86.60	89.60	89.00
5	88.45	89.60	91.20	90.80
10	90.3	91.55	91.95	90.90
20	91.6	92.50	92.85	91.35

As shown in Table 1, hybrid prototype model methodology becomes consistently better in demonstrating improvement. It is evident that extra visual examples enhance adaptation. Across all settings, incorporating text prompts ($\alpha > 0$) proves to be a significant advantage when compared with visual prototypes alone. This suggests that CoCa's textual embeddings, which encode rich semantic knowledge from pre-training, complement the limited visual evidence available in few-shot settings. The gains are most pronounced

at very low shots (e.g., 1-shot), where textual priors substantially boost accuracy, while with more shots the visual prototypes gradually close the gap, leading to the highest overall performance when both cues are combined.

4.2.2 Linear Probing

Table 2: Linear Probing Accuracy under Different Augmentation Levels

N-Shot	Augmentation: False	Augmentation: Low	Augmentation: High
1	66.00	65.20	66.80
3	84.65	84.89	83.00
5	84.80	88.25	84.80
10	90.20	90.05	89.80
20	91.80	91.25	90.90

As shown in Table 2, linear probing fine-tuning shows the expected improvement with more shots, as richer supervision strengthens the classifier.Data augmentation is key in the low-shot scenarios: mild augmentations such as random resized cropping, horizontal flipping, and slight color jitter have improved performance considerably (e.g., at 5-shot) by adding feature diversity. In contrast, more aggressive augmentations sway the results with heavy color jitter and random grayscale. As the number of available examples increases, the contribution of augmentation starts to fade; all variants would finally stabilize at the same level of performance. The best results were achieved with a cosine-decayed learning rate of 1×10^{-4}, AdamW optimizer, cross-entropy loss with label smoothing (0.1), weight decay of 0.01,a warm-up of 10 epochs, and a batch size of 32, providing stable convergence.

4.2.3 LoRA Fine-Tuning

The exploration of **LoRA fine-tuning** as an alternative to linear probing is currently underway. Linear probing only trains the classification head, while LoRA allows parts of the feature extractor to be adapted using low-rank weight updates, thus allowing deeper plug-in task-specific adaptations while maintaining a low number of parameters that participate in training. During this ongoing investigation, we have also tested several loss functions-cross-entropy, supervised contrastive, and prototypical loss-to see how their performance can vary with regard to the number of available examples. Some early observations suggest that loss choice may become considerably data-dependent: metric-based objectives, such as prototypical loss and contrastive loss, appear to have an upper hand in the low-shot scenario, while their counterpart cross-entropy may emerge as a stronger contender when we have more data. These observations still need to be verified with further experimentation, which will allow us to confirm the trends and arrive at a better understanding of the trade-offs involved in each approach.

5 Conclusion and Future Work

This work provides a detailed empirical investigation on few-shot adaptation of the CoCa model, addressing parameter-efficient approaches. The results show a clear hierarchy of methods based on their trade-off capability. The parameter-free hybrid protoyping approach stands out as a strong and simple baseline that capitalizes on the semantic priors from text to boost performance immensely in extremely low-shot settings. Linear

probing adds to the performance, while optimized regularization and data augmentation strategies give large returns on the cost of a only small number of extra parameters.

The exploration of LoRA fine-tuning is still ongoing. Preliminary experiments indicate that adapting parts of the feature extractor through low-rank updates may yield further improvements, but a deeper evaluation particularly of loss functions and training dynamics is required before definitive conclusions can be drawn.

Hence far, the findings point out that there are no one-fit-all solutions: the selection of the best way to adapt depends on different conditions such as the amount of data available, computational limitations, and performance expectations. Future work will focus on completing the LoRA evaluation and extending the study to other multimodal architectures and benchmarks, as well as exploring joint adaptation of the visual and textual encoders using parameter-efficient fine-tuning techniques for more holistic downstream transfer.

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

[4] 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.

[5] C. Finn, P. Abbeel, and S. Levine, "Model-agnostic meta-learning for fast adaptation of deep networks," 2017.

[6] K. Musgrave, S. Belongie, and S.-N. Lim, "A metric learning reality check," 2020.

[7] N. Houlsby, A. Giurgiu, S. Jastrzebski, B. Morrone, Q. de Laroussilhe, A. Gesmundo, M. Attariyan, and S. Gelly, "Parameter-efficient transfer learning for nlp," 2019.

[8] B. Lester, R. Al-Rfou, and N. Constant, "The power of scale for parameter-efficient prompt tuning," 2021.

[9] J. Snell, K. Swersky, and R. S. Zemel, "Prototypical networks for few-shot learning," 2017.

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