Denoising Multimodal Pretraining for Retrieval

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. Denoising Multimodal Pretraining for Retrieval. In . ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

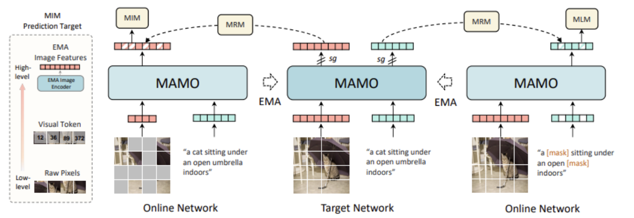

Figure 1: MAMO's architecture: MRM is the masked representation modeling framework, MIM is masked image modeling, and MLM refers to masked language modeling. EMA refers to the exponentially moving average of the target network. High-level image features are predicted in the masked modeling tasks.

1 INTRODUCTION AND RELATED WORKS

Pre-training [9] 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[5, 6, 12, 19], self-distillation [5, 10, 11, 21], canonical correlation analysis [3, 25] or masked modeling[2, 8, 16, 18, 23] methods have been shown to be of tremendous importance in robust feature extraction and augmenting downstream task performance [1].

However, multimodal pretraining with image-text inputs varies significantly from unimodal pretraining primarily because of the shared embedding space between visual and textual modalities. A lot of methods have emerged that take this into account, with contrastive learning to align image-text representations taking the front seat [13, 20] by maximising the InfoNCE loss [17]. Some methods also optimize the image-text matching loss, which determines the probability that an image-text pair corresponds to each other [7, 14, 15, 26]. Recent methods also perform unimodal [15] and joint representation [14] pretraining to optimize unimodal and multimodal representations.

2 OUR APPROACH

Our proposed method builds on MAMO [26], MaskVLM [14] and takes inspiration from MaskFeat [23] and VicREG [3]. Our proposed approach primarily seeks to function in data-limited scenarios.

MAMO has five stages 1:

(1) Masked representation modeling (MRM): 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 (MIM): 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 (MLM): 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.
(5) Image-text matching learning (ITM): match a hard negative sample for each image and text, then calculate the cross

*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) TR R@1 / R@5 / R@10	MSCOCO (5k test set) IR R@1 / R@5 / R@10	Flickr30K (1k test set) TR R@1 / R@5 / R@10	Flickr30K (1k test set) IR 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.

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, instead choosing to do MIM conditioned on clean text, and MLM conditioned on clean image as in [14]. We also seek to reduce the computational complexity by removing the EMA target network and instead choosing to predict features of a low-level trivial visual transform, like HOG or plain pixels, as with [23]. The removal of the EMA part also come from [5] which observed slower learning with EMA self-distillation. The loss shall comprise of the following components:

(1) MIM with text joint modeling: Input a (noisy image, clean text) and match the representations from the fusion encoder to the target HOG representations from a target network.
(2) MLM with image joint modeling: Input a (clean image, noisy text) pair with a higher text masking ratio and predict masked tokens from the joint representation from the fusion encoder.
(3) We shall keep the ITM + ITC losses as is, as they help align the image and text modalities. We shall experiment by ablating the ITC component as the next component aims to address similarities.
(4) Text-image embedding matching, feature variance maximization and feature covariance minimization. This derives from VicREG and will aim to ensure that the feature embeddings are informative and non-redundant.

Wherever decoding is required, we shall use SimMIM's [24] ideology and resort to a lightweight prediction head.

3 EVALUATION

We shall pretrain and evaluate on the MSCoco and Flickr30k datasets in order to assess the learning capability under data-limited conditions. To reduce the compute requirements, we will resort to a ViT-S and BERT-S 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, we assess the performance 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 [26] are shown in Figure 2.

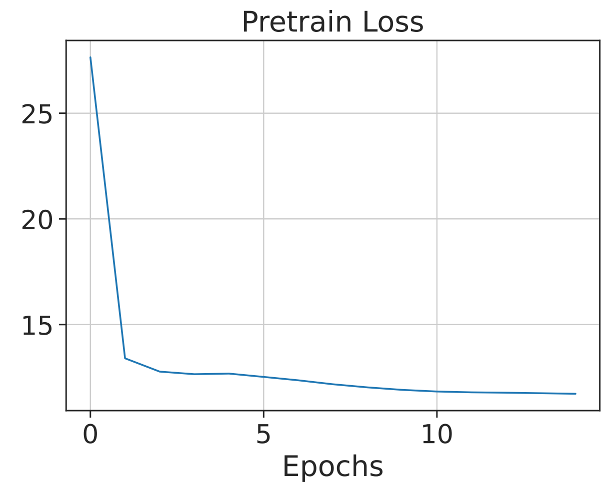

Figure 3: The net loss during pre-training. The loss is composed of Masked Representation Modeling (MRM), Masked Image Modeling (MIM), Masked Language Modeling (MLM), Image-Text Contrastive loss (ITC) and Image-Text Matching (ITM) losses.

4 PROGRESS

The source code for MAMO was not available on any site, so we had to build it from the paper ourselves. For now, we have successfully implemented all aspects in the MAMO paper to build our baseline and pre-trained a model on the Flickr30k dataset because it is computationally cheap compared to the MSCoco dataset. Reference was taken from the ALBEF repository [15] for building parts of the model.

We pre-trained the model for 15 epochs with 3 epochs of warmup. We employed a cosine learning rate schedule with an initial learning rate of $2.5e-4$, warmup learning rate of $1e-6$, and minimum learning rate of $1e-5$ with an AdamW optimizer with decay 0.01, β_1 of 0.9 and β_2 of 0.999. We took the α parameter in the EWMA target network to be 0.995 and used a text-masking and image-masking ratio of 0.25 and 0.75. A batch size of 96 was used as it was the largest that could be accommodated in memory. Standard preprocessing techniques like stopword and punctuation removal, and lowercasing were applied for the captions.

We initialized a ViT-S model trained on the ImageNet-22k dataset [22] and a BERT-S model [4]. We used the first two layers of the BERT-S text encoder for the text embeddings and the last two layers for the multimodal fusion encoder.

The loss curves for the masked representation modeling (MRM), masked image modeling (MIM), masked language modeling (MLM), image-text contrastive learning (ITC), and image-text matching (ITM), when pre-trained on the Flickr30K dataset, are given in figures 4, 3, respectively. The weights for the pre-trained and fine-tuned models can be found **here**.

Preliminary fine-tuning results, consisting of the same optimizer parameters as with pre-training but with a batch size of 144 and warmup and training epochs of 3 and 15, respectively, result in a

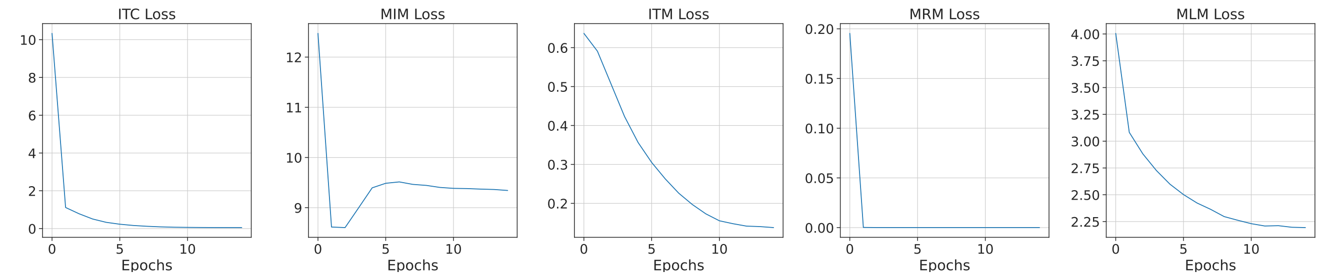

Figure 4: The loss curves of each component contributing to the total pre-training loss. The MRM loss matches masked regions of the joint representations of the online and target networks, the MIM loss matches visual representations of the online and target networks, the MLM component predicts masked words as with BERT pre-training, except with masked locations being replaced by [MASK] tokens, the ITC component pushes similar caption/text pairs together, and the ITM component matches if a caption corresponds to an image.

Dataset	Metrics					
	Image Retrieval			Text Retrieval		
	R@1	R@5	R@10	R@1	R@5	R@10
Flickr30K	55.56	79.14	84.94	67.8	88.1	94.3

Table 1: Performance for MAMO observed after pre-training and finetuning on the Flickr30K dataset. The pre-training and warmup took place for 15 epochs with a warmup period of 3 epochs.

good performance on the test set when retrieved from a set of 128 samples, as shown in table 1.

REFERENCES

[1] Randall Balestriero, Mark Ibrahim, Vlad Sobal, Ari Morcos, Shashank Shekhar, Tom Goldstein, Florian Bordes, Adrien Bardes, Gregoire Mialon, Yuandong Tian, et al. 2023. A cookbook of self-supervised learning. *arXiv preprint arXiv:2304.12210* (2023).

[2] Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. 2021. Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254* (2021).

[3] Adrien Bardes, Jean Ponce, and Yann LeCun. 2021. Vicreg: Variance-invariance-covariance regularization for self-supervised learning. *arXiv preprint arXiv:2105.04906* (2021).

[4] Prajjwal Bhargava, Aleksandr Drozd, and Anna Rogers. 2021. Generalization in NLI: Ways (Not) To Go Beyond Simple Heuristics. arXiv:2110.01518 [cs.CL]

[5] 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.

[6] 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.

[7] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. Uniter: Universal image-text representation learning. In *European conference on computer vision*. Springer, 104–120.

[8] 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).

[9] 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

[10] Tommaso Furlanello, Zachary Lipton, Michael Tschannen, Laurent Itti, and Anima Anandkumar. 2018. Born again neural networks. In *International conference on machine learning*. PMLR, 1607–1616.

[11] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. 2020. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems* 33 (2020), 21271–21284.

[12] 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).

[13] 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.

[14] 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).

[15] Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. 2021. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems* 34 (2021), 9694–9705.

[16] 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).

[17] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748* (2018).

[18] 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).

[19] 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).

[20] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*. PMLR, 8748–8763.

[21] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. 2020. Training data-efficient image transformers & distillation through attention. arXiv 2020. *arXiv preprint arXiv:2012.12877* (2020).

[22] ViT-S by Winkwaks [n. d.]. HuggingFace.

[23] C Wei, H Fan, S Xie, C Wu, AL Yuille, and C Feichtenhofer. 2021. Masked feature prediction for self-supervised visual pre-training. CoRR abs/2112.09133 (2021).

[24] 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.

[25] Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. 2021. Barlow twins: Self-supervised learning via redundancy reduction. In *International conference on machine learning*. PMLR, 12310–12320.

[26] 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.