Answer-Me: Multi-Task Open-Vocabulary Learning for Visual Question-Answering

AJ Piergiovanni, Wei Li, Weicheng Kuo, Mohammad Saffar, Fred Bertsch and Anelia Angelova
Google Research

Abstract

We present Answer-Me, a task-aware multi-task framework which unifies multiple question answering tasks, such as, visual question answering, visual entailment, visual reasoning. In contrast to previous works using contrastive or generative captioning training, we propose a novel and simple recipe to pretrain a vision-language joint model, which is multi-task as well, and uses the entire architecture end-to-end. Our results, which are in the challenging open-vocabulary generative setting, show state-of-the-art performance, zero-shot generalization, robustness to forgetting.

1. Multi-Task Learning for VQA

Visual Question Answering (VQA) is a challenging task as it involves deeper understanding of both visual and language inputs. For intelligent VQA systems it is desirable that they operate with natural questions and answers and are able to generalize to other tasks, not seen during training. Commonly used fine-tuned models [3] tend to exhibit larger rates of catastrophic forgetting on new tasks [6].

We propose 'Answer-Me' which unifies visual question answering tasks and aims to answer a variety of natural language questions towards an image (fig:motivation). The gist of the method is multi-task, task-aware training, which is able to respond according to the question's intent. This is combined with a novel pretraining which trains the entire encoder-decoder vision-language model simultaneously using only noisy data, and is also multi-task itself. This allows for natural language questions and free-form (open-vocabulary) outputs to answer accordingly, without additional prompts. Answer-Me generally outperforms multi-task SOTA methods, despite working in the challenging open-vocabulary setting, generalizing well to novel tasks.

Main architecture. Our model consists of an image encoder (ResNet) and text encoder and a Transformer fusion module. Our experiments are based on a ResNet-50 and T5-base model, and we scale it 3x by using ResNet-101 and T5-large. The image and language features are provided to a fusion module. The output of the fusion module is used as input to the text decoder, which produces free-form text for all Answer-Me tasks. While existing works have proposed similar fusion methods, ([3, 4]), the pretraining method and generalisation abilities without forgetting are new, using only raw images and no region proposals.

Figure 1. Answer-Me performance on unseen datasets (Zero-Shot), comparing a pretrained-only model (PT) and our multi-task learning on 4 and 8 tasks. While pretrained models are powerful they lack understanding of the question intent and are not able to respond to questions as adequately as our multi-task setup does.

1.1. Pretraining for multi-task learning

In order to enable the model to address new tasks, i.e., to respond to unseen question types and answer adequately, we take advantage of a unique pretraining designed to train all the components of a model. Unlike previous work, this pretraining strategy is targeted towards training the entire encoder-decoder model, it exercises various pathways in the model, which makes it suitable for various question-answering tasks. Another key advantage of this approach is that the training loss (cross entropy over the tokens) is simple and shared for all tasks. Specifically, for each sample, we have an (image, text) pair, obtained from the pretraining data. To train all parts of the model, we design a mix of four tasks: **(1) image captioning.** Here the input text is 'caption the image' and the target text is the caption. This task mostly trains the text decoder and fusion layers. **(2) caption completion.** Here the input is 10-40% of the caption text and the target text is the remaining caption. This encourages training of the entire model. **(3) text MLM [5].** Here the input is the caption with 25% of the words masked out, the target text is the missing words. This trains the en-

Approach	VQA(dev)	NLVR2	SNLI-VE	GQA	VizWiz
Specialized	70.2 [12]	53.5 [13]	71.6 [14]	57.5 [9]	57.2 [10]
Multi-task GPV [6]	62.5	-	-	-	-
VL-BART [4]	71.3	70.3	-	60.5	-
VL-T5 [4]	70.3	73.6	-	60.8	-
12-in-1 [11]	72.57	**78.44**	76.78	60.12	-
UniT (Coco init.) [8]	66.97	-	73.16	-	-
AnswerMe	65.1	71.7	77.5	72.8	72.4
AnswerMe, 3x	**73.6**	73.9	**85.8**	**77.5**	**75.3**

Table 1. Comparison of Answer-Me (8 tasks) to SOTA multi-task models.

tire model. **(4) image text matching [3].** Here the input is either the image caption or a random caption, the target text is 'true' or 'false' if the caption matches the image or not.

2. Experiments

Multi-task training. The multi-task training is done by taking a set of N tasks and mixing them together, sampled so that a batch consists of an equal amount of each dataset, i.e., batch size/N samples from each task. Since we use a text generation setting for the tasks, the loss is computed over the tokens, all using the same vocabulary. We use the T5 vocabulary with 32K tokens, for all experiments.

Datasets. We use the following datasets to address a number of VQA tasks: VQA2.0 [1], Visual Entailment (SNLI-VE) [14], Natural Language for Visual Reasoning (NLVR2) [13], GQA [9], and VizWiz [7] which is a VQA dataset collected by visually impaired users. The CC12m [2] is used for pretraining only. **Evaluation.** We follow the evaluation protocols established in prior work, and use standard adopted metrics to measure performance. However, instead of training a classification output layer, we use a large, open vocabulary and generate text answers.

Experimental results. In tab:sota we compare to the state-of-the-art (SOTA) multi-task methods, such as UniT [8], 12-in-1 [11], GPV [6]. Our model generally outperforms others, despite using open-vocabulary.

We then test the capabilities of the Answer-Me models and their potential for skill transfer. I.e., we compare how a model performs when a task is included in the training mix vs. a task outside the mix. tab:main compares Answer-Me trained on single tasks vs. different task mixtures in both standard and Zero-Shot (ZS) evaluation. We observe that the mixtures provide competitive results, outperforming or on par with single fine-tuned (FT) models, while using a single model. As seen, more tasks in the mixture improve the performance across all datasets, so does scaling. Importantly, the same conclusions are observed for ZS, where multi-task is able to improve performance and reduce the gap to supervised training (we make sure there is no 'leakage' to the test set for each experiment). We note how VizWiz has very low ZS results, as it is very challenging.

Answer-Me prevents catastrophic forgetting While pretraining and fine-tuning, as is customarily done in previous works, produces accurate models, it tends to overfit to the new data and immediately forget other datasets/tasks, even when it was previously trained on them. We show that Answer-Me, through the mixture training, is more robust, as it is able to sustain good performance across tasks (tab:oos).

Approach	Model	VQA2.0	NLVR2	SNLI-VE	GQA	VizWiz
Answer-Me, PT, ZS	Single	25.3	32.5	22.7	40.9	2.3
Answer-Me 4 tasks, ZS	Single	30.0	42.5	34.1	42.3	9.7
Answer-Me 8 tasks, ZS	Single	**35.0**	**44.7**	**37.3**	**44.2**	**10.3**
Answer-Me 8 tasks, ZS, 3x	Single	**39.2**	**48.3**	**41.1**	**47.2**	**11.4**
Single-Task (random init)	Mult	49.05	53.5	73.1	68.9	58.5
Single-task, pretrained (PT)	Mult	65.2	70.2	77.72	73.03	70.9
Answer-Me, PT, 4 tasks	Single	64.8	71.5	77.2	72.1	71.5
Answer-Me, PT, 8 tasks	Single	65.1	71.7	77.5	72.8	72.4
Answer-Me, PT, 8 tasks, 3x	Single	**73.6**	**73.9**	**85.8**	**77.5**	**74.3**

Table 2. Experiments comparing Answer-Me multi-task learning.

Approach	Model	VQA2.0	SNLI-VE
3-task + VQA2.0 (no SNLI-VE, ours)	Single	64.3	33.8
3-task + VQA2.0, FT on SNLI	Multiple	35.5	76.9
3-task + VQA2.0, FT on VQA2.0	Multiple	65.2	26.7
3-task + SNLI-VE (no VQA2.0, ours)	Single	33.2	76.5
3-task + SNLI-VE, FT on VQA2.0	Multiple	64.8	24.5
3-task + SNLI-VE, FT on SNLI	Multiple	29.4	77.2
Multi-task (w/o VQA2.0 or SN) Zero-Shot (ours)	Single	27.3	24.2
5-task (ours)	Single	65.1	77.4

Table 3. Fine-tuning vs multi-task Answer-Me. Fine tuned models tend to forget (first/second rows), even if original mix shows good within-data and out-of-sample (Zero-Shot) generalization (first rows). Additional fine-tuning seems to recover the losses within a task (first/third rows), but costs N times the cost in training, and performance on the other task deteriorates again. Interestingly, this model performs even worse than the original out-of-sample mixture on the second task. Training on many tasks in the mix maintains performance for both tasks (last row).

References

[1] Aishwarya Agrawal, Jiasen Lu, Stanislaw Antol, Margaret Mitchell, C. Lawrence Zitnick, Dhruv Batra, and Devi Parikh. Vqa: Visual question answering. In *ICCV*, 2015.

[2] Soravit Changpinyo, Jordi Pont-Tuset, Vittorio Ferrari, and Radu Soricut. Telling the what while pointing to the where: Multimodal queries for image retrieval. In *Arxiv 2102.04980*, 2021.

[3] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Universal image-text representation learning. In *ECCV*, 2020.

[4] Jaemin Cho, Jie Lei, Hao Tan, and Mohit Bansal. Unifying vision-and-language tasks via text generation. In *Arxiv 2102.02779*, 2021.

[5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[6] Tanmay Gupta, Amita Kamath, Aniruddha Kembhavi, and Derek Hoiem2. Towards general purpose vision systems. In *arxiv.org/abs/2104.00743*, 2021.

[7] Danna Gurari, Qing Li, Abigale J Stangl, Anhong Guo, Chi Lin Kristen Grauman Jiebo Luo, and Jeffrey P Bigham. VizWiz grand challenge: Answering visual questions from blind people. In *CVPR*, 2018.

[8] Ronghang Hu and Amanpreet Singh. Unit: Multimodal multitask learning with a unified transformer. In *arxiv.org/abs/2102.10772*, 2021.

[9] Drew A Hudson and Christopher D Manning. Gqa: a new dataset for compositional question answering over realworld images. In *CVPR*, 2019.

[10] Yu Liu, Lianghua Huang, Liuyihang Song, Bin Wang, Yingya Zhang, and Pan Pan. Enhancing textual cues in multi-modal transformers for VQA. In *VizWiz Challenge 2021*, 2021.

[11] Jiasen Lu, Vedanuj Goswami, Marcus Rohrbach, Devi Parikh, and Stefan Lee. 12-in-1: Multi-task vision and language representation learning. In *CVPR*, 2020.

[12] Gao Peng, Zhengkai Jiang, Haoxuan You, Pan Lu, Steven Hoi, Xiaogang Wang, and Hongsheng Li. Dynamic fusion with intra- and inter- modality attention flow for visual question answering. In *CVPR*, 2019.

[13] Alane Suhr, Mike Lewis, James Yeh, and Yoav Artzi. A corpus of natural language for visual reasoning. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, 2017.

[14] Ning Xie, Farley Lai, Derek Doran, and Asim Kadav. Visual entailment: A novel task for fine-grained image understanding. In *https://arxiv.org/abs/1901.06706*, 2019.