Revisiting Denoising as a Pretraining Method for Retrieval Tasks

Aadit Kant Jha*
Indraprastha Institute of Information Technology
New Delhi, India
aadit20001@iiitd.ac.in

Anindya Prithvi*
Indraprastha Institute of Information Technology
New Delhi, India
anindya20024@iiitd.ac.in

Madhava Krishna*
Indraprastha Institute of Information Technology
New Delhi, India
madhava20217@iiitd.ac.in

ABSTRACT

Model pre-training is a crucial aspect of retrieval, positively augmenting the performance of a model in a downstream task. For retrieval, broadly two classes of pre-training methods exist: masked modeling and contrastive learning. MAMO, a framework for pre-training models for retrieval tasks like question answering and captioning, uses the former, boasting state-of-the-art performance on vision-language tasks. Our work attempts to build on MAMO by switching the masked modeling tasks to denoising, which has previously shown good results in pure vision segmentation models in data-limited scenarios. We highlight our proposed framework, inspiration, and evaluation methodologies in this work.

KEYWORDS

Information, Retrieval, Pre-training, Multimodality, Transformers

ACM Reference Format:
Aadit Kant Jha, Anindya Prithvi, and Madhava Krishna. 2024. Revisiting Denoising as a Pretraining Method for Retrieval Tasks. In . ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION AND RELATED WORKS

Pre-training [6] has been a crucial part of modern deep learning, allowing practitioners to exploit the full power of transformers that have changed the landscape of deep learning. Usually done in unsupervised or weakly supervised settings to circumvent the data constraints, pre-training methods for classical image and text classification using contrastive[3, 4, 8, 9, 15] or masked modeling[1, 5, 13, 14, 16] methods have been shown to be of tremendous importance in robust feature extraction and augmenting downstream task performance. This extends to retrieval as well, and such strategies have been implemented and shown to boost performance [7, 10–12]. Our work aims to note the differences in pre-training when denoising is used instead of masked modeling in MAMO [17], a framework for vision-language representation learning for retrieval. We attempt to switch masked image modeling and masked language modeling to an image and text denoising task, respectively, and observe the impacts on the downstream task performance. To our knowledge, this is the first work incorporating denoising as a form of pre-training for retrieval.

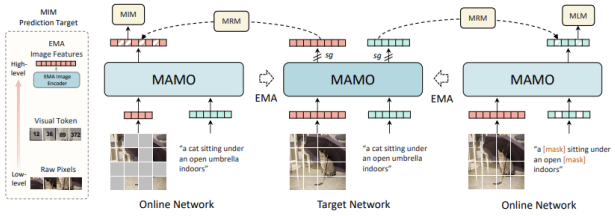

Figure 1: MAMO's architecture: MRM is the masked representation modeling framework, MIM is masked image modeling, and MLM refers to masked language modeling. EMA refers to the exponentially moving average of the target network. High-level image features are predicted in the masked modeling tasks.

2 OUR APPROACH

Our proposed method directly builds on MAMO [17] and takes inspiration from [2] and [16]. [2] uses denoising pretraining to train a decoder for semantic segmentation, outperforming supervised pre-training methods tremendously in data-limited scenarios. [16] is a popular vision-only masked image modeling method that attempts to learn good representations for fine-tuning by masking tokens and reproducing masked patches, utilizing a lightweight prediction head for the reconstruction. Our proposed approach seeks to primarily function in the case of data-limited scenarios.

MAMO has five stages 1:

(1) Masked representation modeling: a target network involving an exponentially weighted moving average (EMA) of the model to be trained, fixed wrt gradient calculations predicts multimodal embeddings. The embeddings are matched for (masked image, clean text) and (clean image, masked text) pairs with the joint embedding of the target network with an L2 loss.
(2) Masked image modeling: the multimodal representation for the (masked image, clean text) pair is matched with the visual representation for (clean image, clean text) with an L1 loss.
(3) Masked language modeling: the multimodal representation for the (clean image, masked text) pair is used to optimize the cross-entropy loss against the original tokens in masked positions.
(4) Image-Text Contrastive Learning (ITC): global alignment between image-text pairs by minimizing the InfoNCE loss.

*All authors contributed equally to this research.

Conference'17, July 2017, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM
https://doi.org/10.1145/nnnnnnn.nnnnnnn

Method	#Images	MSCOCO (5k test set)		Flickr30K (1k test set)	
		TR	IR	TR	IR
		R@1 / R@5 / R@10	R@1 / R@5 / R@10	R@1 / R@5 / R@10	R@1 / R@5 / R@10
UNITER$_{Large}$ [6]	4M	65.7 / 88.6 / 93.8	52.9 / 79.9 / 88.0	87.3 / 98.0 / 99.2	75.6 / 94.1 / 96.8
ALBEF [23]	4M	73.1 / 91.4 / 96.0	56.8 / 81.5 / 89.2	94.3 / 99.4 / 99.8	82.8 / 96.7 / 98.4
TCL [44]	4M	75.6 / 92.8 / 96.7	59.0 / 83.2 / 89.9	94.9 / 99.5 / 99.8	84.0 / 96.7 / 98.5
CODIS [11]	4M	75.3 / 92.6 / 96.6	58.7 / 82.8 / 89.7	95.1 / 99.4 / 99.9	83.3 / 96.1 / 97.8
VLC [15]	5.6M	71.3 / 91.2 / 95.8	50.7 / 78.9 / 88.0	89.2 / 99.2 / 99.8	72.4 / 93.4 / 96.5
VLMo$_{Base}$ [3]	4M	74.8 / 93.1 / 96.9	57.2 / 82.6 / 89.8	92.3 / 99.4 / 99.9	79.3 / 95.7 / 97.8
METER-Swin [10]	4M	73.0 / 92.0 / 96.3	54.9 / 81.4 / 89.3	92.4 / 99.0 / 99.5	79.0 / 95.6 / 98.0
MaskVLM [22]	4M	76.3 / 93.8 / 96.8	60.1 / 83.6 / 90.4	95.6 / 99.4 / 99.9	84.5 / 96.7 / 98.2
ALIGN [18]	1.8B	77.0 / 93.5 / 96.9	59.9 / 83.3 / 89.8	95.3 / **99.8** / **100.0**	84.9 / **97.4** / **98.6**
MAMO	4M	**79.1** / **94.9** / **97.8**	**62.4** / **85.3** / **91.3**	**96.2** / 99.5 / 99.8	**86.1** / 97.0 / 98.4

Figure 2: Fine-tuned image-text retrieval results on MSCOCO and Flickr30K datasets. IR: Image retrieval. TR: Text retrieval.

(5) Image-text matching learning (ITM): match a hard negative sample for each image and text, then calculate the cross entropy loss between the matching probabilities for each image-text pair and the corresponding matching label.

Our proposed method aims to replace the masked representation, image, and language modeling subtasks with representation, text, and image denoising, respectively. This shall be done in the following manner:

(1) Representation denoising: Input a (noisy image, clean text) and (clean image, noisy text) pair, and match the representations with the (clean image, clean text) multimodal embeddings from the EMA model.
(2) Image Denoising: Provide a noisy image by adding Gaussian noise and seek to denoise it using the multimodal embedding, matching it with the clean image reconstructed from the embeddings from the (clean image, clean text) input to the EMA model's image decoder.
(3) Text Denoising: Replace words in the sentence with random words in the corpus and attempt to decipher the noisy words. This serves as a noise prediction task.

Wherever decoding is required, we shall use SimMIM's [16] ideology and resort to a lightweight prediction head. Our premise is that if successful denoising is to be done, the model should know what the clean input should look like, which should augment representation learning for this task.

3 EVALUATION

The evaluation will be conducted using similar parameters and hyperparameters as given in the main paper, with compute-saving measures to be used wherever necessary. We shall use a subset of the MS-COCO and the FLICKR30K datasets for pre-training, in order to assess the learning capability under data-limited conditions. To reduce the compute requirements, we will resort to a ViT-T and BERT-T model instead of the ViT-B and BERT-B vision and text encoders, respectively, mentioned in the base MAMO paper. The performance metric for our study will be retrieval at 1, 5, and 10. Specifically, the pre-training will be assessed by fine-tuning the pre-trained model on image-retrieval and text-retrieval on the FLICKR30K and MSCOCO datasets. If time allows, we shall conduct ablative analyses and compare how our changes affect the pretraining method. We shall also attempt to determine the data percentage scaling for the number of pretraining samples. The baseline results of MAMO reported in [17] are shown in Figure 2.

REFERENCES

[1] Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. 2021. Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254* (2021).
[2] Emmanuel Asiedu Brempong, Simon Kornblith, Ting Chen, Niki Parmar, Matthias Minderer, and Mohammad Norouzi. 2022. Denoising pretraining for semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 4175–4186.
[3] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*. PMLR, 1597–1607.
[4] Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey E Hinton. 2020. Big self-supervised models are strong semi-supervised learners. *Advances in neural information processing systems* 33 (2020), 22243–22255.
[5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).
[6] Dumitru Erhan, Aaron Courville, Yoshua Bengio, and Pascal Vincent. 2010. Why Does Unsupervised Pre-training Help Deep Learning?. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics (Proceedings of Machine Learning Research, Vol. 9)*, Yee Whye Teh and Mike Titterington (Eds.). PMLR, Chia Laguna Resort, Sardinia, Italy, 201–208. https://proceedings.mlr.press/v9/erhan10a.html
[7] Valentin Gabeur, Arsha Nagrani, Chen Sun, Karteek Alahari, and Cordelia Schmid. 2022. Masking modalities for cross-modal video retrieval. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 1766–1775.
[8] Beliz Gunel, Jingfei Du, Alexis Conneau, and Ves Stoyanov. 2020. Supervised contrastive learning for pre-trained language model fine-tuning. *arXiv preprint arXiv:2011.01403* (2020).
[9] Dan Iter, Kelvin Guu, Larry Lansing, and Dan Jurafsky. 2020. Pretraining with contrastive sentence objectives improves discourse performance of language models. *arXiv preprint arXiv:2005.10389* (2020).
[10] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*. PMLR, 4904–4916.
[11] Gukyeong Kwon, Zhaowei Cai, Avinash Ravichandran, Erhan Bas, Rahul Bhotika, and Stefano Soatto. 2022. Masked vision and language modeling for multi-modal representation learning. *arXiv preprint arXiv:2208.02131* (2022).
[12] Yizhi Li, Zhenghao Liu, Chenyan Xiong, and Zhiyuan Liu. 2021. More robust dense retrieval with contrastive dual learning. In *Proceedings of the 2021 ACM SIGIR International Conference on Theory of Information Retrieval*. 287–296.
[13] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692* (2019).
[14] Zhiliang Peng, Li Dong, Hangbo Bao, Qixiang Ye, and Furu Wei. 2022. Beit v2: Masked image modeling with vector-quantized visual tokenizers. *arXiv preprint arXiv:2208.06366* (2022).
[15] Yujia Qin, Yankai Lin, Ryuichi Takanobu, Zhiyuan Liu, Peng Li, Heng Ji, Minlie Huang, Maosong Sun, and Jie Zhou. 2020. ERICA: improving entity and relation understanding for pre-trained language models via contrastive learning. *arXiv preprint arXiv:2012.15022* (2020).
[16] Zhenda Xie, Zheng Zhang, Yue Cao, Yutong Lin, Jianmin Bao, Zhuliang Yao, Qi Dai, and Han Hu. 2022. Simmim: A simple framework for masked image modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 9653–9663.
[17] Zijia Zhao, Longteng Guo, Xingjian He, Shuai Shao, Zehuan Yuan, and Jing Liu. 2023. MAMO: Fine-Grained Vision-Language Representations Learning with Masked Multimodal Modeling. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1528–1538.