I-TUNING: TUNING FROZEN LANGUAGE MODELS WITH IMAGE FOR LIGHTWEIGHT IMAGE CAPTIONING

Ziyang Luo[⋆]   *Zhipeng Hu*[†‡]   *Yadong Xi*[‡]   *Rongsheng Zhang*[‡]   *Jing Ma*[⋆✉]

[⋆] Department of Computer Science, Hong Kong Baptist University, Hong Kong SAR, China
[†] College of Computer Science and Technology, Zhejiang University, Hangzhou, China
[‡] Fuxi AI Lab, NetEase Inc., Hangzhou, China

ABSTRACT

Image Captioning is a traditional vision-and-language task that aims to generate the language description of an image. Recent studies focus on scaling up the model size and the number of training data, which significantly increase the cost of model training. Different to these heavy-cost models, we introduce a lightweight image captioning framework (**I-Tuning**), which contains a small number of trainable parameters. We design a novel I-Tuning cross-attention module to connect the non-trainable pre-trained language decoder GPT2 and vision encoder CLIP-ViT. Since most parameters are not required to be updated during training, our framework is lightweight and fast. Experimental results conducted on three image captioning benchmarks reveal that our framework achieves comparable or better performance than the large-scale baseline systems. But our models contain up to **10 times fewer** trainable parameters and require much fewer data for training compared with state-of-the-art baselines.

Index Terms— Lightweight image captioning, Language models, Transformer, Cross-Modal

1. INTRODUCTION

Image Captioning is a critical task in the field of cross-modal, which focus on natural language generation to depict an image. Recent years have witnessed the success of applying large-scale pre-trained models on the task of image captioning, which generally scale up the number of trainable parameters and training data to achieve state-of-the-art performances [1, 2, 3, 4]. For example, a recent proposed OSCAR model [1] contains more than 135M trainable parameters and requires around 4M images during pre-training. Therefore, in spite of the performances, the heavy demands for extra computational resources and massive data for model training have become an urgent issue.

Recent studies showed that parameter-efficient pre-trained language models (PLMs) tuning [5] can effectively reduce the cost during training, where most parameters are frozen (i.e.,
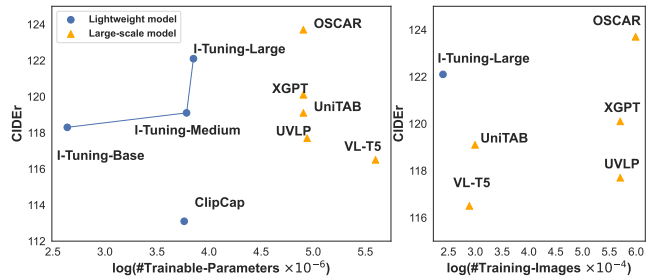

✉: Corresponding author (majing@hkbu.edu.hk)

Fig. 1. Image captioning performances on MSCOCO dataset based on the lightweight and large-scale models, where I-Tuning-Base/-Medium/-Large respectively denote our model with base/medium/large number of trainable parameters.

not updated during training) and the rest small set are trainable. [6] recently introduced a ClipCap model that transforms images into fixed-length vectors and prompts a frozen GPT2 for image captioning. However, the learned vectors cannot capture accurate visual information to enhance the caption generation. To overcome these shortcomings, we propose a novel lightweight image captioning framework (**I-Tuning**) to alleviate the cost in terms of computational resource and training data. We design an I-Tuning module to connect the pre-trained vision encoder (i.e., CLIP-ViT [7]) and the language decoder (i.e., GPT2 [8]). To align between the language and vision modals, it serves as a cross-modal filter that automatically picks the visual information from the output of the vision encoder and adjusts the output hidden states of the language decoder. During training, we only update the newly introduced parameters in the I-Tuning module, and the parameters of the two pre-trained models are frozen.

Figure 1 examplifies the CIDEr scores of our lightweight models and large-scale baselines. In terms of model training, our basic model I-Tuning-Base only contains around 14M trainable parameters, namely **10 times fewer** than the other large-scale models such as OSCAR. In terms of data, even our I-Tuning-Large model can achieve comparable performances with relatively less training data. We evaluate our proposed framework on 3 image captioning benchmarks (i.e.,

MSCOCO [9], Flickr30k [10] and NoCaps [11]). The results show that our **I-Tuning** framework achieves comparable or even better performances than large-scale baselines with up to **10 times fewer** trainable parameters and much fewer cross-modal training data. Moreover, our I-Tuning model is agnostic to the pre-trained language models, suggesting a broadly applicable framework.

2. RELATED WORK

CLIP-ViT and GPT2. CLIP-ViT [7] is the state-of-the-art vision encoder. It is pre-trained with contrastive loss [12] to supervise the vision encoder with language description. GPT2 [8] is the state-of-the-art language decoder, which is pre-trained with large-scale text data. In this work, we propose a lightweight image captioning framework **I-Tuning** to leverage these two off-the-shelf pre-trained models.

Image Captioning. Generating the language descriptions from images is an important task to examine the vision-and-language representation ability of a cross-modal model. The recent works choose to increase the model size and the number of training data to further boost the performance [1, 2, 3, 4, 13]. The training process of these models is heavy. As an alternative, the ClipCap model [6] proposes a lightweight captioning model by connecting the off-the-shelf CLIP-ViT and GPT2. However, their method cannot filter the relevant visual information to adjust the output hidden states of GPT2, leading to poor image captioning performance.

Parameter-efficient PLMs Tuning. Recently, the model size of a pre-trained model becomes larger and larger, which makes us hard to fully fine-tune such models. To make use of them without updating all parameters, researchers propose several great ideas, such as Prefix tuning [14], Adapter tuning [5] and Prompt tuning [15]. However, most of them only focus on the NLP area. Our **I-Tuning** extends the parameter-efficient PLMs tuning idea to the cross-modal setting.

3. THE PROPOSED I-TUNING FRAMEWORK

Overview. Our framework contains three components, the non-trainable vision encoder (CLIP-ViT), the non-trainable language decoder (GPT2), and the trainable I-Tuning Module. During training, our framework is trained with the parallel image-caption data and only updated the parameters of the lightweight I-Tuning Module.

During inference, a frozen visual encoder first generates the visual embeddings V of a given image. Then the **I-Tuning** module serves as a lightweight filter to pick the relevant visual information to tune the output hidden states of the frozen language model. As a result, the language generation is conditioned with the given image.

Visual Encoder and Language Decoder. In our framework, we adopt the state-of-the-art vision pre-trained transformer, CLIP-ViT [7] to generate an image's visual embed-
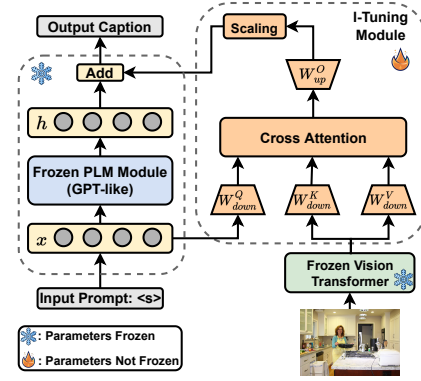

Fig. 2. **An overview of our I-Tuning framework for Lightweight Image Captioning.**

dings V. Such model takes a sequence of image patches as input and visual representations for each patch as output. For the Language Decoder, we leverage the state-of-the-art auto-regressive pre-trained language model (PLM), GPT2 [8], which is a multi-layer Transformer Decoder model [16] with remarkable language generation ability.

I-Tuning Module. In our framework, the **I-Tuning** module is the key component to extract the relevant visual information from the visual embeddings, which is parallel to a specific PLM module (feedforward) in each Transformer layer. Such module is a bottleneck neural network, sharing a similar structure as the Adapter module [5], but the non-linear activation function is replaced by a cross-attention network (see Figure 2) to filter the visual information from images. The calculation process is as follows:

$$Q_L = W_{down}^Q(X) + b^Q, \qquad (1)$$

$$K_V = W_{down}^K(V) + b^K, \qquad (2)$$

$$V_V = W_{down}^V(V) + b^V, \qquad (3)$$

where X is the input hidden states of a specific PLM module. Then we can get the attention scores across the visual embeddings:

$$S = softmax(Q_L K_V^T). \qquad (4)$$

Based on the scores, we can get the final **I-Tuning** output to adjust the output hidden states of the PLM module:

$$\Delta h = \lambda W_{up}^O \left(\sum_i s_i V_{Vi} \right) + b^O, \qquad (5)$$

where $\lambda \geq 1$ is a scaling hyper-parameter.

Since the lower layers of PLMs have weaker representation ability, we also propose **I-Tuning Dropping** to remove the **I-Tuning** modules in the first-few layers. As a result, backpropagating through fewer layers can further improve the training efficiency of our models.

Training Objective. The objective is the auto-regressive language modeling conditioned on the visual information:

Model	#Images	#Params	MSCOCO (test)				Flickr (test)			
			CIDEr	BLUE@4	METER	SPICE	CIDEr	BLUE@4	METER	SPICE
Large-scale Cross-Modal Pre-trained Image Captioning Model										
OSCAR$_{base}$(no tags) [1]	4M	135M	115.6	34.5	29.1	21.9	-	-	-	-
OSCAR$_{base}$♠ [1]	4M	135M	123.7	36.5	30.3	23.1	-	-	-	-
Unified VLP [2]	3M	135M	117.7	36.5	28.4	21.3	67.4	30.1	23.0	17.0
XGPT [3]	3M	135M	120.1	37.2	28.6	21.8	70.9	31.8	23.6	17.6
UniTAB [4]	200k	135M	119.1	35.8	28.4	21.5	70.1	30.7	23.7	17.4
VL-T5 [17]	180k	270M	116.5	34.5	28.7	21.9	-	-	-	-
Lightweight Image Captioning Model										
ClipCap(GPT2-Large) [6]	0	43M	113.1	33.5	27.5	21.1	-	-	-	-
Our Lightweight Models w/o VLP										
I-Tuning(GPT2-Base)	0	14M	116.7	34.8	28.3	21.8	61.5	25.2	22.8	16.9
I-Tuning(GPT2-Medium)	0	44M	120.0	35.5	28.8	22.0	72.3	28.8	24.6	19.0
I-Tuning(GPT2-Large)	0	95M	119.4	34.8	29.3	22.4	75.4	29.8	25.1	19.2
Our Lightweight Models w/ VLP										
I-Tuning(GPT2-Base)	110k	14M	118.3	35.2	28.5	22.0	68.4	27.5	24.0	18.4
I-Tuning(GPT2-Medium)	110k	44M	119.1	34.8	29.2	22.2	73.2	29.1	25.2	19.9
I-Tuning(GPT2-Large)	110k	95M	**122.2**	35.9	**29.5**	**22.6**	77.2	30.0	**25.5**	**20.2**
Our Lightweight Models w/ I-Tuning Dropping										
I-Tuning(GPT2-Large)	110k	47M	122.1	**36.1**	29.4	**22.6**	**79.2**	**31.1**	25.3	19.9

Table 1. Evaluations on MSCOCO and Flickr Image Captioning. "-" represents that the model does not report such result in its original paper. **Bold** indicates the best scores of our models. #Images represents the number of distinct images during VLP. #Params represents the number of trainable parameters. ♠: Extra training data are needed to generate the object tags.)

$\mathcal{L} = -\Sigma_{t=1}^{T} \log P(x_t | x_{<t}, V)$, where V represents the visual embeddings encoded by the frozen visual encoder, T denotes the length of a sequence and $x_{<t} = (x_0, ..., x_{t-1})$.

4. EXPERIMENT

4.1. Dataset and Setup

We adopts CLIP-ViT B/16 as our visual encoder and GPT2 as language decoder. All of them are frozen during training. We include 3 different GPT2 model sizes, including Base, Medium and Large. For **I-Tuning** modules, the parameters are randomly initialized and updated during training. For VLP, we adopt the cross-modal dataset, Visual Genome [18], which contains 110k distinct images. To evaluate our methods, we use three datasets, namely **MSCOCO** [9], **Flickr30k** [10] and **NoCaps** [11]. For the first two datasets, we follow the Karpathy's split [19] to split 113.2k/5k/5k and 29.8k/1k/1k images for train/val/test, respectively. We adopt CIDEr [20], BLEU@4 [21], METEOR [22] and SPICE [23] as metrics to evaluate the generated captions. We train our models with the AdamW [24] and 4k batch size. For VLP, our models are pre-trained with 10 epochs. For training on downstream tasks, our models are trained with 30 epochs. For inference, we use the beam search (beam size = 5) to generate captions.

4.2. Result Analysis

Table 1–2 reveal that our lightweight image captioning framework achieves comparable or better performance than all the large-scale baselines, but contains up to 10 times fewer trainable parameters and/or consume much fewer VLP data.

I-Tuning without VLP. As shown in Table 2, our method outperform the large-scale baselines even without VLP. Especially, the overall CIDEr score of the OSCAR model on the NoCaps even lags behind the frozen GPT2-base with our **I-Tuning** modules by around 2 points, while our model contains around 120M fewer trainable parameters. With the larger GPT2, the performance gap becomes larger. Moreover, our method is also sample efficient. Without any cross-modal pre-training, our **I-Tuning** (GPT2-Medium) already outperforms some baselines with VLP. For example, VL-T5 is pre-trained with 180k distinct cross-modal images, but the CIDEr scores are around 3.4 lower than ours on MSCOCO.

I-Tuning with VLP. Table 1 reveals that after cross-modal pre-training, our **I-Tuning** method achieves better overall performance than all the baseline systems (except OS-CAR w/ object tags). Especially, our **I-Tuning** can achieve a CIDEr score of 122.1 on MSCOCO test set, surpassing the XGPT model by 2.0 points, while our method requires less trainable parameters and training data. For the OSCAR model, it requires object tags during pre-training and fine-tuning. Additional supervision is needed to generate these tags. One can find that our lightweight **I-Tuning** framework still reaches comparable performance with only 1.5 CIDEr score lower, while our model requires around 90M less trainable parameters and 30 times less distinct VLP images. Without the help of object tags, OSCAR even lags behind our models without VLP.

I-Tuning with Dropping. Since the lower layers of GPT2 have weaker representation ability, we investigate whether we

Model	#Params	in-domain		near-domain		out-of-domain		Overall	
		CIDEr	SPICE	CIDEr	SPICE	CIDEr	SPICE	CIDEr	SPICE
OSCAR$_{base}$ [1]	135M	79.6	12.3	66.1	11.5	45.3	9.7	63.8	11.2
ClipCap(GPT2-Large) [6]	43M	84.9	12.1	66.8	10.9	49.1	9.6	65.8	10.9
Our Models									
I-Tuning(GPT2-Base)	14M	83.9	12.4	70.3	11.7	48.1	9.5	67.8	11.4
I-Tuning(GPT2-Medium)	44M	**89.6**	12.9	77.4	12.2	58.8	10.5	75.4	12.0
I-Tuning(GPT2-Large)	95M	**89.6**	**13.3**	80.4	**12.6**	64.8	**11.0**	78.5	**12.4**
Our Models w/ Dropping									
I-Tuning(GPT2-Large)	47M	88.3	12.7	**80.8**	**12.6**	**66.1**	10.8	**78.9**	12.3

Table 2. Evaluations on NoCaps image captioning. Models are only trained with MSCOCO training set without VLP. #Params represents the number of trainable parameters. **Bold** indicates the best scores.

Image				
Golden Captions	(1) A man with a red helmet on a small moped on a dirt road. (2) Man riding a motor bike on a dirt road on the countryside.	(1) A young girl inhales with the intent of blowing out a candle. (2) A young girl is preparing to blow out her candle.	(1) A man on a bicycle riding next to a train. (2) A person is riding a bicycle but there is a train in the background.	(1) A kitchen is shown with a variety of items on the counters. (2) A kitchen has the windows open and plaid curtains.
Model	**Generated Caption**			
ClipCap (GPT2-Large)	a man is riding a motorbike on a dirt road.	a young girl sitting at a table with a cup of cake.	a man is standing next to a train.	a kitchen with a sink and a window
OSCAR$_{base}$	a man riding a motorcycle down a dirt road.	a woman sitting at a table with a plate of food.	a woman riding a bike down a street next to a train.	a kitchen with a sink, dishwasher and a window
I-Tuning (GPT2-Large)	A man riding a motorcycle on a dirt road.	A little girl blowing out a candle on a birthday cake.	A man standing next to a train on a train track.	A kitchen sink sitting under a window next to a window.

Table 3. Examples of our **I-Tuning**, OSCAR$_{base}$ and ClipCap for the first 4 images in the MSCOCO test set. (Red = inaccurate)

can drop the **I-Tuning** modules in the first few layers, so that we can further reduce the computational overhead during training and inference. Table 1 shows that it is not necessary to include the **I-Tuning** models in all layers. Especially, dropping the **I-Tuning** modules in the first-18 layers can even improve the performance of our model on some evaluation metrics, while the number of trainable parameters is reduced by 50%, improving the efficiency of our models.

Qualitative Evaluation. Table 3 presents the image captioning examples of **I-Tuning**, OSCAR and ClipCap for the first 4 images in the MSCOCO test set. The generated captions of **I-Tuning** depict the image successfully, which can identify the movement of the people in the image. For example, our model can recognize that the little girl is blowing the candles, while ClipCap and OSCAR cannot.

4.3. Cross-Attention Visualization

We visualize the cross-attention maps of **I-Tuning** to examine whether it learns the cross-modal information alignment implicitly. We randomly choose an image in the MSCOCO dataset and present the cross-attention heatmaps in the final **I-Tuning** module of GPT2-Large. Figure 3 shows that our **I-Tuning** module can correctly attend to the corresponding image regions given different tokens. These examples reveal that our method can learn visual grounding implicitly.
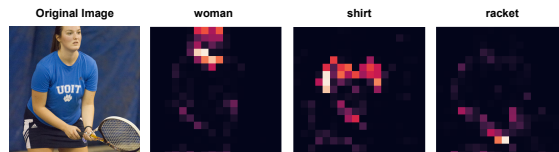
Fig. 3. Visualization of the cross-attention maps of text tokens in the caption "A woman in a blue shirt holding a racket".

5. CONCLUSION

In this paper, we present a novel lightweight image captioning framework, **I-Tuning**, which efficiently tunes the frozen PLMs with images. Extensive experiments are conducted to verify the effectiveness of our method. Compared with the baseline systems, our method achieves comparable or even better performance, while our models require up to **10 times fewer** trainable parameters and **much fewer** training data.

Acknowledgment

6. REFERENCES

[1] Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, Yejin Choi, and Jianfeng Gao, "Oscar: Object-semantics aligned pre-training for vision-language tasks," in *ECCV 2020*.

[2] Luowei Zhou, Hamid Palangi, Lei Zhang, Houdong Hu, Jason Corso, and Jianfeng Gao, "Unified vision-language pre-training for image captioning and vqa," *AAAI*, 2020.

[3] Qiaolin Xia, Haoyang Huang, Nan Duan, Dongdong Zhang, Lei Ji, Zhifang Sui, Edward Cui, Taroon Bharti, and Ming Zhou, "Xgpt: Cross-modal generative pre-training for image captioning," in *NLPCC*, 2021.

[4] Zhengyuan Yang, Zhe Gan, Jianfeng Wang, Xiaowei Hu, Faisal Ahmed, Zicheng Liu, Yumao Lu, and Lijuan Wang, "Unitab: Unifying text and box outputs for grounded vision-language modeling," in *ECCV*, 2022.

[5] Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin de Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly, "Parameter-efficient transfer learning for NLP," in *ICML*, 2019.

[6] Ron Mokady, Amir Hertz, and Amit H. Bermano, "Clipcap: Clip prefix for image captioning," 2021.

[7] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever, "Learning transferable visual models from natural language supervision," in *ICML*, 2021.

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

[9] Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár, "Microsoft coco: Common objects in context," 2015.

[10] Bryan A. Plummer, Liwei Wang, Chris M. Cervantes, Juan C. Caicedo, Julia Hockenmaier, and Svetlana Lazebnik, "Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models," in *ICCV*, 2015.

[11] Harsh Agrawal, Karan Desai, Yufei Wang, Xinlei Chen, Rishabh Jain, Mark Johnson, Dhruv Batra, Devi Parikh, Stefan Lee, and Peter Anderson, "Nocaps: novel object captioning at scale," *ICCV*, 2019.

[12] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross B. Girshick, "Momentum contrast for unsupervised visual representation learning," in *CVPR*, 2020.

[13] Zirui Wang, Jiahui Yu, Adams Wei Yu, Zihang Dai, Yulia Tsvetkov, and Yuan Cao, "SimVLM: Simple visual language model pretraining with weak supervision," in *ICLR*, 2022.

[14] Xiang Lisa Li and Percy Liang, "Prefix-tuning: Optimizing continuous prompts for generation," in *ACL*, 2021.

[15] Xiao Liu, Yanan Zheng, Zhengxiao Du, Ming Ding, Yujie Qian, Zhilin Yang, and Jie Tang, "Gpt understands, too," *arXiv:2103.10385*, 2021.

[16] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin, "Attention is all you need," in *Neurips*, 2017.

[17] Jaemin Cho, Jie Lei, Hao Tan, and Mohit Bansal, "Unifying vision-and-language tasks via text generation," in *ICML*, 2021.

[18] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Li Fei-Fei, "Visual genome: Connecting language and vision using crowdsourced dense image annotations," *IJCV*, vol. 123, pp. 32–73, 2016.

[19] Andrej Karpathy and Fei-Fei Li, "Deep visual-semantic alignments for generating image descriptions," in *CVPR*, 2015.

[20] Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh, "Cider: Consensus-based image description evaluation," in *CVPR*, 2015.

[21] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu, "Bleu: a method for automatic evaluation of machine translation," in *ACL*, 2002.

[22] Satanjeev Banerjee and Alon Lavie, "METEOR: An automatic metric for MT evaluation with improved correlation with human judgments," in *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, 2005.

[23] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould, "Spice: Semantic propositional image caption evaluation," 2016.

[24] Ilya Loshchilov and Frank Hutter, "Decoupled weight decay regularization," in *ICLR*, 2019.