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

Ivona Najdenkoska[1], Xiantong Zhen[1,2], Marcel Worring[1]
[1]AIM Lab, University of Amsterdam, [2]Inception Institute of Artificial Intelligence
{i.najdenkoska, x.zhen, m.worring} @uva.nl

APPENDIX

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

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 1, 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

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

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 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

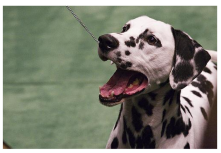
In this section, we provide more results, both quantitative and qualitative, and we discuss few additional observations. Regarding the quantitative results, we provide the results for the complete settings, with all shots: {1, 3, 5}, in Table 1 and 2. These experiments demonstrate that our model shows 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

Figure 1: 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.

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 2 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 3 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*.

Next, Figure 4 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 5 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.


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.

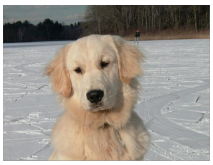

Q: This is a?
Ground-truth: This is a hourglass.
Ours: This is a hourglass.


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

Figure 2: 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.

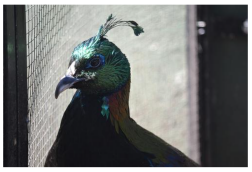

Q: This is a?
Ground-truth: This is a vase.
Ours: This is a birthday cake vase.


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 3: 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.

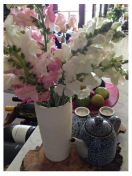

Q: What is the bird next to?
Ground-truth: A fence.
Ours: This is a fence.


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 4: 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 5: 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 1: 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 2: 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**

REFERENCES

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.

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.

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.

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

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.

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.

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.