Enhancing Generalization in Medical Visual Question Answering Tasks via Gradient-Guided Model Perturbation

Gang Liu, Hongyang Li, Zerui He, Shenjun Zhong

Harbin Engineering University, China; Monash University, Australia

Introduction

- In the specialized field of Medical VQA, a significant challenge is the restricted size of publicly accessible datasets, often due to privacy considerations. A widely used strategy to address this limitation involves a pre-training and finetuning approach, where medical image caption datasets are typically used to learn a vision-language model which is subsequently transferred to medical VQA tasks, as demonstrated by the methods like M3AE and M2I2.

- To prevent overfitting, auxiliary regularization techniques are also commonly used in the training process, such as noise injection. Some works that add noise or perturbations to the model's weights are mainly designed to improve the model's robustness against adversarial attacks.

- In the context of a pre-training and fine-tuning framework, we further explore the options of introducing dynamic perturbations to model weights during training, to improve the model generalization on downstream medical VQA tasks. Our proposed method generates gradient-guided perturbations and integrates them into the visual encoder of the multi-modality model tailored for medical VQA tasks, in both pre-training and finetuning phases.

Methods

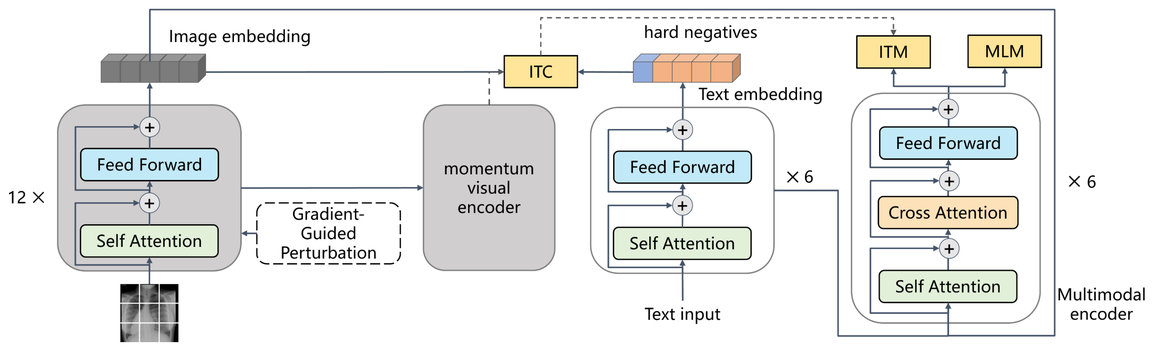

Fig. 1: Pre-training model architecture and objectives of the proposed method.

In our method, we specifically design the perturbations for the model weights of the visual encoder. The perturbations are formed as a minor offset along the direction of the moving averaged gradients of past model updates. In other words, the perturbations introduced to the visual encoder are against the direction of the most recent updates made by the optimizer in the optimization landscape. The gradient-guided perturbation, denoted as r_t at a given time t can obtained by the following equation:

$$r_t = \delta \cdot \frac{\|\theta_t\|}{\|\nabla_t\|} \cdot \nabla_t$$

As the gradients and model parameters update, an adaptive perturbation, r_i is added to the visual encoder in each iteration. The objective here is to modify the model weights in the opposite direction of the prior optimizations in the loss landscape. To minimize the risk of gradient explosion, we further apply clipping on the perturbed model weights as per the following equation:

$$\theta_t^{'} = \theta_{t-1} + \mathcal{C}lip(r_i, -\epsilon \cdot |\theta_{t-1}|, \epsilon \cdot |\theta_{t-1}|)$$

The associated pseudo-code for adding perturbations to the training process is presented in Algorithm 1. During each iteration, we obtain the first-order moment from the AdamW optimizer to generate the adaptive weight perturbation. The losses are calculated using the perturbed model, rather than the original weights. Subsequently, the gradients are updated to the unperturbed weights that are stored prior to each iteration.:

Algorithm 1 Pseudo code of perturbation in pre-training

Require: Training samples $\mathcal{D} = \{(\boldsymbol{x}_{img}, \boldsymbol{x}_{text}, \boldsymbol{y})\}$

for epoch $= 1 \ldots N$ **do**
  for minibatch $B \subset D$ **do**
    Take a snapshot of θ_{t-1}
    Compute the first-order gradient moment ∇_t
    Generate the perturbation r_t
    Get perturbed model $\boldsymbol{\theta}_t^{'}$, given $\boldsymbol{\theta}_{t-1}, r_t$
    Compute losses with perturbed model $\boldsymbol{\theta}_t^{'}$
    Update model $\boldsymbol{\theta}_{t-1} \rightarrow \boldsymbol{\theta}_t$
  end for
end for

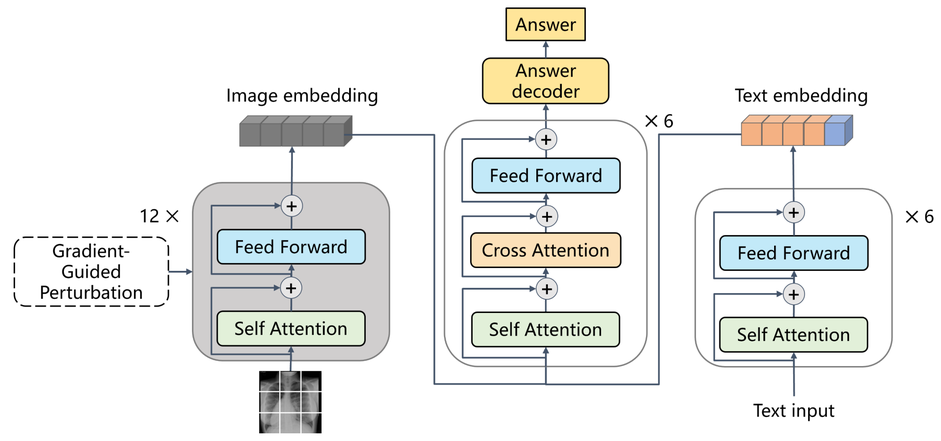

Fig. 2: Finetuning model structure..

Similar to the pre-training phase, we also introduce model perturbations in the fine-tuning process. The gradients for these perturbations are computed on the fly, from the conditional language-modeling loss.

Result

Table 1: Comparisons with the state-of-the-art methods on the VQA-RAD and SLAKE test set.

Method	Pre-Training Data	VQA-RAD			SLAKE		
		Open	Closed	Overall	Open	Closed	Overall
MEVF[16]	-	43.9	75.1	62.6	-	-	-
CPRD[17]	-	61.1	80.4	72.7	81.2	83.4	82.1
AMAM[18]	-	63.8	80.3	73.3	-	-	-
M2I2[2]	91k	66.5	83.5	76.8	74.7	91.1	81.2
M3AE[1]	401k	67.2	83.5	77.0	80.3	87.8	83.2
PMC-VQA[19]	177k	69.3	84.2	78.2	**88.2**	87.7	**88.0**
ours	81k	**70.39**	**84.56**	**78.94**	82.17	**89.9**	85.2

Table 2: The ablation study validated on VQA-RAD and SLAKE dataset. PT and FT mean perturbing the model in the pre-training and fine-tuning phases respectively, and APM means adaptive perturbation magnitude.

PT	FT	APM	VQA-RAD	SLAKE
			77.16	82.09
✓		✓	77.38	82.66
	✓	✓	78.05	83.88
✓	✓		77.83	82.94
✓	✓	✓	**78.94**	**85.2**

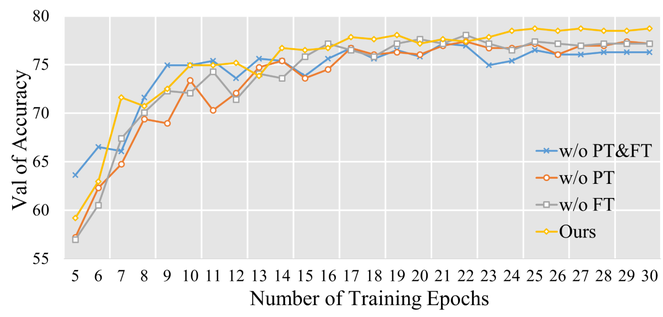

Fig. 3: Validation accuracies on VQA-RAD dataset during downstream VQA training, with different components enabled.

The method is designed as a regularization technique, therefore, we visualize the validation accuracy during training in Fig 3, to demonstrate its effectiveness. The plot shows that introducing gradient-guided perturbations in both pre-training and downstream tasks leads to a higher and more stable validation accuracy growth. This is especially notable when the performance gap becomes consistent after 24 epochs in the downstream VQA tasks.

Conclusion

In this paper, we propose a novel regularization technique for vision-language pre-training, that introduce adaptive perturbations to the visual encoder of the multi-modality model to enhance model generalization.

The method can be applied to both pre-training and downstream medical VQA tasks. The results show that the method improves the performance on downstream medical VQA tasks, while requiring less data for pre-training.

Importantly, our approach is model-agnostic, making it potentially useful in other uni-modal or multi-modal applications.