META LEARNING TO BRIDGE VISION AND LANGUAGE MODELS FOR MULTIMODAL FEW-SHOT LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal few-shot learning is challenging due to the large domain gap between vision and language modalities. Existing methods are trying to communicate visual concepts as prompts to frozen language models, but rely on hand-engineered task induction to reduce the hypothesis space. To make the whole process learnable, we introduce a multimodal meta-learning approach. Specifically, our approach decomposes the training of the model into a set of related multimodal few-shot tasks. We define a meta-mapper network, acting as a meta-learner, to efficiently bridge frozen large-scale vision and language models and leverage their already learned capacity. By updating the learnable parameters only of the meta-mapper, it learns to accrue shared meta-knowledge among these tasks. Thus, it can rapidly adapt to newly presented samples with only a few gradient updates. Importantly, it induces the task in a completely data-driven manner, with no need for a hand-engineered task induction. We evaluate our approach on recently proposed multimodal few-shot benchmarks, measuring how rapidly the model can bind novel visual concepts to words and answer visual questions by observing only a limited set of labeled examples. The experimental results show that our meta-learning approach outperforms the baseline across multiple datasets and various training settings, while being computationally more efficient.

1 INTRODUCTION

Learning quickly by observing a few examples in a multimodal environment is an integral part of human intelligence (Schmidhuber, 1987; Bengio et al., 1991). Yet, it is quite challenging for current vision and language models to deal with a limited labeled space, when performing multimodal few-shot learning (Tsimpoukelli et al., 2021; Alayrac et al., 2022). On the contrary, language-only models have already flourished over the past years, especially when being transferred to a limited labeled space (Brown et al., 2020; Perez et al., 2021; Min et al., 2022), as a result of the large-scale pre-training and huge model capacity. Such advances in natural language processing inspired similar efforts in the vision domain, yielding large vision models with impressive few-shot and zero-shot image classification capabilities (Radford et al., 2021; Zhai et al., 2021; Jia et al., 2021). However, due to the large domain gap between vision and language modalities, it is non-trivial to directly embody few-shot capabilities into multimodal settings, which is the main motivation of this work.

One of the main challenges for bridging this gap is finding a proper mechanism for communicating visual concepts to a language model, by accruing shared knowledge from sequences of multimodal tasks. The Frozen model (Tsimpoukelli et al., 2021) is the first multimodal few-shot learner which tackles this challenge by taking inspiration from language models able to do in-context learning (Brown et al., 2020). This requires a task induction provided as a sentence followed by context data samples, to reduce the hypothesis space for open-ended image interpretation. This might not be a shortcoming for simpler tasks, like binary decisions; however it becomes an obstacle for more complicated tasks (Li & Liang, 2021), since the task induction has to be engineered each time. We hypothesize that the task can be induced from the data itself in a completely learnable manner.

Meta-learning or *learning to learn* (Schmidhuber, 1987; Thrun & Pratt, 2012; Andrychowicz et al., 2016) comes as a natural solution to any few-shot settings. While it has been extensively studied in unimodal settings, particularly for few-shot image classification (Vinyals et al., 2016; Ravi & Larochelle, 2017; Finn et al., 2017; Raghu et al., 2019; Ye et al., 2020), meta-learning remains almost

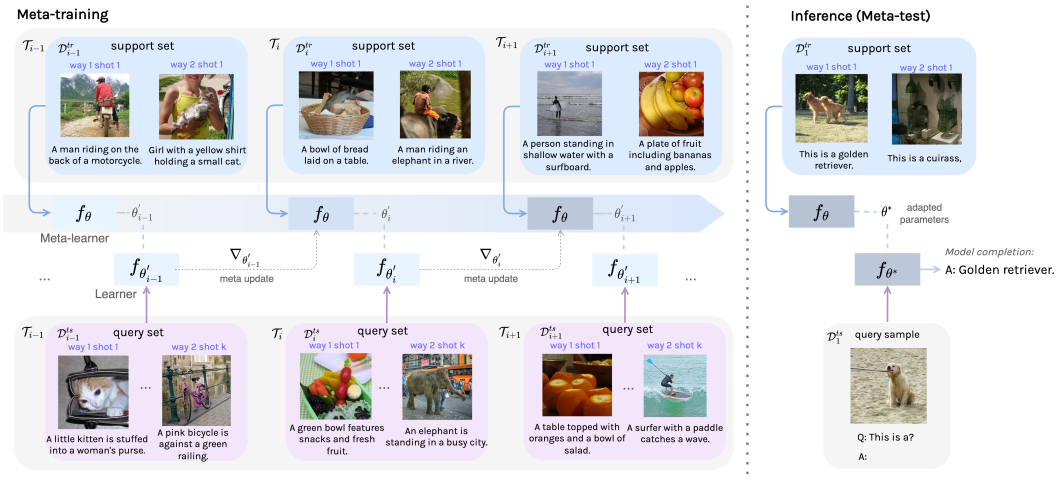


Figure 1: Multimodal few-shot meta-learning task for an example of a 2-way 1-shot setting, with two categories (ways) present in the support set images, each represented with one sample (shot). Given a batch of tasks $\mathcal{T}_i$, the support set is first used to obtain task-specific model parameters $\theta_i^{'}$ for each task by few gradient-step updates, which are then used together with the query set samples to perform a meta-update step for updating the meta-parameters θ. After the meta-training is finished, for a new given task, the meta-trained model is used for inference by further adapting the meta-learned model with the support set, and measuring the performance on unseen query samples.

unexplored for multimodal few-shot settings. Similar to unimodal settings, empowering a multimodal few-shot learner with the ability to accrue knowledge across tasks, would assist in building internal multimodal representations broadly suitable for many tasks. These representations could serve as a bridge between the different modalities and assist in learning quickly new tasks by observing only limited labeled examples.

Motivated by this, we define a novel multimodal few-shot meta-learning approach, to bridge the gap between vision and language modalities, illustrated in Figure 5. Specifically, our approach adopts publicly available pre-trained large vision encoders (Radford et al., 2021) and language models (Brown et al., 2020), which are kept frozen to take advantages of their already-learned reasoning capability (Tsimpoukelli et al., 2021; Mokady et al., 2021; Zhou et al., 2022; Jia et al., 2022; Tewel et al., 2021; Zhai et al., 2021). In doing so, our method avoids the huge computational resources required by those models during training and their dependency on large datasets. Unlike prior multimodal few-shot learners (Tsimpoukelli et al., 2021; Alayrac et al., 2022), our approach decomposes the training of the model into sequential observing of multimodal few-shot tasks, in the spirit of meta-learning. During the meta-training stage, the model translates the visual representations into a visual prefix for the language model, by using a lightweight meta-mapper network. This network serves as a multimodal bridge between the large vision and language models and is entirely built from self-attention layers (Lee et al., 2019). Essentially, the aim of the meta-mapper is to collect shared meta-knowledge from related tasks, by mapping the visual representation into the latent space of the language model. Then, during inference, or meta-test according to the meta-learning parlance, the model is able to induce the task in a fully data-driven manner by observing few labeled examples and thus entirely removes the need for hand-engineered task inductions.

In summary, we contribute in four major aspects: (*i*) We introduce meta-learning to multimodal few-shot learning, which enables fast adaptation and efficient learning of multimodal few-shot tasks. (*ii*) We present a multimodal meta learner by using a lightweight meta-mapper, which learns to bridge a large frozen vision and language backbones. To the best of our knowledge, this is the first meta-learning based model to solve multimodal few-shot tasks. (*iii*) We design a new, meta-learning setting for multimodal few-shot learning by re-organizing existing datasets and following suitable benchmarks. (*iv*) We demonstrate through systematic experiments on those benchmarks that our model, using only the small trainable meta-mapper with frozen backbones, yields strong multimodal few-shot performance, while being computationally more efficient.

2 RELATED WORKS

Large-scale language models emerged since the introduction of attention (Bahdanau et al., 2015) and Transformers (Vaswani et al., 2017) to successfully deal with long-range dependencies in sequences. The field has seen significant progress over the last years (Devlin et al., 2019; Radford et al., 2019; Brown et al., 2020), which also initiated the development of similar strategies for vision (Dosovitskiy et al., 2021; Radford et al., 2021; Zhai et al., 2021) and multimodal models (Lu et al., 2019; Li et al., 2020; Wang et al., 2021; Jia et al., 2021; Tewel et al., 2021; Mokady et al., 2021; Eichenberg et al., 2021; Shen et al., 2021). In current few-shot scenarios, the language models are steered into producing a desired output by using the concept of prompting (Brown et al., 2020; Tsimpoukelli et al., 2021; Alayrac et al., 2022). The approach is to prepend fixed task induction with a few examples as a prompt and then generate the output from the language model. Instead of optimizing over a fixed set of task inductions and examples, also known as in-context learning, we aim to optimize the context representation (Li & Liang, 2021; Lester et al., 2021), in the form of meta-learned visual prefix.

Meta-learning for few-shot learning (Ravi & Larochelle, 2017; Santoro et al., 2016b; Finn et al., 2017; Vinyals et al., 2016; Snell et al., 2017) addresses the fundamental challenge of generalizing across tasks with limited labelled data. The aim is learning to acquire inductive biases and perform fast adaptation to new tasks (Grant et al., 2018). Meta-learning algorithms are typically categorized as: (i) metric-based, focusing on learning a common embedding space and deriving prototypes as meta-knowledge (Vinyals et al., 2016; Snell et al., 2017; Sung et al., 2018) (ii) memory-based, using an external memory as meta-knowledge to quickly adapt to new tasks (Mishra et al., 2018; Santoro et al., 2016a; Munkhdalai & Yu, 2017; Munkhdalai et al., 2018) 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 samples (Ravi & Larochelle, 2017; Finn et al., 2017; Grant et al., 2018). Our approach is positioned in this last category, as it offers greater flexibility and is modality-agnostic.

Multimodal few-shot learning across vision and language modalities has emerged recently with the introduction of the first multimodal few-shot learner Frozen (Tsimpoukelli et al., 2021), followed by other innovative prompting and in-context learning strategies (Song et al., 2022; Jin et al., 2021). Particularly, Frozen performs in-context learning, which is considered as one of the possible approaches to deal with multimodal few-shot learning settings. Another recently proposed model named Flamingo (Alayrac et al., 2022) follows a similar in-context learning paradigm, by using a visual-language model of 70 billion parameters, which is proven to be successful due to its scale. Our goals differ since our aim is to perform fast adaptation to new tasks by acquiring meta-knowledge across tasks, with a lightweight model of less than two million trainable parameters. In particular, we define optimization-based meta-learning steps (Ravi & Larochelle, 2017; Finn et al., 2017; Grant et al., 2018), by using the context samples to adapt a small trainable mapper and evaluate on the query samples.

3 METHODOLOGY

We aim to train a model that can learn and quickly adapt to new multimodal tasks with limited labeled data under the meta-learning setting. In the next sections, we will first define the multimodal meta-learning setting, then explain our architecture, and finally describe how it is used during training and inference time.

3.1 PROBLEM FORMULATION

We start by defining our approach for multimodal few-shot settings, following a meta-learning approach. Different from standard supervised learning, in meta-learning we are dealing with a collection of meta-datasets, which are split into disjoint $\mathcal{D}_{meta-train}$ and $\mathcal{D}_{meta-test}$ partitions. We define a meta-training stage, where the model is trained on the meta-train $\mathcal{D}_{meta-train}$ partition, and a separate meta-test stage where the performance is measured on the $\mathcal{D}_{meta-test}$ partition. In particular, the meta-train $\mathcal{D}_{meta-train}$ set consists of meta-datasets $\mathcal{D}_i$, each one containing a pair of separate inner-train set $\mathcal{D}_i^{tr}$, also referred to as a support set and inner-test set $\mathcal{D}_i^{ts}$, referred to as a query set. This means that $\mathcal{D}_{meta-train} = \{(\mathcal{D}_1^{tr}, \mathcal{D}_1^{ts}), \ldots (\mathcal{D}_n^{tr}, \mathcal{D}_n^{ts})\}$. We refer to each meta-dataset pair (D_i^{tr}, D_i^{ts}) as a meta-task $\mathcal{T}_i$, following (Finn et al., 2017).

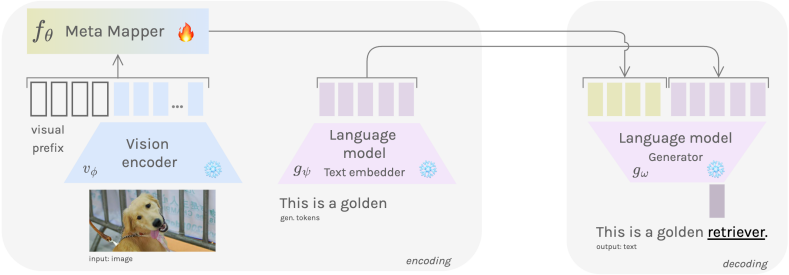


Figure 2: The architecture of the multimodal meta few-shot learner. It consists of three parts: frozen vision encoder v_ϕ; frozen language model with a text embedder g_ψ and a generator g_ω; and a meta-mapper f_θ with trainable meta-parameters θ. In the example shown, the model is generating the last word *retriever*, in an autoregressive manner.

When considering a k-shot, N-way setting, the support set $\mathcal{D}_i^{tr}$ of a single meta-task $\mathcal{T}_i$ consists of k labeled samples for each of the N-ways, where N is the number of object categories i.e. ways in few-shot parlance, meaning $\mathcal{D}_i^{tr} = \{(x_1^i, y_1^i) \dots (x_k^i, y_k^i)\}$, where x_j^i represents the image and y_j^i represents the caption. The query set $\mathcal{D}_i^{ts}$ is defined as $\mathcal{D}_i^{ts} = \{(x_1^i, q_1^i, a_1^i) \dots (x_m^i, q_m^i, a_m^i)\}$, where x_j^i is the image, q_m^i is an optional slot for an input text representation, like a question, and a_m^i is the output text representation, i.e an answer to the question.

We consider a distribution over tasks $p(\mathcal{T})$, where a single batch of tasks $\mathcal{T}_i \sim p(\mathcal{T})$ for the k-shot N-way problem, is sampled in the following fashion: firstly N object categories are randomly sampled from the pool of all meta-train $\mathcal{D}_{meta-train}$ samples. For the support set there are k samples chosen for each of the N categories, and for the query set there are m samples per category, with $m > k$.

3.2 Model Architecture

The architecture that we present in this paper is modular and consists of three components: a vision encoder, the meta-mapper and a language model, illustrated in Figure 2. It is trained in a fully autoregressive manner, meaning that it offers flexibility for the choice of a downstream generative multimodal task.

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

Meta-mapper network To map the visual encoding into the latent space of the language model in our multimodal few-shot setting, we use a set of l learnable parameters $p_i \in \mathbb{R}^{d_e}$. These parameters are also called visual prefix for the language model, with the same dimension d_e as the language embeddings. We prepend this visual prefix to the encoded visual features, yielding the following sequence $[p_1 \dots p_l, x_1, \dots x_n]$. Particularly, we view this representation as an ordered set of elements and we employ self-attention to simultaneously encode the whole set. For this purpose, we adopt the set multi-head attention block (Lee et al., 2019), which has the role of a meta-mapper with trainable meta-parameters θ, and is defined as:

$$\text{MetaMap}_\theta(Q, K, V) = \sigma(QK^T) * V, \tag{1}$$

where the pairwise dot-product QK^T measures the similarity amongst features, and is used for feature weighting computed through an activation function σ. Intuitively, each feature of V will get more weight if the dot-product between Q and K is larger. In Eq. 1, we have $Q = K = V = [p_1 \dots p_l, x_1, \dots x_n]$, meaning that the input of the meta-mapper is the sequence of features, and the output is a vector of learned parameters $p_1^* \dots p_k^*$, namely the visual prefix for the language model, as expressed in:

$$p_1^* \dots p_l^* = \text{MetaMap}_\theta([p_1 \dots p_l, x_1, \dots x_n]). \tag{2}$$

The self-attention layer, as the building block of this module, allows to retrieve meaningful information from the visual features $x_1, \dots x_n$, and accumulate it into $p_1^* \dots p_l^*$, due to its inherent weighting

mechanism of pairwise similarities between elements in the sequence. The meta-parameters of the meta-mapper are learned and shared across all tasks $\mathcal{T}$ in $D_{meta-train}$.

Language model The language model is defined as a function g_ω, which parametrizes a probability distribution over a text sequence $\mathbf{y}$. It consists of an embedding function g_ψ, with $\psi \in \mathbb{R}^{d_e}$, to embed each word token y_i of the sequence $\mathbf{y}$ into a word token embedding t_i, followed by a Transformer block conducting an autoregressive text generation. More precisely, 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 tokens conditioned on the prefix in an autoregressive manner:

$$t_{i+1} = g_\omega([p_1^* \dots p_l^*, t_1, \dots t_i]), i < m. \tag{3}$$

Since our approach aims to learn in a limited labeled space which might distort the learned parameters if we update them, we initialize the ω parameters with pre-trained parameters from a large-scale language model and keep them entirely frozen.

3.3 Meta-Training & Inference

To meta-train the model, we sample batches of multimodal few-shot learning tasks $\mathcal{T}_i \sim p(\mathcal{T})$, which consist of a support set denoted as D_i^{tr} and a query set D_i^{ts}. Here, for simplicity, we assume that our full model described in 3.2 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, defined as:

$$\mathcal{L}_{\mathcal{T}_i}(f_\theta) = \sum_{\mathbf{x}^j, \mathbf{y}^j \sim \mathcal{D}_i^{tr}} \mathbf{y}^j \log f_\theta(\mathbf{x}^j) + (1 - \mathbf{y}^j) \log(1 - f_\theta(\mathbf{x}^j)). \tag{4}$$

When adapting to a new task $\mathcal{T}_i$, the trainable meta parameters θ become *task-specific* parameters, namely $\theta_i^{'}$. These task-specific parameters are computed with N gradient-step updates, similar as in (Finn et al., 2017), with the following rule for one gradient update: $\theta_i^{'} = \theta - \alpha \nabla_\theta \mathcal{L}_{\mathcal{T}_i}(f_\theta)$. This is referred as the inner-loop update, where α is the hyperparameter for the step size. Next, the model meta-parameters θ are optimized for the performance of $f_{\theta_i^{'}}$, using the query set D_i^{ts} samples and the task-specific parameters $\theta_i^{'}$ as initialization of the model:

$$\min_\theta \sum_{\mathbf{x}^j, \mathbf{y}^j \sim \mathcal{D}_i^{ts}} \mathcal{L}_{\mathcal{T}_i}(f_{\theta_i^{'}}) = \sum_{\mathbf{x}^j, \mathbf{y}^j \sim \mathcal{D}_i^{ts}} \mathcal{L}_{\mathcal{T}_i}(f_{\theta - \alpha \nabla_\theta \mathcal{L}_{\mathcal{T}_i}(f_\theta)}), \tag{5}$$

which is called outer-loop optimization. The meta-optimization across all tasks $\mathcal{T}_i$ is performed using stochastic gradient descent update rule, as follows: $\theta \leftarrow \theta - \beta \nabla_\theta \sum_{\mathbf{x}^j, \mathbf{y}^j \sim \mathcal{D}_i^{ts}} \mathcal{L}_{\mathcal{T}_i}(f_{\theta_i^{'}})$, where β is the step size hyperparameter.

In the meta-test stage, we consider new multimodal few-shot tasks $\mathcal{T}_i$ with previously unseen objects, which also have a support set: $\mathcal{D}_i^{tr}$ for fast adaptation by fine-tuning the model meta-parameters θ to a given task, and a query set $\mathcal{D}_i^{ts}$ to evaluate the model on this task. Note that the generation of the answer for each query set sample is done in an open-ended autoregressive manner. Specifically, we use top-k nucleus sampling (Holtzman et al., 2019) to sample from the language model given the visual prefix of the image.

4 Experiments

4.1 Experimental Setup

Datasets To design a meta-learning setting for multimodal few-shot learning, the datasets have to be structured into sequences of tasks, as explained in section 3.1. The datasets introduced in Frozen (Tsimpoukelli et al., 2021) are representative examples for this type of multimodal few-shot datasets, where each task consists of a few labeled instances as a support set and unlabeled ones as a query set. 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. Therefore, for meta-training, we use the COCO2017 captioning dataset (Lin et al., 2014) and restructure it to construct tasks in N-way, k-shot manner based on the N object categories present in the images. For meta-test, we use the four

Table 1: Comparison with the Frozen (Tsimpoukelli et al., 2021) baselines on Real-Name and Open-Ended miniImageNet 2- and 5-way setting; expressed in accuracy(%). ANIL (Raghu et al., 2019) is used as an upper bound, since it is a discriminative approach as opposed to our open-ended generative one. Our episodically trained models are outperforming the Frozen baselines, both with and without domain-shift. The overall best performance is denoted in bold, whereas the same settings as the baseline are denoted in italic & bold.

			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 (Tsimpoukelli et al., 2021) baselines on 2- and 5-way Real-Name and Open-Ended miniImageNet for the 1-shot setting, with different number of repeats (reps); expressed in accuracy(%). Increasing the repeats in the support set does not show a huge benefit compared to when the number of shots are increased. The overall best performance is denoted in bold, whereas the same settings as the baseline are denoted in italic & bold.

			Real-Name 2-way		**Open-Ended 2-way**		**Real-Name 5-way**		**Open-Ended 5-way**	
Methods	episodic	cross-domain	1-rep	5-reps	1-rep	5-reps	1-rep	5-reps	1-rep	5-reps
Frozen w/ task ind	✗	✓	**63.0**	63.7	51.1	58.5	**33.8**	32.8	21.4	20.9
Ours	✗	✗	36.0	36.7	51.2	52.0	15.5	15.9	18.5	19.0
	✗	✓	*37.9*	*38.5*	***53.0***	*54.2*	*20.3*	*22.5*	***25.6***	***26.1***
	✓	✓	46.8	47.5	56.4	58.0	26.7	27.1	26.5	28.1
	✓	✗	62.1	**64.3**	**59.8**	**60.8**	31.3	**33.7**	**26.8**	**27.5**

mentioned datasets, namely, Real-Name miniImageNet, Real-Fast VQA, Open-Ended miniImageNet and Fast-VQA datasets, as they test the characteristics that make a good multimodal few-shot learner, such as fast binding of visual concepts to words and adaptation to new tasks.

Training & inference procedures We consider two different training procedures: *(i)* standard mini-batched training, which we refer to as a non-episodic; *(ii)* our proposed procedure by observing batches of multimodal tasks, which we refer to as episodic training; These training procedures are considered within two scenarios determined by the domain shift emerging during training and inference: *(i)* cross-domain multimodal few-shot learning, when the training and test datasets are of different distribution. Here the model is trained on COCO2017 (both in episodic and non-episodic manner) and the performance is measured on the full multimodal few-shot datasets; *(ii)* in-domain multimodal few-shot learning (standard meta-learning) setup, where the training and test partitions are of the same domain. This means that the training and test partitions, namely meta-training and meta-test tasks are constructed from the same multimodal few-shot dataset.

Implementation details The vision encoder is implemented as CLIP (Radford et al., 2021) with the Vision Transformer (ViT/B-32) (Dosovitskiy et al., 2021) as a backbone model, yielding visual features of size 512. The language model is implemented as GPT-2 (Radford et al., 2019), with word embeddings of size 768. The learnable visual prefix is empirically set to length 4 with dimensions of 768, which are specific to the language model. The meta-mapper is initialized following Xavier uniform initialization (Glorot & Bengio, 2010). For the meta-learning specific hyperparameters, we empirically determined to have five inner-loop gradient-updates with a learning rate of 0.01. The meta-update is performed with AdamW (Loshchilov & Hutter, 2018) with a meta-learning rate of 0.001 and 4 tasks in each meta-batch. All hyperparameters are tuned using the query sets in the meta-training partition which acts as a validation set.

Regarding the computational resources, our model is trained end-to-end on one NVIDIA GTX 1080Ti GPU, in less than 2 hours, showing the benefit of the lightweight framework. The total number of trainable parameters is approximately two million, which is order of magnitudes lower than the

Table 3: Comparison with the Frozen baseline (Tsimpoukelli et al., 2021) on 2-way Real-Fast VQA and Fast-VQA, in accuracy(%). Our episodically trained models outperform their counterparts, both with and without domain shift.

			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**

Table 4: Benefit of the meta-knowledge in accuracy(%) on the four multimodal few-shot datasets under the 2-way setting. Accruing meta-knowledge is crucial for our approach.

	Real-Name		**Open-Ended**		**Real-Fast VQA**		**Fast-VQA**	
Variants	1-shot	5-shots	1-shot	5-shot	1-shot	5-shots	1-shot	5-shot
w/o meta-knowledge	0.0	0.06	0.0	0.05	0.0	0.03	0.0	0.06
w/ meta-knowledge	**48.2**	**72.3**	**58.7**	**65.8**	**9.7**	**13.2**	**5.7**	**9.3**

Frozen model. The full implementation is in PyTorch and HuggingFace (Wolf et al., 2020) and will be publicly realised at https://github.com/<redacted>/<redacted>.

4.2 Results & Discussion

Binding of visual concepts and words The experiments conducted 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. We observe that our multimodal meta-learner is able to largely outperform Frozen, even without using an engineered 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 an engineered task instruction. Having this adaptable component, makes it easier for the model to later adapt to unseen visual concepts, without being explicitly instructed to do so. Also, similar to Frozen, we observe an upward trend when increasing the number of shots, for both the 2-way and 5-way settings, which confirms that having more distinct examples of the task increases the performance.

We believe that the open-ended approach is promising due to its flexibility in reasoning about visual concepts in an unconstrained manner, instead of relying on a pre-defined closed set of concepts. However, due to the magnitudes larger hypothesis space in the open-ended text generation (the full vocabulary of the language model), compared to the finite set of possible answers of conventional classifiers, such as ANIL (Raghu et al., 2019), it is not possible to perform a fair comparison to them. Therefore, we use their results as upper bound to our approach.

Increasing the number of repeats of support samples As an additional analysis, we experiment with the number of times each shot is repeated in the support set presented to the model, during inference i.e. the meta-test time. We observe from Table 2 that increasing the repeats in the support set, does not show as large benefit as increasing the number of distinct shots. This means that the meta-learner is able to accumulate more useful meta-knowledge by seeing different examples of images, even only once, rather than multiple repetitions of the same ones. Contrary to this observation, Frozen benefits more from these repetitions since they are used in combination with a hand-engineered task induction, which is crucial for their method.

Visual question-answering with limited labeled samples The aim of the experiments on the Real-Fast and Fast-VQA 2-ways benchmarks, presented in Table 3 with 1 and 5 shots, is to evaluate the abilities of the multimodal meta-learner to answer more complex queries 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

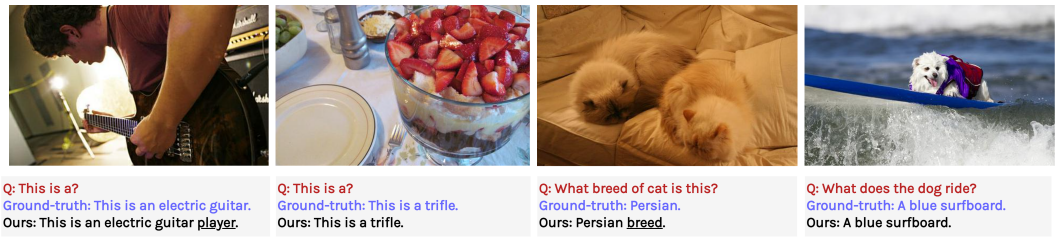


Figure 4: Qualitative examples of query set images from Real-Name miniImageNet (first two) and Real-Fast VQA (last two), with their question, ground-truth and answers generated by our model.

image, while the question is addressed in one of them. Moreover, the emphasis is on the ability of the meta-learner to reason about particular object properties, without being explicitly meta-trained to do so. As we observe from the results, our multimodal meta-learner achieves improvements over the Frozen baseline, showing once more the benefit of accruing meta-knowledge across tasks. Again, there is an upward trend in the performance when increasing the number of shots.

Episodic training of multimodal few-shot learners Another important observation from Tables 1 and 3 is that the episodic training improves the overall performance. The simple reason for this is having matched training and inference time conditions, as in standard supervised learning. More precisely, the multimodal few-shot datasets that are employed at inference time, are structured into tasks, so training the model in a similar manner by observing the tasks improves the performance.

Learning across domains Tables 1 and 3 provide insights into the more challenging cross-domain multimodal few-shot setting. Essentially, meta-training on the COCO2017 captioning dataset, helps the model to bind visual concepts with richer semantic descriptions, compared to when training on the multimodal few-shot datasets (which have simpler captions). However, according to the evaluation procedure that we follow, also employed by Frozen, the model is expected to generate the exact word as the ground-truth, penalizing any paraphrasing of words. Therefore, a COCO2017 meta-trained model exhibits lower accuracy from a quantitative perspective when transferred to a miniImageNet-based dataset. However, in many cases the generated sentences are qualitatively better and more detailed, as we discuss in the next paragraph.

Qualitative analysis In Figure 4, we show a examples of query images with the questions and answers at inference time, by the best version of our approach in Tables 1 and 3. 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 *electric guitar*, as stated in the ground-truth, but also to the word *player*. This demonstrates that the model is correctly steered by the meta-learned visual prefix to leverage additional information about image, not necessarily present in the ground-truth. An important observation from these examples is that the generated answers have some differences w.r.t the ground-truth, which is however penalized during evaluation, since we count as correct only words that match, to be able to compare to Frozen. The evaluation should, however, measure whether the generated sentence matches the image in a more general manner, ideally without using ground-truth references (Hessel et al., 2021), which is left for future work.

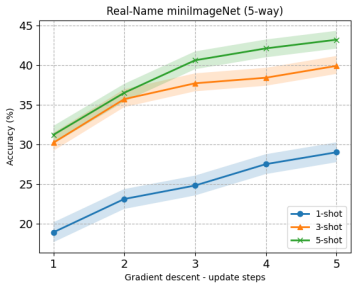


Figure 3: Relationship between consecutive steps of gradient-based updates and the accuracy on Real-Name miniImageNet on 5-way tasks, performed during the meta-test stage.

4.3 Ablation Study

What is the benefit of accruing meta-knowledge? To test how useful the meta-knowledge is, we run an ablation experiment where the meta-knowledge accrued from the previous batch of tasks is erased before observing the next one. This essentially means to randomly initialize the meta-mapper, which collects the meta-knowledge, before observing the next batch of tasks. As it can be observed

Table 5: Benefit of the self-attention structure of the meta-mapper in accuracy (%) on the four multimodal few-shot datasets under the 2-way setting. Self-attention meta-mapper is more critical than the MLP-based one for our approach.

	Real-Name		**Open-Ended**		**Real-Fast VQA**		**Fast-VQA**	
Variants	1-shot	5-shots	1-shot	5-shot	1-shot	5-shots	1-shot	5-shot
MLP-based meta-mapper	42.5	62.6	45.3	58.3	4.3	11.9	4.5	7.5
Self-attention meta-mapper	**48.2**	**72.3**	**58.7**	**65.8**	**9.7**	**13.2**	**5.7**	**9.3**

Table 6: Effect of using a fixed, hand-engineered task induction vs. only learning it from support sets, in accuracy (%) on the four multimodal few-shot datasets under the 2-way setting. Learning the task induction in a data-driven manner improves the performance.

	Real-Name		**Open-Ended**		**Real-Fast VQA**		**Fast-VQA**	
Variants	1-shot	5-shots	1-shot	5-shot	1-shot	5-shots	1-shot	5-shot
w/ fixed task ind	47.9	71.9	58.0	65.0	9.5	**13.2**	5.0	8.9
w/o fixed task ind	**48.2**	**72.3**	**58.7**	**65.8**	**9.7**	**13.2**	**5.7**	**9.3**

in Table 4, this erasing yields deteriorating performance, across all datasets, ways and shots. This shows that training the meta-mapper to accumulate useful representations from sequences of tasks is crucial for the success of our meta-approach.

What is the effect of the multimodal inner loop optimization? To analyse the effect of the gradient-based updates during test time, namely the inner loop optimization, we illustrate the relationship between the update steps and the accuracy in Figure 3. We observe increased performance as the number of gradient-based steps increases, for $\{1, 3, 5\}$ shots on 5-way Real-Name miniImageNet tasks. This demonstrates that the mechanism of the meta-learning framework to adapt the meta-parameters on support samples, allows the model to respond promptly to a new task, which yields increased performance on the query samples.

Does the structure of the meta-mapper influence the performance? To test the performance of our approach with a different structure of the meta-mapper, we run experiments with an alternative to the proposed self-attention-based one. Specifically, we implement a multi-layer perceptron (MLP) with one hidden layer. It can be observed from Table 5, that the self-attention meta-mapper brings notable improvements over MLP. This is attributed to the fact that, different from MLP, self-attention layers are able to select and extract the most relevant features from an input representation.

Is there an effect of fixed task induction in the meta-learning setting? We analyse the effect of a fixed hand-engineered task induction, as opposed to learning the induction in a data-driven manner. In particular, we choose the best variant of our meta-learner from Table 1, and add a fixed sentence describing the task in natural language, for instance, *"Answer with lion or dog."*, similar as in Frozen. It can be observed from Table 6, that there is no significant increase in the performance, meaning that only fine-tuning the meta-mapper on the support set is good enough to induce the task.

5 Conclusion

We present a novel meta-learning approach for multimodal few-shot learning. We design our model by introducing a lightweight trainable meta-mapper that acts as a bridge between large frozen vision and language models. This meta-mapper accrues shared meta-knowledge by sequentially observing tasks into a learnable visual prefix, which is then used to steer the language model into generating relevant linguistic interpretations. Unlike prior works, we use the support samples to adapt our model to unseen tasks, which shows to be beneficial as it improves performance without using hand-engineered task induction. Moreover, it operates with low computational costs, as it smartly combines large pre-trained models without the need to train them from scratch. This makes our architecture flexible enough to incorporate additional modalities, which is left to tackle in future work. Last but not least, our experiments verify the effectiveness of our approach by outperforming the baseline on several benchmarks and fostering further exploration of multimodal few-shot meta-learning.

ETHICS STATEMENT

Our proposed model uses pre-trained large-scale models, both for the vision encoder and language decoder modules, namely CLIP (Radford et al., 2021) and GPT-2 (Brown et al., 2020). So, our model inherits any problematic biases and limitations that these pre-trained models may contain, meaning we should be careful when using the results especially in sensitive settings.
More specifically, our model has a generative nature with the ability to produce language, which might lead to generation of disinformation or potential hallucinations. But, as our model is conditional on the visual features, it enforces the text generation to follow the image contents. This serves as a control mechanism of the otherwise free text generation, and can mitigate the generation of linguistic disinformation. Additionally, as can be seen in some of the qualitative examples in Figure 4, our proposed approach takes advantage of the complementarities of the two backbone models. This suggests that it might be useful to mitigate biases potentially present in one of the pre-trained models.

REPRODUCIBILITY STATEMENT

To help the reproducibility of the paper, we provide the formal definition of the meta-learning setup in Section 3.1, the implementation details in Section 4.1, additional details about the design of the setting and the formal definition of the algorithms during the meta-training and inferences stages in Appendix C. Also, we release the source code to reproduce the results and evaluate the performance of the model at: https://github.com/<redacted>/<redacted>.

REFERENCES

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.

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.

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

Y. Bengio, S. Bengio, and J. Cloutier. Learning a synaptic learning rule. In *IJCNN-91-Seattle International Joint Conference on Neural Networks*, volume ii, pp. 969 vol.2–, 1991. doi: 10.1109/IJCNN.1991.155621.

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.

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.

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.

Constantin Eichenberg, Sidney Black, Samuel Weinbach, Letitia Parcalabescu, and Anette Frank. Magma–multimodal augmentation of generative models through adapter-based finetuning. *arXiv preprint arXiv:2112.05253*, 2021.

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

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

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

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*, pp. 770–778, 2016.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 2021. doi: 10.18653/v1/2021.emnlp-main.595. URL http://dx.doi.org/10.18653/v1/2021.emnlp-main.595.

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

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*, pp. 4904–4916. PMLR, 2021.

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.

Woojeong Jin, Yu Cheng, Yelong Shen, Weizhu Chen, and Xiang Ren. A good prompt is worth millions of parameters: Low-resource prompt-based learning for vision-language models, 2021.

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*, pp. 3744–3753. PMLR, 2019.

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. doi: 10.18653/v1/2021.emnlp-main.243. URL http://dx.doi.org/10.18653/v1/2021.emnlp-main.243.

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

Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *European Conference on Computer Vision*, pp. 121–137. Springer, 2020.

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*, pp. 740–755. Springer, 2014.

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

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, 32, 2019.

Sewon Min, Mike Lewis, Luke Zettlemoyer, and Hannaneh Hajishirzi. Metaicl: Learning to learn in context. *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2022.

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

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

Tsendsuren Munkhdalai and Hong Yu. Meta networks. In *ICML*, pp. 2554–2563, 2017.

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

Ethan Perez, Douwe Kiela, and Kyunghyun Cho. True few-shot learning with language models. *Advances in Neural Information Processing Systems*, 34:11054–11070, 2021.

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

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*, pp. 8748–8763. PMLR, 2021.

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.

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

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

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

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

Sheng Shen, Liunian Harold Li, Hao Tan, Mohit Bansal, Anna Rohrbach, Kai-Wei Chang, Zhewei Yao, and Kurt Keutzer. How much can clip benefit vision-and-language tasks? In *International Conference on Learning Representations*, 2021.

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

Haoyu Song, Li Dong, Wei-Nan Zhang, Ting Liu, and Furu Wei. Clip models are few-shot learners: Empirical studies on vqa and visual entailment, 2022.

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*, pp. 1199–1208, 2018.

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.

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

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.

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.

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

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.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 38–45, Online, October 2020. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/2020.emnlp-demos.6.

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*, pp. 8808–8817, 2020.

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.

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

APPENDIX

The appendix consists of the following sections: A. Vision encoder details, B. Language model details, C. Multimodal meta-learning details, D. Additional results, E. Additional visualizations of the effect of gradient-based updates.

A VISION ENCODER DETAILS

For the pre-trained vision encoder we adopt CLIP (Radford et al., 2021), 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 (He et al., 2016) or a Vision Transformer (ViT) (Dosovitskiy et al., 2021) and 2) a text encoder implemented as a Transformer (Radford et al., 2019). The pre-training is done in a contrastive manner, on a large dataset of 400 million pairs of image-caption, with the aim is to minimize the distance of same images and captions pairs in the embedding space and to maximize the distance between different image and captions.

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.

B LANGUAGE MODEL DETAILS

The pre-trained language model that we employ is GPT-2 (Radford et al., 2019), particularly the small version with 117M parameters. Its architecture is following a Transformer decoder (Vaswani et al., 2017) with 12 layers and word embedding dimensions of 768. The model is pre-trained on a very large corpus of English data in a self-supervised fashion with 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 also 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 technique as in Holtzman et al. (2019). To build a more efficient architecture, the language model is kept entirely frozen, same as the vision stream.

C 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 (Ravi & Larochelle, 2017). In particular, we use either COCO2017 captioning dataset (Lin et al., 2014) to obtain cross-domain experimental setup, or the multimodal few-shot datasets Tsimpoukelli et al. (2021) for the standard meta-learning setup. The partitioning into meta-training and meta-test tasks is illustrated in Figure 5, using the Real-Name miniImageNet dataset as an example. 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 to those objects. Note that the samples in the query set during meta-training should be at least 15 per category, following (Finn et al., 2017), since the optimization of the meta-parameters is done based on those samples.

The detailed optimization process in the meta-training stage is described in Algorithm 1. Similarly, the inference stage using the meta-test partitions is described in Algorithm 2. In particular, the meta-training and inference stages in our approach are less computationally-expensive compared to the large-scale pre-training of the vision encoder of Frozen (Tsimpoukelli et al., 2021), while achieving comparable performance. To be more specific, Frozen pre-trains the vision encoder from

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 θ 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

scratch and then uses it to extract image features for prompting a frozen language model. By contrast, we freeze the backbones entirely and only train the meta-mapper. This results in a more flexible architecture which is also less computationally intensive and independent of the specific pre-training of large-scale models. However, during inference time, we are fine-tuning the meta-mapper, by using the support set to adapt its parameters to the given task. This is different from Frozen, which performs direct prompting of the language model with no gradient-step updates. Although we perform a few gradient-step updates at inference time - due to the nature of the meta-learning algorithm - this adds a minimal complexity compared to training a whole vision encoder from scratch.

D ADDITIONAL RESULTS

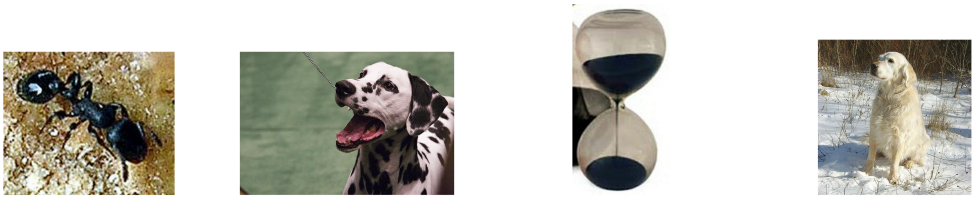
In this section, we provide more results, both quantitative and qualitative, and we discuss few additional observations. Regarding the quantitative results, we provide the results for the complete settings, with all shots: {1, 3, 5}, in Table 7 and 8. These experiments demonstrate that our model shows higher performance for the in-domain setup, compared to the cross-domain one, as commonly observed when the training and test partitions come from the same distribution. Frozen is not considering this in-domain setting, nor has released the pre-trained model or code, thus we could not include such results in our tables. We consider the in-domain setting as a relevant one in *any* few-shot settings, thus we incorporate it in our experimental design. As for the additional qualitative results, Figure 6 shows few successful cases when our multimodal few-shot learner is trained and tested on Real-Name miniImageNet (Tsimpoukelli et al., 2021). On the other hand, Figure 7 shows few cases when the correct answers were either missed or slightly different compared to the ground-truth ones. It is interesting to see that even though the model is not producing a correct answer which matches to the ground-truth, it is still able to grasp the visual concepts in the image and to map them with the meta-mapper into the visual prefix. For instance, the golden retriever is described as a *large* retriever and the dalmatian is described as a king *dog*.

Figure 5: Example of the new multimodal few-shot meta-learning setting, illustrating the 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 meta-test step is defined in a similar way, with the major difference that it contains new categories of objects that are never seen at meta-training time.

Next, Figure 8 shows few successful cases when our multimodal few-shot learner is trained and tested on Real-Fast VQA. As it can be noticed, this dataset contains more complicated questions, often asking about some attribute or relation between the objects in the images. Here, we again observe that the model can describe some extra visual information, for instance, in the case with the elephant (additionally described as *walking*). An additional observation is that the model tends to generate the answer starting with "This is a", since all images in the support set, used for adapting, have a caption starting with "This is a ... ". In addition, Figure 9 shows examples of few cases where the generated answers are not exactly matching the ground-truth. For instance, the ground-truth answer for the first image is a *teapot*, whereas our model generates *a wine bottle* as an answer. Although different from the ground-truth answer, it can be noticed that there is also a wine bottle behind the vase. This means that some visual information was correctly learned in the visual prefix and used to steer the language model into generating the corresponding words.

E Additional visualizations of the effect of gradient-based updates

In this section, we provide additional visualizations of the relationship between the performance in accuracy and the gradient-based update steps in the multimodal inner loop, similar to Figure 3. Particularly, we use the Real-Fast VQA and Fast-VQA datasets, for the 2-way settings, to visualize this relationship. Similar to Figure 3, we observe increased performance as the number of gradient-descent steps increases, suggesting the benefit of the meta-learned mapper and its fine-tuning during inference to adapt to a new unseen task. Additionally, we observe a slightly larger variance in the



Q: This is a?
Ground-truth: This is a ant.
Ours: This is a ant.

Q: This is a?
Ground-truth: This is a dalmatian.
Ours: This is a dalmatian barking.

Q: This is a?
Ground-truth: This is a hourglass.
Ours: This is a hourglass.

Q: This is a?
Ground-truth: This is a golden retriever.
Ours: This is a golden retriever.

Figure 6: Qualitative results from our multimodal few-shot learner, evaluated on Real-Name miniImageNet (Tsimpoukelli et al., 2021), with 5-way, 5-shot tasks, showing few successful cases.

Q: This is a?
Ground-truth: This is a vase.
Ours: This is a birthday cake vase.

Q: This is a?
Ground-truth: This is an hourglass.
Ours: This is a diamond hourglass.

Q: This is a?
Ground-truth: This is a golden retriever.
Our: This is a large retriever.

Q: This is a?
Ground-truth: This is a dalmatian.
Ours: This is a king dog.

Figure 7: Qualitative results from our multimodal few-shot learner, evaluated on Real-Name miniImageNet (Tsimpoukelli et al., 2021), with 5 way, 5-shot tasks, showing cases when the correct answers were either missed or slightly different compared to the ground-truth.

Q: What is the bird next to?
Ground-truth: A fence.
Ours: This is a fence.

Q: What animal is gray in color?
Ground-truth: A elephant.
Ours: This is a elephant walking.

Q: What does the dog appear to have in mouth?
Ground-truth: Boy's shoe.
Ours: This is a shoe.

Q: What is the clock attached to?
Ground-truth: A tower.
Ours: A tower.

Figure 8: Qualitative results from our multimodal few-shot learner, evaluated on Real-Fast VQA (Tsimpoukelli et al., 2021), with 5-way, 2-shot tasks, showing successful cases.

Q: What kind of pot is by the vase?
Ground-truth: Teapot.
Ours: wine bottle.

Q: What is the glass dish sitting on?
Ground-truth: A table.
Ours: A dishwasher.

Q: What is the tall building in the background?
Ground-truth: A skyscraper.
Ours: This is a building image.

Q: What is the pizza in?
Ground-truth: A cardboard.
Ours: A box.

Figure 9: Qualitative results from our multimodal few-shot learner, evaluated on Real-Fast VQA (Tsimpoukelli et al., 2021), with 5-way, 2-shot tasks, showing cases when the correct answers were either missed or slightly different compared to the ground-truth.

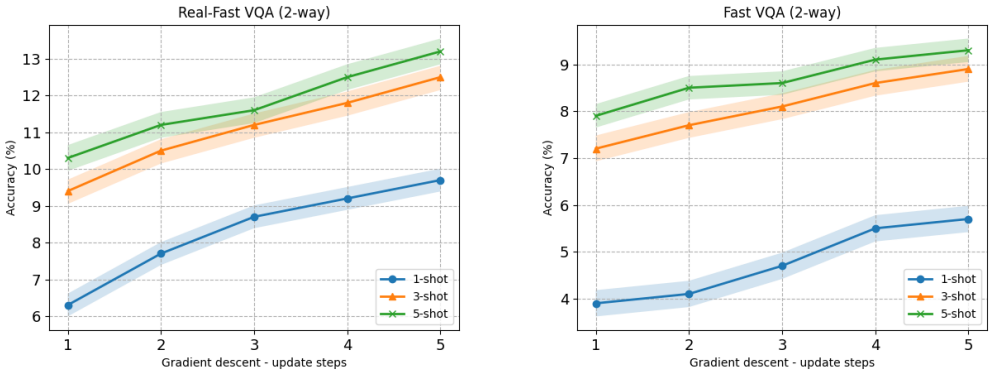
Table 7: Comparison with the Frozen (Tsimpoukelli et al., 2021) baselines on Real-Name and Open-Ended miniImageNet 2- and 5-way setting, including all shots available: {1, 3, 5}. ANIL (Raghu et al., 2019) is used as an upper bound, since it is a discriminative approach as opposed to our open-ended generative one. Our episodically trained models are outperforming the Frozen baselines, both with and without domain-shift. The overall best performance is denoted in bold, whereas the same settings as the baseline are denoted in italic & bold.

			Real-Name 2-way			**Open-Ended 2-way**			**Real-Name 5-way**			**Open-Ended 5-way**		
Methods	episodic	cross-domain	1-shot	3-shots	5-shots	1-shot	3-shots	5-shots	1-shot	3-shots	5-shots	1-shot	3-shots	5-shots
Frozen w/o task ind	✗	✓	1.7	-	-	29.0	-	-	0.9	-	-	18.0	-	-
Frozen w/ task ind	✗	✓	33.7	66	66	53.4	57.9	58.9	14.5	34.7	33.8	20.2	22.3	21.3
Ours	✗	✗	35.6	65.3	65.7	50.2	54.6	57.5	15.2	35.2	39.6	18.9	20.6	22
	✗	✓	***37.3***	*65.2*	***66***	*52.5*	*55.6*	***59***	***19.2***	***37.5***	***40.3***	***20.9***	***22.6***	***25.0***
	✓	✓	45.3	68.3	69.8	56.7	60	63.4	24.7	37.9	41.8	24.8	26.9	28.0
	✓	✗	**48.2**	**70.7**	**72.3**	**58.7**	**62.2**	**65.8**	**29.0**	**39.9**	**43.2**	**25.1**	**27.6**	**29.6**
ANIL upper-bound	-	-	73.9	81.7	84.2	-	-	-	45.5	57.7	62.6	-	-	-

Table 8: Comparison with the Frozen baseline (Tsimpoukelli et al., 2021) on 2-way Real-Fast VQA and Fast-VQA, including all shots available: {1, 3, 5}. Our episodically trained models outperform their counterparts, both with and without domain shift. The overall best performance is denoted in bold, whereas the same settings as the baseline are denoted in italic & bold.

			Real-Fast VQA 2-way			**Fast-VQA 2-way**		
Methods	episodic	cross-domain	1-shot	3-shots	5-shots	1-shot	3-shots	5-shots
Frozen	✗	✓	7.8	10.1	10.5	2.8	7.0	7.9
Ours	✗	✗	5.4	8.4	9.1	2.5	6.4	7.1
	✗	✓	*6.9*	*10.0*	***10.7***	***3***	***7.1***	***8***
	✓	✓	8.5	11.2	13	5.2	7.5	8.6
	✓	✗	**9.7**	**12.5**	**13.2**	**5.7**	**8.9**	**9.3**

accuracy, which is hinting at something also relevant for any unimodal few-shot scenarios: namely, the magnitude of the variance in few-shot learning benchmarks largely depends on the total number of few-shot tasks in the meta-test stage, rather than the number of support set samples per task. This means that if the few-shot datasets have more tasks to average over, naturally the variance in the reported accuracy will be smaller. For instance, in few-shot image classification datasets (Vinyals et al., 2016), the number of tasks is a lot larger compared to the multimodal few-shot datasets (Tsimpoukelli et al., 2021), thus the few-shot models trained on them exhibit smaller variance in their performance.



(a) Real-Fast VQA (b) Fast-VQA

Figure 10: Relationship between 5 consecutive steps of gradient-based updates and the accuracy on Real-Fast VQA and Fast-VQA on 2-way tasks, performed during the meta-test stage.