Meta-Learning Makes a Better Multimodal Few-shot Learner

Ivona Najdenkoska, Xiantong Zhen, Marcel Worring
AIMLab, University of Amsterdam
Amsterdam, the Netherlands
i.najdenkoska@uva.nl, x.zhen@uva.nl, m.worring@uva.nl

Abstract

Multimodal few-shot learning is challenging due to the large domain gap between vision and language modalities. As an effort to bridge this gap, we introduce a meta-learning approach for multimodal few-shot learning, to leverage its strong ability of accruing knowledge across tasks. The full model is based on frozen foundation vision and language models to benefit from their already learned capacity. To translate the visual features into the latent space of the language model, we introduce a light-weight meta-mapper acting as a meta-learner. By updating only the parameters of the meta-mapper, our model learns to quickly adapt to unseen samples with only a few gradient-step updates. Unlike prior multimodal few-shot learners, which need a hand-engineered task induction, our model is able to induce the task in a completely data-driven manner. Experiments on recent multimodal few-shot benchmarks demonstrate that compared to its counterparts our meta-learning approach yields better multimodal few-shot learners, while being computationally more efficient.

1 Introduction

Learning quickly from a few observations in a multimodal environment is an integral part of human intelligence [1, 2]. Yet, it is quite challenging for current foundation vision and language models to perform multimodal few-shot learning [3, 4] due the limited number of labeled samples. The challenges arise from the fact that current vision-only and language-only models are trained separately on different datasets and optimize different objectives, which results in inconsistent latent representations. The Frozen model [3] is the first multimodal few-shot learner, trying to bridge the gap between these models, by taking inspiration from how language models [5] perform in-context learning. This requires prompting of the language model with a hand-engineered task description, followed by a few demonstrations of the task. While being a good approach for simpler tasks, like binary decisions, it is not optimal to hand-engineer the prompt each time [6], especially when it comes to more complex multimodal tasks involving reasoning [3, 4].

Meta-learning or *learning to learn* [1, 7, 8] comes as a natural solution to any few-shot settings. Notably, it can be deployed to accrue shared knowledge from related tasks and rapidly learn new tasks by observing only limited labeled data. While it has been extensively studied in unimodal settings, particularly for few-shot image classification [9, 10, 11, 12, 13], meta-learning remains almost unexplored for multimodal few-shot settings. We hypothesize that empowering a multimodal few-shot learner with the ability to do meta-learning would assist in building internal representations broadly suitable for many tasks, while inducing the task in a data-driven manner could fill in the gap between the different foundation models.

Motivated by this, we define a novel multimodal few-shot meta-learning approach, illustrated in Figure 1. Instead of training foundation models from scratch, our architecture, shown in Figure 2,

6th Workshop on Meta-Learning at NeurIPS 2022, New Orleans.

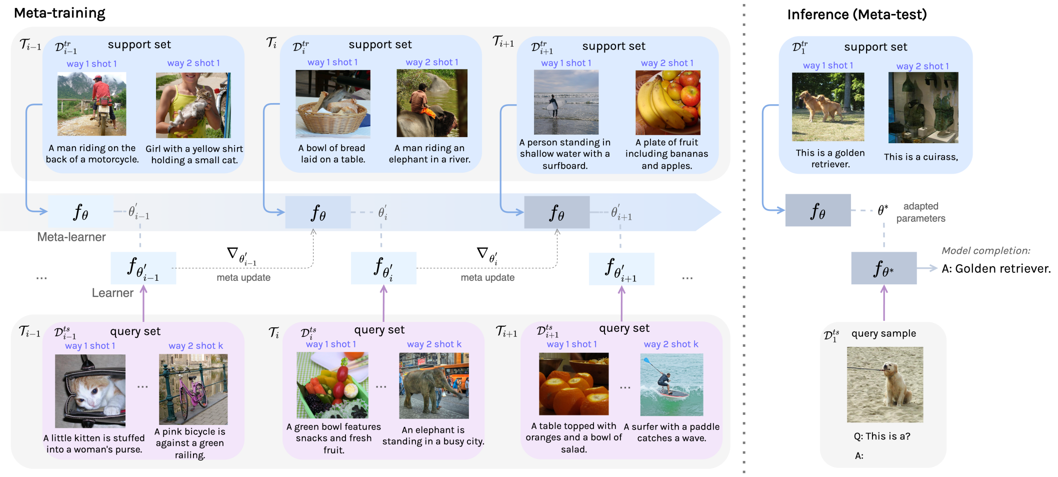
Meta-training
$\mathcal{T}_{i-1}$ $\mathcal{D}^{tr}_{i-1}$ support set
way 1 shot 1
A man riding on the back of a motorcycle.
way 2 shot 1
Girl with a yellow shirt holding a small cat.
$\mathcal{T}_{i}$ $\mathcal{D}^{tr}_{i}$ support set
way 1 shot 1
A bowl of bread laid on a table.
way 2 shot 1
A man riding an elephant in a river.
$\mathcal{T}_{i+1}$ $\mathcal{D}^{tr}_{i+1}$ support set
way 1 shot 1
A person standing in shallow water with a surfboard.
way 2 shot 1
A plate of fruit including bananas and apples.
f_θ
θ'_{i-1}
Meta-learner
f_θ
θ'_{i}
f_θ
θ'_{i+1}
...
$f_{\theta'_{i-1}}$
Learner
$\nabla_{\theta'_{i-1}}$
meta update
$f_{\theta'_{i}}$
$\nabla_{\theta'_{i}}$
meta update
$f_{\theta'_{i+1}}$
...
$\mathcal{T}_{i-1}$ $\mathcal{D}^{ts}_{i-1}$ query set
way 1 shot 1
A little kitten is stuffed into a woman's purse.
...
way 2 shot k
A pink bicycle is against a green railing.
$\mathcal{T}_{i}$ $\mathcal{D}^{ts}_{i}$ query set
way 1 shot 1
A green bowl features snacks and fresh fruit.
...
way 2 shot k
An elephant is standing in a busy city.
$\mathcal{T}_{i+1}$ $\mathcal{D}^{ts}_{i+1}$ query set
way 1 shot 1
A table topped with oranges and a bowl of salad.
...
way 2 shot k
A surfer with a paddle catches a wave.
Inference (Meta-test)
$\mathcal{D}^{tr}_{1}$ support set
way 1 shot 1
This is a golden retriever.
way 2 shot 1
This is a cuirass,
f_θ
θ^* adapted parameters
f_{θ^*}
Model completion:
A: Golden retriever.
$\mathcal{D}^{ts}_{1}$ query sample
Q: This is a?
A:

Figure 1: Multimodal few-shot meta-learning task for an example of a 2-way 1-shot setting, with two categories (ways) present in the support set, each represented with one sample (shot).

adopts a pre-trained vision encoder [14] and language model [5], which are kept entirely frozen during training, in order to not distort their learned parameters [15, 16, 17, 18]. Another major reason for smartly re-using trained models instead of training them, is the huge computational burden they create during training, and their dependency on large-scale datasets. The multimodal bridge between the frozen vision and language backbones is defined as a light-weight meta-mapper and built entirely from self-attention layers. By updating only the meta-mapper during meta-training, the model learns to map the visual features into a visual prefix corresponding to the latent space of the language model. This diminishes the need for fixed task inductions, since the model is able to accrue shared meta-knowledge from related tasks and induce the task for the query samples in a data-driven manner.

To summarize, our contributions are as follows: (*i*) We introduce meta-learning to perform multimodal few-shot learning, which enables fast adaptation and efficient learning of multimodal few-shot tasks. (*ii*) We present a multimodal few-shot meta-learner, which bridges a frozen vision encoder with a language model by using a light-weight meta-mapper, aiming to meta-learn a learnable visual prefix. (*iii*) We design a new multimodal meta-learning setting and experimentally demonstrate on existing multimodal few-shot benchmarks that our model yields a strong performance, while being computationally very efficient.

2 Methodology

Our objective is to train a model that can learn and quickly adapt to new multimodal tasks with limited labeled data. First, we define the multimodal meta-learning setting, then we explain our architecture in details, and finally how it used during training and at inference time.

2.1 Multimodal Meta-learning Setting

Different from standard supervised learning, in meta-learning settings we are dealing with a collection of meta-datasets split into disjoint partitions, namely, meta-training and meta-test. Both partitions consist of meta-datasets $\mathcal{D}_i$ containing a pair of a separate inner-train set $\mathcal{D}_i^{tr}$ i.e. a support set and an inner-test set $\mathcal{D}_i^{ts}$, i.e. a query set, meaning $\{(\mathcal{D}_1^{tr}, \mathcal{D}_1^{ts}), \dots (\mathcal{D}_n^{tr}, \mathcal{D}_n^{ts})\}$. Each pair (D_i^{tr}, D_i^{ts}) is referred to as a meta-task $\mathcal{T}_i$, following [11].

When considering a k-shot, N-way setting, a single support set $\mathcal{D}_i^{tr}$ consists of k labeled samples for each of the N-ways, where N is the number of object categories. This means that $\mathcal{D}_i^{tr} = \{(x_1^i, y_1^i) \dots (x_k^i, y_k^i)\}$, where x_j^i represents the image and y_j^i represents the corresponding caption. The query set $\mathcal{D}_i^{ts}$ is similarly defined, but has more samples than the support set, as they are needed for the inner-optimization step, as described in the next sections.

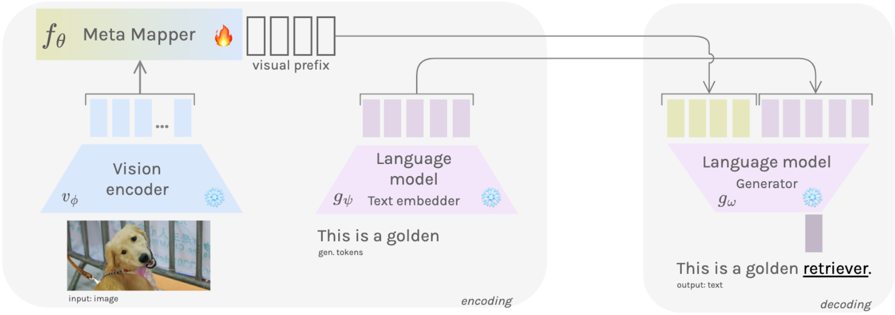

Figure 2: The architecture of the multimodal meta few-shot learner. It consists of three parts: frozen vision encoder v_ϕ; a meta-mapper f_θ with trainable meta-parameters θ; frozen language model with a text embedder g_ψ and a generator g_ω;

2.2 Model Architecture

Vision encoder The vision encoder is defined as a function v_ϕ, with fixed parameters $\phi \in \mathbb{R}^{d_v}$. The input is a raw image $\mathbf{x}$ and the outputs are the extracted visual features $x_1, \dots x_n = v_\phi(\mathbf{x})$.

Meta-mapper To map the visual features $x_1, \dots x_n$ into the language space, we use a set of l learnable parameters $p_i \in \mathbb{R}^{d_e}$, namely the visual prefix for the language model. We prepend these learnable parameters to the visual features, and we view it as an ordered set of elements: $[p_1 \dots p_l, x_1, \dots x_n]$. Then we employ multi-head self-attention to simultaneously encode the whole set [19], i.e., $\mathrm{MetaMap}_\theta(Q, K, V) = \sigma(QK^T) * V$. The pairwise dot-product QK^T measures the similarity amongst features, and is used as feature weighting computed through an activation function σ. The output of the meta-mapper are the learned parameters i.e. the visual prefix $p_1^* \dots p_k^*$, meaning that $p_1^* \dots p_l^* = \mathrm{MetaMap}_\theta([p_1 \dots p_l, x_1, \dots x_n])$.

Language model The language model uses an embedding function g_ψ to embed each generated token into a word embedding t_i, followed by a Transformer decoder defined as a function g_ω, to perform the text generation. During the meta-training, the language model receives the visual prefix $p_1^* \dots p_k^*$ concatenated with the token embeddings $t_1, \dots t_m$, and outputs the next token conditioned on the prefix: $t_{i+1} = g_\omega([p_1^* \dots p_l^*, t_1, \dots t_i]), i < m$, in an autoregressive manner.

2.3 Meta-Training & Inference

First, for simplicity, we assume that our full architecture described in 2.2 is defined as a function f_θ, which receives an image $\mathbf{x}$ as input and produces $\mathbf{y}$ as output. The loss function, optimized per task during training, is a cross-entropy loss $\mathcal{L}_{\mathcal{T}_i}(f_\theta)$, defined as:

$$\mathcal{L}_{\mathcal{T}_i}(f_\theta) = \sum_{\mathbf{x}^j, \mathbf{y}^j \sim \mathcal{D}_i^{tr}} \mathbf{y}^j \log f_\theta(\mathbf{x}^j) + (1 - \mathbf{y}^j) \log(1 - f_\theta(\mathbf{x}^j)). \tag{1}$$

When adapting to a new task $\mathcal{T}_i$, the trainable model parameters θ become *task-specific* parameters, namely $\hat{\theta}_i$. These task-specific parameters are computed with n gradient-step updates, with the following rule for one gradient update: $\hat{\theta}_i = \theta - \alpha \nabla_\theta \mathcal{L}_{\mathcal{T}_i}(f_\theta)$, which is the inner-loop update and α is the hyperparameter for the step size. Next, the model meta-parameters θ are optimized for the performance of $f_{\hat{\theta}_i}$, using the query set D_i^{ts} and the task-specific parameters $\hat{\theta}_i$ as initialization of the model:

$$\min_\theta \sum_{\mathbf{x}^j, \mathbf{y}^j \sim \mathcal{D}_i^{ts}} \mathcal{L}_{\mathcal{T}_i}(f_{\hat{\theta}_i}) = \sum_{\mathbf{x}^j, \mathbf{y}^j \sim \mathcal{D}_i^{ts}} \mathcal{L}_{\mathcal{T}_i}(f_{\theta - \alpha \nabla_\theta \mathcal{L}_{\mathcal{T}_i}(f_\theta)}), \tag{2}$$

akin to [11] but in a multimodal setting. The meta-optimization across all tasks $\mathcal{T}_i$ is performed using the stochastic gradient descent update rule, as follows: $\theta \leftarrow \theta - \beta \nabla_\theta \sum_{\mathbf{x}^j, \mathbf{y}^j \sim \mathcal{D}_i^{ts}} \mathcal{L}_{\mathcal{T}_i(f_{\hat{\theta}_i})}$, where β is the step size hyperparameter.

During inference, or the meta-test stage in meta-learning parlance, we are given new multimodal few-shot tasks with previously unseen objects. The support set is used for fast adaptation of the meta-parameters θ to the new task, followed by measuring the performance on the query set. Conditioned

Table 1: Comparison with the Frozen [3] baselines on Real-Name and Open-Ended miniImageNet 2- and 5-way setting; expressed in accuracy(%). The episodically trained models are outperforming the Frozen baselines, both for cross-domain and in-domain few-shot settings.

			Real-Name 2-way		**Open-Ended 2-way**		**Real-Name 5-way**		**Open-Ended 5-way**	
Methods	episodic	cross-domain	1-shot	5-shots	1-shot	5-shots	1-shot	5-shots	1-shot	5-shots
Frozen w/o task ind	✗	✓	1.7	-	29.0	-	0.9	-	18.0	-
Frozen w/ task ind	✗	✓	33.7	66.0	53.4	58.9	14.5	33.8	20.2	21.3
Ours	✗	✗	35.6	65.7	50.2	57.5	15.2	39.6	18.9	22.0
	✗	✓	37.3	66.0	52.5	59.0	19.2	40.3	20.9	25.0
	✓	✓	45.3	69.8	53.6	63.4	24.7	41.8	24.8	28.5
	✓	✗	**48.2**	**72.3**	**58.7**	**65.8**	**29.0**	**43.2**	**25.1**	**29.6**
ANIL upper-bound	-	-	73.9	84.2	-	-	45.5	62.6	-	-

Table 2: Comparison with the Frozen baseline [3] on Real-Fast VQA and Fast-VQA 2-way settings, in accuracy(%). Our episodically trained models outperform their counterparts, both for cross-domain and in-domain few-shot settings.

			Real-Fast VQA		**Fast-VQA**	
Methods	episodic	cross-domain	1-shot	5-shots	1-shot	5-shots
Frozen	✗	✓	7.8	10.5	2.8	7.9
Ours	✗	✗	5.4	9.1	2.5	7.1
	✗	✓	6.9	10.7	3	8
	✓	✓	8.5	13	5.2	8.6
	✓	✗	**9.7**	**13.2**	**5.7**	**9.3**

on the context, the generation of the answer for each query sample is done in an open-ended autoregressive manner, by using top-k nucleus sampling [20] for sampling words from the language model. To obtain the final performance, we take the average of the accuracy over all query samples from all meta-test tasks.

3 Experiments

We conduct systematic experiments to evaluate how a meta-learned model performs in multimodal few-shot settings. Specifically, we test the ability of fast adaptation as a main characteristic of meta-learning [11] by quantifying the fast binding of visual concepts and words and visual-question answering with limited examples.

3.1 Experimental Setup

Datasets & Settings To design a meta-learning setting for multimodal few-shot learning, the datasets have to be structured into sequences of tasks, as explained in 2.1. In practise, any dataset can be suited for few-shot meta-learning, as long as there is an available object information based on which the tasks can be constructed. For meta-training, we use the COCO2017 captioning dataset [21] and restructure it to construct tasks in N-way, k-shot manner based on the N object categories present in the support set. Then, for meta-test we use the four datasets introduced in [3], namely, Real-Name and Open-Ended miniImageNet; and Real-Fast and Fast-VQA. This is an example for *cross-domain* few-shot learning setting. We also consider *in-domain* few-shot setting, where the meta-training and meta-test partitions are entirely derived from the mentioned multimodal few-shot datasets [3]. Additionally we experiment with two different training procedures: the proposed *episodic* meta-learning and a standard mini-batched, *non-episodic* one.

Implementation Details The vision encoder is implemented as CLIP [14] with the Vision Transformer (ViT/B-32) [22] as a backbone model, yielding visual features of size 512. The language model is implemented as GPT-2 [23], with word embeddings of size 768, which is also the size of the visual prefix. The meta-mapper is initialized following Xavier uniform initialization [24]. For the meta-learning specific hyperparameters, we empirically determined to have five gradient-update steps with a learning rate of 0.01. The meta-update is performed with AdamW [25] with a meta-learning rate of 0.001 and 4 tasks in each meta-batch. The model is trained end-to-end on one NVIDIA GTX

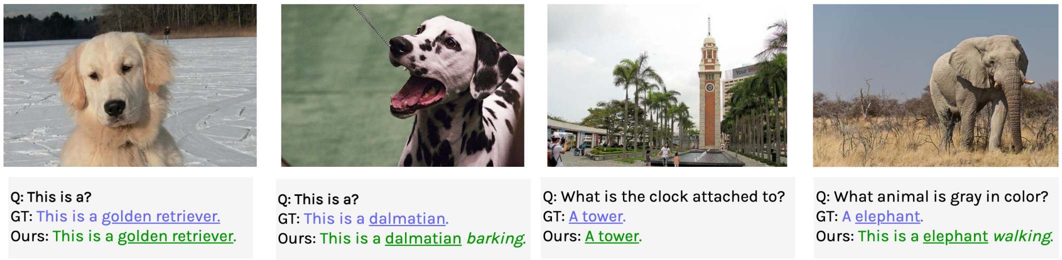
Q: This is a?
GT: This is a golden retriever.
Ours: This is a golden retriever.
Q: This is a?
GT: This is a dalmatian.
Ours: This is a dalmatian *barking*.
Q: What is the clock attached to?
GT: A tower.
Ours: A tower.
Q: What animal is gray in color?
GT: A elephant.
Ours: This is a elephant *walking*.

Figure 3: Examples of query set images from Real-Name miniImageNet (first two) and Real-Fast VQA (last two), with their question (Q), ground-truth (GT) and answers generated by our model (Ours).

1080Ti GPU, in less than 2 hours, which shows the benefit of the light-weight framework. The total number of trainable parameters of our model is less than two million, which is orders of magnitude lower than Frozen.

3.2 Results & Discussion

Fast binding of visual concepts to words The experiments on Real-Name and Open-Ended miniImageNet measure to what extent the multimodal meta-learner is able to bind visual concepts to words. Table 1 shows the 2-way and 5-way accuracy in 1 and 5 shots on both datasets. From the tables, we observe that our multimodal meta-learner is able to largely outperform Frozen [3], even without using a fixed task induction. This shows the advantage of having a meta-learned visual prefix, in contrast to just reshaping the vision encoder output as a prefix to language models. Specifically, the meta-learned prefix is able to collect shared meta-knowledge from related instances in the tasks, which is useful for narrowing down the search space in a learnable manner, instead of using a hand-engineered task induction. We believe that the open-ended approach is more promising due to its flexibility in reasoning about visual concepts, instead of relying on a pre-defined closed set of concepts. However, due to the magnitudes larger search space in text generation, compared to the one of conventional classifiers, it is still not possible to compete with their performance. Therefore, we use results from ANIL [12] as a reference upper bound to our approach.

VQA with limited labeled context The aim of the experiments on the Real-Fast and Fast-VQA 2-ways benchmarks is to evaluate the abilities of the multimodal meta-learner to answer more complex questions about the objects in the image. There is an implicit testing of the binding of visual concepts and words, since the query samples are designed in such a way to contain both categories from the support set in the query image, while the question is addressed to one of them. As we observe from the results in Table 2 with different number of shots, our multimodal meta-learner achieves improvements over Frozen [3], showing once more the benefit of the meta-knowledge and the ability to adapt fast to new tasks.

Qualitative results In Figure 3, we show examples of query images with the questions and answers at inference time. The capability of the multimodal meta-learner to bind visual concepts to words is apparent: the model is able to connect the visual concepts in the image not only to dalmatian as stated in the ground-truth, but also to the word barking. This observation suggests that the model can leverage visual concepts, not necessarily represented by the ground-truth sentence, which are still accurate and in many cases capture the image contents even better.

4 Conclusion

In this paper, we present a novel meta-learning approach for multimodal few-shot learning. Particularly, we introduce a light-weight meta-mapper which acts as a bridge between frozen vision and language models, and is trained in a meta-learning manner. The meta-mapper accrues shared meta-knowledge from related tasks into a learnable visual prefix, which is used to steer the language model into generating relevant outputs, without using a hand-engineered task induction. Our experiments verify the effectiveness of our method by outperforming the baseline on several multimodal few-shot benchmarks, fostering further research in multimodal few-shot meta-learning.

Acknowledgement

This work is financially supported by the Inception Institute of Artificial Intelligence, the University of Amsterdam and the allowance Top consortia for Knowledge and Innovation (TKIs) from the Netherlands Ministry of Economic Affairs and Climate Policy.

References

[1] Jürgen Schmidhuber. *Evolutionary principles in self-referential learning, or on learning how to learn: the meta-meta-... hook.* PhD thesis, Technische Universität München, 1987.

[2] Y. Bengio, S. Bengio, and J. Cloutier. Learning a synaptic learning rule. In *IJCNN-91-Seattle International Joint Conference on Neural Networks*, volume ii, pages 969 vol.2–, 1991.

[3] Maria Tsimpoukelli, Jacob L Menick, Serkan Cabi, SM Eslami, Oriol Vinyals, and Felix Hill. Multimodal few-shot learning with frozen language models. *Advances in Neural Information Processing Systems*, 34:200–212, 2021.

[4] 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.

[5] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[6] Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for generation. *arXiv preprint arXiv:2101.00190*, 2021.

[7] Sebastian Thrun and Lorien Pratt. *Learning to learn*. Springer Science & Business Media, 2012.

[8] Marcin Andrychowicz, Misha Denil, Sergio Gomez, Matthew W Hoffman, David Pfau, Tom Schaul, Brendan Shillingford, and Nando De Freitas. Learning to learn by gradient descent by gradient descent. *Advances in neural information processing systems*, 29, 2016.

[9] Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Daan Wierstra, et al. Matching networks for one shot learning. *Advances in neural information processing systems*, 29, 2016.

[10] Sachin Ravi and H. Larochelle. Optimization as a model for few-shot learning. In *ICLR*, 2017.

[11] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *International conference on machine learning*, pages 1126–1135. PMLR, 2017.

[12] Aniruddh Raghu, Maithra Raghu, Samy Bengio, and Oriol Vinyals. Rapid learning or feature reuse? towards understanding the effectiveness of maml. In *International Conference on Learning Representations*, 2019.

[13] Han-Jia Ye, Hexiang Hu, De-Chuan Zhan, and Fei Sha. Few-shot learning via embedding adaptation with set-to-set functions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8808–8817, 2020.

[14] 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*, pages 8748–8763. PMLR, 2021.

[15] Ron Mokady, Amir Hertz, and Amit H Bermano. Clipcap: Clip prefix for image captioning. *arXiv preprint arXiv:2111.09734*, 2021.

[16] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *CVPR*, 2022.

[17] Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. *arXiv preprint arXiv:2203.12119*, 2022.

[18] Yoad Tewel, Yoav Shalev, Idan Schwartz, and Lior Wolf. Zero-shot image-to-text generation for visual-semantic arithmetic. *arXiv preprint arXiv:2111.14447*, 2021.

[19] Juho Lee, Yoonho Lee, Jungtaek Kim, Adam Kosiorek, Seungjin Choi, and Yee Whye Teh. Set transformer: A framework for attention-based permutation-invariant neural networks. In *International Conference on Machine Learning*, pages 3744–3753. PMLR, 2019.

[20] Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. The curious case of neural text degeneration. *arXiv preprint arXiv:1904.09751*, 2019.

[21] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

[22] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

[23] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2019.

[24] Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In Yee Whye Teh and Mike Titterington, editors, *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, volume 9 of *Proceedings of Machine Learning Research*, pages 249–256, Chia Laguna Resort, Sardinia, Italy, 13–15 May 2010. PMLR.

[25] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2018.

[26] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[27] Dzmitry Bahdanau, KyungHyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *ICLR*, 2015.

[28] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of the 2019 Conference of the North*, 2019.

[29] Xiaohua Zhai, Xiao Wang, Basil Mustafa, Andreas Steiner, Daniel Keysers, Alexander Kolesnikov, and Lucas Beyer. Lit: Zero-shot transfer with locked-image text tuning. *arXiv preprint arXiv:2111.07991*, 2021.

[30] Zirui Wang, Jiahui Yu, Adams Wei Yu, Zihang Dai, Yulia Tsvetkov, and Yuan Cao. Simvlm: Simple visual language model pretraining with weak supervision. *ArXiv*, abs/2108.10904, 2021.

[31] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning*, pages 4904–4916. PMLR, 2021.

[32] Jaemin Cho, Jie Lei, Hao Tan, and Mohit Bansal. Unifying vision-and-language tasks via text generation. In *International Conference on Machine Learning*, pages 1931–1942. PMLR, 2021.

[33] Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt tuning. *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 2021.

[34] Adam Santoro, Sergey Bartunov, Matthew Botvinick, Daan Wierstra, and Timothy Lillicrap. Meta-learning with memory-augmented neural networks. In *ICML*, 2016.

[35] Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. *Advances in neural information processing systems*, 30, 2017.

[36] Erin Grant, Chelsea Finn, Sergey Levine, Trevor Darrell, and Thomas Griffiths. Recasting gradient-based meta-learning as hierarchical bayes. In *ICLR*, 2018.

[37] Flood Sung, Yongxin Yang, Li Zhang, Tao Xiang, Philip HS Torr, and Timothy M Hospedales. Learning to compare: Relation network for few-shot learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1199–1208, 2018.

[38] Nikhil Mishra, Mostafa Rohaninejad, Xi Chen, and Pieter Abbeel. A simple neural attentive meta-learner. *ICLR*, 2018.

[39] Adam Santoro, Sergey Bartunov, Matthew Botvinick, Daan Wierstra, and Timothy Lillicrap. Meta-learning with memory-augmented neural networks. 2016.

[40] Tsendsuren Munkhdalai and Hong Yu. Meta networks. In *ICML*, pages 2554–2563, 2017.

[41] Tsendsuren Munkhdalai, Xingdi Yuan, Soroush Mehri, and Adam Trischler. Rapid adaptation with conditionally shifted neurons. In *ICML*, 2018.

[42] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

Appendix

A Related Works

Large-scale vision and language models Large-scale language models emerged since the introduction of Transformers [26] and the attention mechanism [27] to successfully deal with long-range dependencies in sequences. The field has seen significant progress over the last years [28, 23, 5], which also initiated the development of similar strategies for vision [14, 29, 22] and multimodal models [30, 31, 18, 15, 4]. Unlike these approaches, which are trained with specific objectives and tasks, our work focuses on unifying them through meta-learning with a single objective for text generation [32]. Similarly as [3], our approach is modular and can accommodate any separately pre-trained vision encoder and pre-trained language model. In few-shot scenarios, these large models are steered into producing a desired output by using the idea of prompting [5, 3]. Standard prompting prepends fixed task instructions and a few examples as prompt and then generates the output from the language model. Instead of optimizing over a fixed set of examples, prefix tuning [6, 33] aims to optimize the instruction as learnable embeddings. Motivated by this, and by the need to develop an efficient way of bridging large models, we adopt a learnable visual prefix, which in our case is meta-learned.

Meta-learning for few-shot image classification tasks Meta-learning for few-shot learning [10, 34, 11, 9, 35] addresses the fundamental challenge of generalizing across tasks with limited labelled data. Meta-learning approaches for few-shot learning acquire inductive biases and adopt them for individual tasks in different ways [36]. Existing meta-learning algorithms are typically categorized as follows: (i) metric-based, focusing on learning a common embedding space and deriving prototypes as meta-knowledge [9, 35, 37] (ii) memory-based, using an external memory module as meta-knowledge to quickly adapt to new tasks [38, 39, 40, 41], and (iii) optimization-based, aiming to learn a good model initialization across tasks as meta-knowledge, which can be used to efficiently adapt to new tasks [10, 11, 11, 36]. Our approach is positioned in this last category, as it is modality-agnostic and offers greater flexibility.

Multimodal few-shot learning Few-shot learning by combining both vision and language, has only emerged recently with the introduction of the multimodal few-shot learner Frozen [3]. In particular, Frozen is based on the idea of in-context few-shot prompting of language models, meaning that the prompt consists of task induction and interleaved sequences of images and their captions, representing the context. This in-context few-shot prompting paradigm is considered as one of the possible approaches to deal with few-shot learning scenarios in language models [3, 5, 4]. A recently proposed model [4], is following similar multimodal few-shot learning settings. However, they are training a vision-language model of 70 billion parameters, which is proven to be successful due to its scale and the amount of pre-training data. Our goals highly differ since our model aims to handle a limited labeled space during training and to adapt to new tasks. In particular, we define optimization-based meta-learning steps [10, 11, 11, 36], by using the context samples to fine-tune the meta-mapper and evaluate on the query samples.

B Vision encoder details

For the pre-trained vision encoder we adopt CLIP [14], due to its already proven performance and large web-scale multimodal pre-training. CLIP is considered as a multimodal architecture, as it consists of 1) a vision encoder, which can be a ResNet [42] or a Vision Transformer (ViT) [22] and 2) a text encoder implemented as a Transformer [23]. The pre-training is done in a contrastive manner, on a large dataset of 400 million pairs of image-caption, with the aim to minimize the distance of corresponding image-caption pairs in the embedding space and to maximize the distance for non-corresponding pairs.

In this work, we use the vision encoder stream with a base ViT backbone, comprised of 12 layers, 512-dimensions wide, each one with 12 attention heads. The size of the input images is 224×224 and are split into image patches, each one with dimensions 32×32, yielding 49 flattened patches and one leading special token. We keep the backbone entirely frozen and use the special token as a visual encoding, since it holistically represents the image.

Algorithm 1 Meta-training the multimodal few-shot meta-learner

```
Require: $p(\mathcal{T})$: distribution over N-way, k-shot tasks
Require: $\theta \leftarrow$ random initialization
while not done do
    Sample a batch of tasks $\mathcal{T}_i \sim p(\mathcal{T})$,
    for all $\mathcal{T}_i$ do
        $\mathcal{D}_i^{tr}, \mathcal{D}_i^{ts} \leftarrow \mathcal{T}_i$
        Evaluate $\nabla_\theta \mathcal{L}_{\mathcal{T}_i}(f_\theta)$ using $\mathcal{D}_i^{tr}$.
        for $i = 1$ to $n$ do                                ▷ $n$ is number of gradient steps
            Compute adapted parameters $\hat{\theta}_i$ with a gradient-descent step $\hat{\theta}_i = \theta - \alpha \nabla_\theta \mathcal{L}_{\mathcal{T}_i}(f_\theta)$.
        end for
        Use adapted parameters $\hat{\theta}_i$ and $\mathcal{D}_i^{ts}$ for meta-optimization.
    end for
    Update meta-parameters $\theta$ across all tasks $\mathcal{T}_i$ with $\theta \leftarrow \theta - \beta \nabla_\theta \sum_{\mathbf{x}^j, \mathbf{y}^j \sim \mathcal{T}_i} \mathcal{L}_{\mathcal{T}_i(f_{\hat{\theta}_i})}$.
end while
```

Algorithm 2 Meta-test the multimodal few-shot meta-learner

```
Require: $p(\mathcal{T})$: distribution over N-way, k-shot tasks
Require: $\theta \leftarrow$ meta-learned parameters in the meta-training stage
while not done do
    Sample a task $\mathcal{T}_i \sim p(\mathcal{T})$,
    $\mathcal{D}_i^{tr}, \mathcal{D}_i^{ts} \leftarrow \mathcal{T}_i$                  ▷ support set and query accordingly
    Evaluate $\nabla_\theta \mathcal{L}_{\mathcal{T}_i}(f_\theta)$ using $\mathcal{D}_i^{tr}$      ▷ $\mathcal{L}_{\mathcal{T}_i}(f_\theta)$ is the cross-entropy loss
    for $i = 1$ to $n$ do                                    ▷ $n$ is number of gradient steps
        Compute adapted parameters $\hat{\theta}_i$ with a gradient-descent step $\hat{\theta}_i = \theta - \alpha \nabla_\theta \mathcal{L}_{\mathcal{T}_i}(f_\theta)$.
    end for
    Use adapted parameters $\hat{\theta}_i$ and $\mathcal{D}_i^{ts}$ for computing the final accuracy
end while
```

C Language model details

The pre-trained language model that we employ is GPT-2 [23], particularly the small version with 117M parameters. Its architecture is following a Transformer decoder [26] with 12 layers and a word embedding dimension of 768. The model is pre-trained on a very large corpus of English data in a self-supervised fashion with a standard language modelling objective. Since the model performs best at what it was pre-trained for, which is generating text from a given prompt in an autoregressive manner, we employ it in a similar fashion. In particular, we use the word embedding layer to transform each word token into a continuous word embedding, and the full stack of Transformer decoder layers to parameterize the probability distribution over the vocabulary word tokens. To obtain the next word token we sample from the probability distribution over the vocabulary with top-k nucleus sampling as in [20]. To build a more efficient architecture, similar to the vision stream, the language model is kept entirely frozen.

D Multimodal meta-learning details

To design a meta-learning setting for the multimodal few-shot learning, we re-purpose an image captioning dataset, with available meta-data about the object categories, to fit the meta-learning criteria [10]. In particular, we use either COCO2017 captioning dataset [21] to obtain cross-domain experimental setup, or the multimodal few-shot datasets [3] for the standard in-domain meta-learning setup. The partitioning into meta-training and meta-test tasks is illustrated in Figure 4. We start by splitting the full dataset into task partitions according to the object categories in the images in the scope of their own meta-training and meta-test partitions. The sampling of tasks for both stages is straightforward due to the provided object information and the captions targeted for those objects. Note that following [11] the samples in the query set during meta-training should be at least 15 per category, since the optimization of the meta-parameters is done based on those samples.

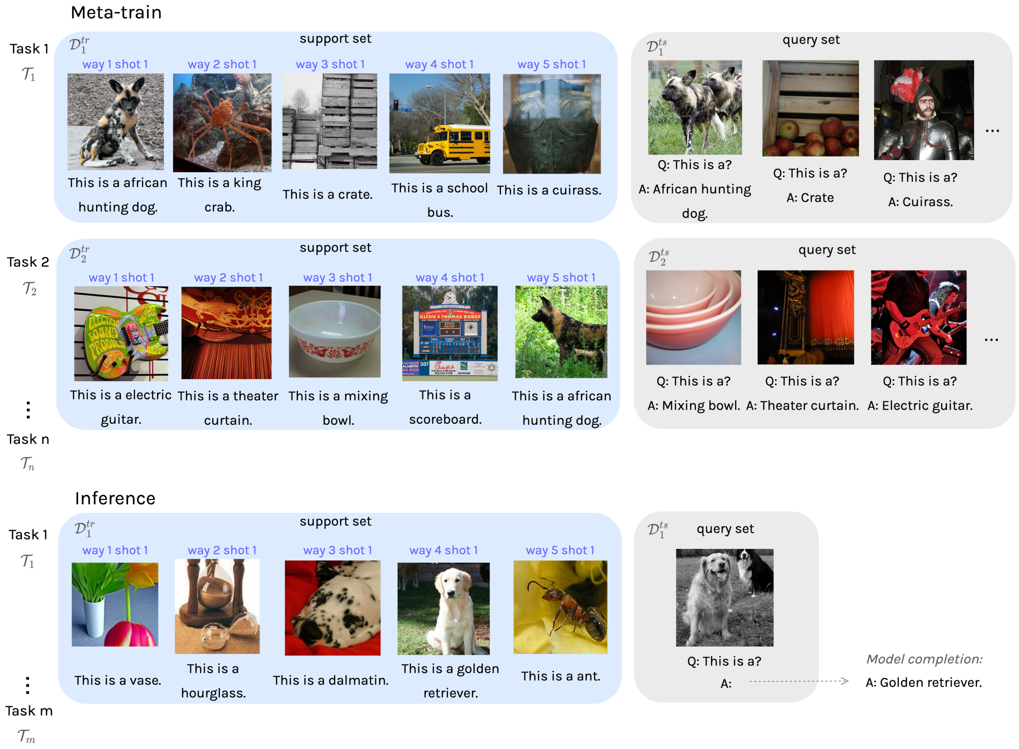

Figure 4: Example of the new multimodal few-shot meta-learning setting, illustrating the in-domain 2-way 1-shot problem with the Real-Name miniImageNet. The top represents the meta-training stage and the bottom part is the meta-test stage. In meta-training, the blue box indicates the support set samples which consist of an image-caption pair. The gray box indicates the query set samples.

The detailed optimization process in the meta-training stage is described in Algorithm 1. Similarly, the adaptation stage using the meta-test partitions is described in Algorithm 2.