Multimodal Masked Autoencoders Learn Transferable Representations

Anonymous Authors[1]

Abstract

Building scalable models to learn from diverse, multimodal data remains an open challenge. For vision-language data, the dominant approaches are based on contrastive learning objectives that train a separate encoder for each modality. While effective, contrastive learning approaches introduce sampling bias depending on the data augmentations used, which can degrade performance on downstream tasks. Moreover, these methods are limited to paired image-text data, and cannot leverage widely-available unpaired data. In this paper, we investigate whether a large multimodal model trained purely via masked token prediction, without using modality-specific encoders or contrastive learning, can learn transferable representations for downstream tasks. We propose a simple and scalable network architecture, the Multimodal Masked Autoencoder (M3AE), which learns a unified encoder for both vision and language data via masked token prediction. We provide an empirical study of M3AE trained on a large-scale image-text dataset, and find that M3AE is able to learn generalizable representations that transfer well to downstream tasks. We demonstrate the scalability of M3AE with larger model size and training time, and its flexibility to train on both paired image-text data as well as unpaired data.

1. Introduction

With the rapid advances in neural architectures [41] and hardware performance, self-supervised pre-training has made tremendous progress in natural language processing (NLP) and vision [17, 10, 2, 3]. The underlying idea, often referred as masked token prediction, is conceptually simple: the model learns to predict a removed portion of the data. Masked token prediction has enabled highly successful methods for pre-training in NLP and vision, including Transformer [41], GPT [3], BERT [10], and MAE [17]. These pre-trained representations have been shown to generalize well to various downstream tasks.

Driven by these successes, there has been significant interest in improving visual representation learning by training on large and diverse multimodal datasets that contains both images and text. These datasets, such as CC12M [4] and YFCC100M [36], are often much more scalable than explicitly labeled datasets such as ImageNet [9], and the diverse language data can provide rich supervision to train more generalizable representations.

The dominant paradigm for multimodal pre-training is cross-modal contrastive learning, such as CLIP [30] and ALIGN [22]. These methods show that cross-modal contrastive learning models, trained on giant corpora of paired image-and-text, can generalize well to various downstream tasks. Despite these progresses, a major limitation for contrastive learning is that it requires paired image-and-text data and therefore cannot leverage widely available unpaired data. In addition, contrastive learning based methods use separate encoders for image and text, making it difficult to access information from different modalities at the same time.

To address the above limitations for visual representation learning, we propose a simple and scalable architecture called the multimodal masked autoencoders (M3AE) for learning a single unified model on large image and language data, without using modality-specific encoders or contrastive learning. Based on MAE [17], M3AE is trained purely via masked token prediction. Our key idea is to treat an image-and-text pair as a long sequence of tokens consisting of embeddings of image patches and text. M3AE is trained simply by masking random patches of the input image and language tokens, and learning to reconstruct the masked pixels and text.

In this paper, we provide an empirical study of M3AE trained on the multimodal CC12M [4] dataset, and find that M3AE is able to learn generalizable representations that transfer well to downstream tasks such as image clas-

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the First Workshop of Pre-training: Perspectives, Pitfalls, and Paths Forward at ICML 2022. Do not distribute.

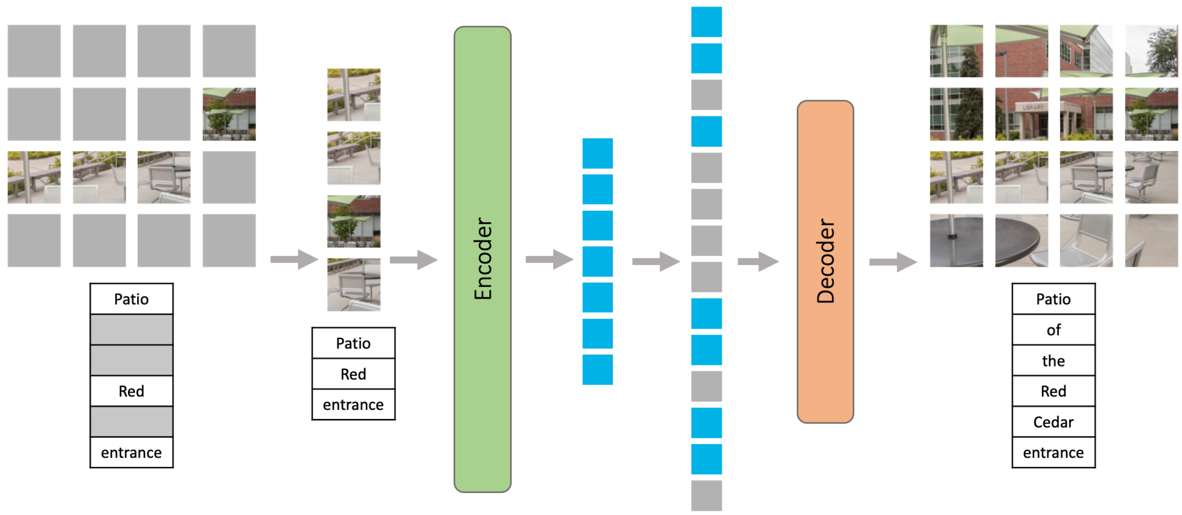
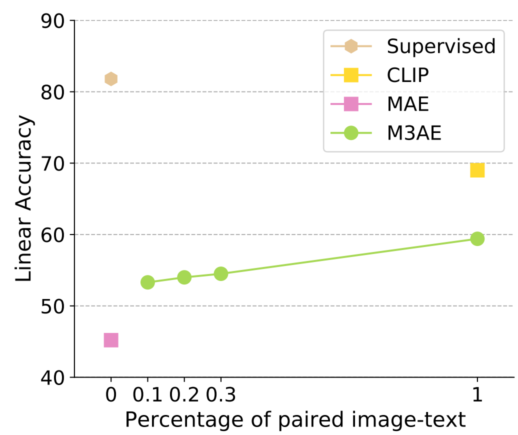


Figure 1. Multimodal masked autoencoder (M3AE) consists of an encoder that maps language tokens and image patches to a shared representation space, and a decoder that reconstructs the original image and language from the representation.

sification and out-of-distribution detection. We find that multimodal pre-training of M3AE on CC12M achieves significantly higher performance on the ImageNet-1k linear classification benchmark [33] compared to pre-training on images only (MAE). Our strong results for M3AE demonstrate the generalization benefits of multimodal training for learning transferable representations across datasets.

Surprisingly, we find that M3AE performs best when we apply a high mask ratio (75%) on language, while in contrast, language models like BERT [10] conventionally use a low mask ratio (15%) because language data are highly semantic and information-dense. We hypothesize that M3AE benefits from a higher mask ratio on text because it enforces a better joint understanding of vision and language during masked token prediction. We also provide qualitative analysis showing that the learned representation incorporates meaningful information from both image and language. Furthermore, we demonstrate the scalability of M3AE with larger model size and training time, as well as its flexibility to train on both paired image-text data as well as unpaired data.



Model	MAE	M3AE	CLIP	Supv
Accuracy	45.2	59.4	69.0	81.8
M3AE text ratio	10%	20%	30%	100%
Accuracy	53.3	54.0	54.5	59.4

Figure 2. Comparison of M3AE, MAE, and CLIP on ImageNet. M3AE significantly outperforms MAE. M3AE can flexibly leverage a combination of paired image-text data and unpaired image only data. All models are ViT-B. MAE and M3AE are pretrained on CC12M for 100 epochs.

2. MultiModal Masked Autoencoder (M3AE)

In this section we introduce our method, multimodal masked autoencoder (M3AE). M3AE consists of an encoder that maps image and language to representation space, and a decoder that reconstructs the original image and language from the representation. We summarize the main architecture and training process of M3AE in Figure 1.

Image-language masking. The first step of M3AE is to combine the language and image input into a single sequence. Following standard natural language processing practice [10], we tokenize the input text into a sequence of discrete tokens. For image input, we divide it into regular non-overlapping patches of pixels, following the practice of ViT [12]. Text tokens and image patches are then concatenated into a single sequence.

For patches and tokens, we sample s random subset without replacement from a uniform distribution, and mask (*i.e.*, remove) the remaining ones. A *high* masking ratio is applied to both text tokens and image patches, in order to eliminate information redundancy and make a sufficiently difficult task that cannot be easily solved by extrapolation from visible neighboring patches and tokens.

M3AE encoder. The M3AE architecture consists of two networks: an encoder and a decoder. The encoder is a large Transformer, following the architecuture of ViT [12] and BERT [10]. The encoder takes only *unmasked (visible)* language tokens and image patches as input. For language tokens, we first convert it into learnable embedding vectors and then apply 1D positional encodings, following the standard practice [10]. For image patches, we use a learnable linear projection to convert them to image embeddings that have the same dimension as the language embeddings, and then apply 2D positional encodings, following the practice of MAE [17]. In order to distinguish the two different modalities, we add two learnable vectors that represent language and images respectively to the corresponding modalities' embeddings. We call these "modality type encodings". Additionally, a learnable CLS embedding [10] is prepended to the beginning of the sequence. The combined language and image embeddings are then processed by a series of Transformer blocks to obtain the final representation. Although the input consists of long sequences of image patches and text tokens, we can still train very large Transformer encoders efficiently because the same only operates on a small subset (*e.g.*, 25%) of the full set.

M3AE decoder. Following MAE [17], we use a lightweight Transformer-based decoder on the full set of tokens consisting of (i) encoded visible image patches, (ii) encoded visible text tokens, and (iii) mask tokens. Each mask token is a shared, learned vector that indicates the presence of a missing patch or token to be predicted. We add positional embeddings to all tokens in this full set in order to encode location information in mask tokens. We also add a different set of modality type embeddings to visible tokens, similar to the encoder. After the decoder Transformer, we apply two linear projection output heads to compute the reconstruction. The image output head projects the decoder output corresponding to image patches to the same dimension as pixels in the original image patches. The language output head projects the decoder output of language to token logits. These output heads are then used for supervision during the self-supervised training of M3AE.

Self-supervised training of M3AE. Our M3AE reconstructs the input by predicting the *pixel* values for masked image patches and the token probabilities for masked language tokens. For image reconstruction, we compute the mean squared error (MSE) between the reconstructed and original images in the pixel space. For language reconstruction, we apply the cross entropy loss between the reconstructed and original text. Our loss is a weighted sum of the image loss and the text loss. Similar to MAE [17] and BERT [10], we compute the loss only on the masked image patches and language tokens. Since M3AE processes image and language data uniformly by combining them into a single sequence, a natural advantage for our model is that it can be trained with the exact same loss on a mixture of paired and unpaired data, significantly extending the applicability of our model beyond what is possible with contrastive learning.

3. Experiments

In this section, we study the representation quality of M3AE. We aim to answer the following questions in our experiments: **(1)** Can M3AE learn generalizable visual representations that transfer well to downstream tasks? **(2)** Does the learned representation incorporate meaningful information from both images and language? **(3)** Does M3AE scale well with model size and training time?

To answer these questions, we first pre-train the M3AE model on a diverse image-and-language dataset and evalute its performance for downstream classification and out-of-distribution detection. We further evaluate the scalability of the model with respect to training epochs and model size. Finally, we provide a detailed ablation study and qualitative analyses to inspect the quality of the learned representations.

3.1. Experiment Setup

Pre-training datasets. M3AE is trained on Conceptual 12M (CC12M) [4]. The original dataset images are provided in the form of internet URLs. Note that due to some expired URLs and non-English captions, we did not obtain the complete data in the dataset. For language data, we use the BERT tokenizer to tokenize the text. We provide more details about data preprocessing in Section C.1.

Downstream datasets. We assess model performance in a wider variety of distributions and tasks. We evaluate the image encoder transferability on ImageNet [33]. We report top-1 validation accuracy of a single 256×256 crop.

Network architectures. Following MAE, we use ViT [12] as the model architecture and consider three different sizes of ViT for the M3AE image and text encoder. We use the original ViT-B/16 and ViT-L/16 architectures [12] for our encoder, as well as ViT-S/16 [37] which is comparable to ResNet-50 in FLOPs and parameters. Following MAE [17], our decoder is lightweight and has 8 blocks of width 512. More details can be found in Section C.3.

Pre-training setup. For comparison with MAE, we train our model from scratch for the same number of epochs. The loss weights of image prediction and text prediction are set to 1 and 0.5. The mask ratio for image and text are both set to 0.75. Refers to Section C.4 for more details.

Downstream evaluation setup. We evaluate our model transferability by performing linear classification on frozen features, *i.e.*, the pre-trained image encoder is fixed and serves as a feature extractor. After feature extraction, we train the linear classifier with the AdamW [27] optimizer as the same in He et al. [17]. More details can be found in Section C.5

Epoch Scaling	25	50	100
M3AE	62.9	64.1	64.8
MAE	47.8	50.1	55.2
Model Scaling	ViT-S	ViT-B	ViT-L
M3AE	38.2	58.2	64.1
MAE	32.1	44.2	50.1

Table 1. **Top**. ViT-L/16 with longer pre-training schedules (25/50/100 epochs). **Bottom**. Comparing ViT model variants of different capacities (ViT-S/B/L). All models are pre-trained for 50 epochs. We see that M3AE scales well with model size and training epochs, outperforming MAE in every setting.

3.2. Results

ImageNet Classification. We evaluate performance on ImageNet under the linear classification setting. Linear classification, also called linear probing, is a standard evaluation method used to evaluate unsupervised or self-supervised representations. A randomly initialized final classification layer is trained while all other model weights are frozen.

Figure 2 shows the results of linear classification. We report the results of ViT-B trained on ImageNet and CLIP [30] pre-trained on CC12M from prior work [37, 29].To study the flexibility of M3AE, we remove the text for a portion of image-text pairs, *i.e.*, 30% of paired image-text examples means 70% of CC12M image-text pairs become images only. A lower percentage of paired image-text data contains less information and therefore makes the task more difficult, since the model has to infer the relation between visual and language concepts based on limited paired data.

The comparison between M3AE and the baselines are shown in Figure 2. M3AE significantly outperforms MAE by nearly 10 percent. CLIP is a strong baseline based on cross-modal contrastive learning. While it achieves higher accuracy than M3AE, it is less flexible than our model since it can only use paired image-text data. In contrast, M3AE can leverage both paired image-text and unpaired image data without modifying the training procedure, as shown in Figure 2, giving our model strong potential to leverage a diverse combination of unpaired single modality and multi-modal data. Notably, with M3AE pre-training, even adding 10% noise to the text annotation leads to a significant boost in accuracy over MAE (53.3% vs 45.2%).

Text mask ratio	0.15	0.5	0.75	0.9
Accuracy	53.3	62.5	64.1	62.6

Table 2. Comparing M3AE with different text mask ratio. We see that M3AE performs the best with a surprisingly high text mask ratio of 75%.

We make an important note that the linear classification performance of MAE pre-trained on CC12M is much lower than MAE pre-trained on ImageNet, and we hypothesize that such a difference is caused by the **large domain gap between the two datasets**. To confirm this hypothesis, we pre-trained a ViT-L MAE on ImageNet for 800 epochs using the same hyperparameters on top of our implementation, and obtained 73.5% accuracy on linear classification, which exactly matches the original reported performance [17]. Thus, while our results cannot be directly compared to the original MAE results [17] pre-trained on ImageNet due to distribution mismatch, they demonstrate the strengths of multimodal training of M3AE for learning transferable representations across datasets.

Model scaling and epochs scaling. We also investigate the scaling behavior of M3AE with more training epochs and larger Transformer models. We note that because CC12M is 10 times larger than ImageNet-1K, the number of gradient steps in 100 epochs of pre-training corresponds to around 800 epochs on ImageNet-1K. Table 1 show holding model size fixed (ViT-B/16) and training for longer as well as training different model sizes for an extended training schedule (50 epochs). Our results indicate that M3AE scales well with both longer training and larger models.

Ablation on text mask ratio. We also investigate the performance of M3AE under various text mask ratios. Table 2 shows holding the image patch mask ratio fixed (75%) and training for various text mask ratios. Surprisingly, the results indicate that M3AE benefits from a high text mask ratio (50%-90%), contrary to BERT [10] whose typical masking ratio is 15%. We believe that this is the result of joint training of two modalities of data, where the masked language prediction can make use of information from both the visible language tokens and image patches.

4. Conclusion and Future Work

In this paper, we propose M3AE, a simple but effective model that learns multimodal representations from image and language without the need for contrastive objectives. We show that by pre-training with diverse image and language data, our model can learn shared representations that generalize well to downstream tasks. Due to its flexibility and scalability, M3AE is especially suitable for learning from extremely large-scale datasets, and we envision that such pre-trained models can be broadly applicable in many practical downstream tasks, such as visual reasoning [11], dialog systems [1] and language guided image generation [31, 32].

References

[1] Alayrac, J.-B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., Ring, R., Rutherford, E., Cabi, S., Han, T., Gong, Z., Samangooei, S., Monteiro, M., Menick, J., Borgeaud, S., Brock, A., Nematzadeh, A., Sharifzadeh, S., Binkowski, M., Barreira, R., Vinyals, O., Zisserman, A., and Simonyan, K. Flamingo: a visual language model for few-shot learning. *arXiv preprint arXiv: Arxiv-2204.14198*, 2022.

[2] Bao, H., Dong, L., and Wei, F. Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254*, 2021.

[3] Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[4] Changpinyo, S., Sharma, P., Ding, N., and Soricut, R. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3558–3568, 2021.

[5] Chen, M., Radford, A., Child, R., Wu, J., Jun, H., Luan, D., and Sutskever, I. Generative pretraining from pixels. In *International Conference on Machine Learning*, pp. 1691–1703. PMLR, 2020.

[6] Chen, T., Kornblith, S., Norouzi, M., and Hinton, G. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pp. 1597–1607. PMLR, 2020.

[7] Chen, T., Kornblith, S., Norouzi, M., and Hinton, G. E. A simple framework for contrastive learning of visual representations. *ArXiv*, abs/2002.05709, 2020.

[8] Chen, Y.-C., Li, L., Yu, L., Kholy, A. E., Ahmed, F., Gan, Z., Cheng, Y., and Liu, J. Uniter: Universal image-text representation learning. *arXiv preprint arXiv: Arxiv-1909.11740*, 2019.

[9] Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

[10] Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[11] Ding, D., Hill, F., Santoro, A., Reynolds, M., and Botvinick, M. Attention over learned object embeddings enables complex visual reasoning. *Advances in Neural Information Processing Systems*, 34, 2021.

[12] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[13] Fort, S., Ren, J., and Lakshminarayanan, B. Exploring the limits of out-of-distribution detection. *arXiv preprint arXiv: Arxiv-2106.03004*, 2021.

[14] Gao, T., Yao, X., and Chen, D. Simcse: Simple contrastive learning of sentence embeddings. *ArXiv*, abs/2104.08821, 2021.

[15] Grill, J.-B., Strub, F., Altch'e, F., Tallec, C., Richemond, P. H., Buchatskaya, E., Doersch, C., Pires, B. Á., Guo, Z. D., Azar, M. G., Piot, B., Kavukcuoglu, K., Munos, R., and Valko, M. Bootstrap your own latent: A new approach to self-supervised learning. *ArXiv*, abs/2006.07733, 2020.

[16] He, K., Fan, H., Wu, Y., Xie, S., and Girshick, R. B. Momentum contrast for unsupervised visual representation learning. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9726–9735, 2020.

[17] He, K., Chen, X., Xie, S., Li, Y., Dollár, P., and Girshick, R. Masked autoencoders are scalable vision learners. *arXiv preprint arXiv:2111.06377*, 2021.

[18] Hendrycks, D., Mazeika, M., and Dietterich, T. Deep anomaly detection with outlier exposure. *arXiv preprint arXiv:1812.04606*, 2018.

[19] Hendrycks, D., Lee, K., and Mazeika, M. Using pre-training can improve model robustness and uncertainty. In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pp. 2712–2721. PMLR, 2019.

[20] Hendrycks, D., Liu, X., Wallace, E., Dziedzic, A., Krishnan, R., and Song, D. Pretrained transformers improve out-of-distribution robustness. *arXiv preprint arXiv: Arxiv-2004.06100*, 2020.

[21] Jaegle, A., Gimeno, F., Brock, A., Zisserman, A., Vinyals, O., and Carreira, J. Perceiver: General perception with iterative attention. *arXiv preprint arXiv: Arxiv-2103.03206*, 2021.

[22] Jia, C., Yang, Y., Xia, Y., Chen, Y.-T., Parekh, Z., Pham, H., Le, Q., Sung, Y.-H., Li, Z., and Duerig, T. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning*, pp. 4904–4916. PMLR, 2021.

[23] Krizhevsky, A., Hinton, G., et al. Learning multiple layers of features from tiny images. 2009.

[24] Lee, K., Lee, K., Lee, H., and Shin, J. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. *Advances in neural information processing systems*, 31, 2018.

[25] Li, J., Li, D., Xiong, C., and Hoi, S. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. *arXiv preprint arXiv:2201.12086*, 2022.

[26] Li, Y., Liang, F., Zhao, L., Cui, Y., Ouyang, W., Shao, J., Yu, F., and Yan, J. Supervision exists everywhere: A data efficient contrastive language-image pre-training paradigm. *arXiv preprint arXiv:2110.05208*, 2021.

[27] Loshchilov, I. and Hutter, F. Fixing weight decay regularization in adam. 2018.

[28] Lu, J., Batra, D., Parikh, D., and Lee, S. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *arXiv preprint arXiv: Arxiv-1908.02265*, 2019.

[29] Mu, N., Kirillov, A., Wagner, D., and Xie, S. Slip: Self-supervision meets language-image pre-training. *arXiv preprint arXiv:2112.12750*, 2021.

[30] Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pp. 8748–8763. PMLR, 2021.

[31] Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M., and Sutskever, I. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, pp. 8821–8831. PMLR, 2021.

[32] Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., and Chen, M. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022.

[33] Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al. Imagenet large scale visual recognition challenge (2014). *arXiv preprint arXiv:1409.0575*, 2(3), 2014.

[34] Singh, A., Hu, R., Goswami, V., Couairon, G., Galuba, W., Rohrbach, M., and Kiela, D. Flava: A foundational language and vision alignment model. *arXiv preprint arXiv:2112.04482*, 2021.

[35] Tan, H. and Bansal, M. Lxmert: Learning cross-modality encoder representations from transformers. *arXiv preprint arXiv: Arxiv-1908.07490*, 2019.

[36] Thomee, B., Shamma, D. A., Friedland, G., Elizalde, B., Ni, K., Poland, D., Borth, D., and Li, L.-J. Yfcc100m: The new data in multimedia research. *arXiv preprint arXiv: Arxiv-1503.01817*, 2015.

[37] Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., and Jégou, H. Training data-efficient image transformers & distillation through attention. In *International Conference on Machine Learning*, pp. 10347–10357. PMLR, 2021.

[38] Van Den Oord, A., Vinyals, O., et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017.

[39] van den Oord, A., Li, Y., and Vinyals, O. Representation learning with contrastive predictive coding. *ArXiv*, abs/1807.03748, 2018.

[40] Van der Maaten, L. and Hinton, G. Visualizing data using t-sne. *Journal of machine learning research*, 9 (11), 2008.

[41] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[42] Wu, Z., Xiong, Y., Yu, S. X., and Lin, D. Unsupervised feature learning via non-parametric instance discrimination. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3733–3742, 2018.

[43] Yang, J. and Zhang, Y. Ncrf++: An open-source neural sequence labeling toolkit. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, 2018. URL `http://aclweb.org/anthology/P18-4013`.

[44] Yu, J., Wang, Z., Vasudevan, V., Yeung, L., Seyedhosseini, M., and Wu, Y. Coca: Contrastive captioners are image-text foundation models. *arXiv preprint arXiv: Arxiv-2205.01917*, 2022.

A. Additional Experiments and Analysis

Out-of-distribution detection. Some prior work demonstrated self-supervised learning approaches significantly improve OOD detection performance [19, 20, 13], where their self-supervised pre-training heavily relies on domain-specific data augmentations. We expect MAE to perform well on OOD benchmarks and want to study how M3AE performs compared with MAE.

We consider the difficult near-OOD as this is a more challenging and realistic problem; many methods can achieve high AUROC on the easier far-OOD benchmarks, but do not perform as well in near-OOD tasks. The results are shown in Figure 3, M3AE outperforms MAE in terms of both Mahalanobis outlier score [24] and max over softmax score [18].

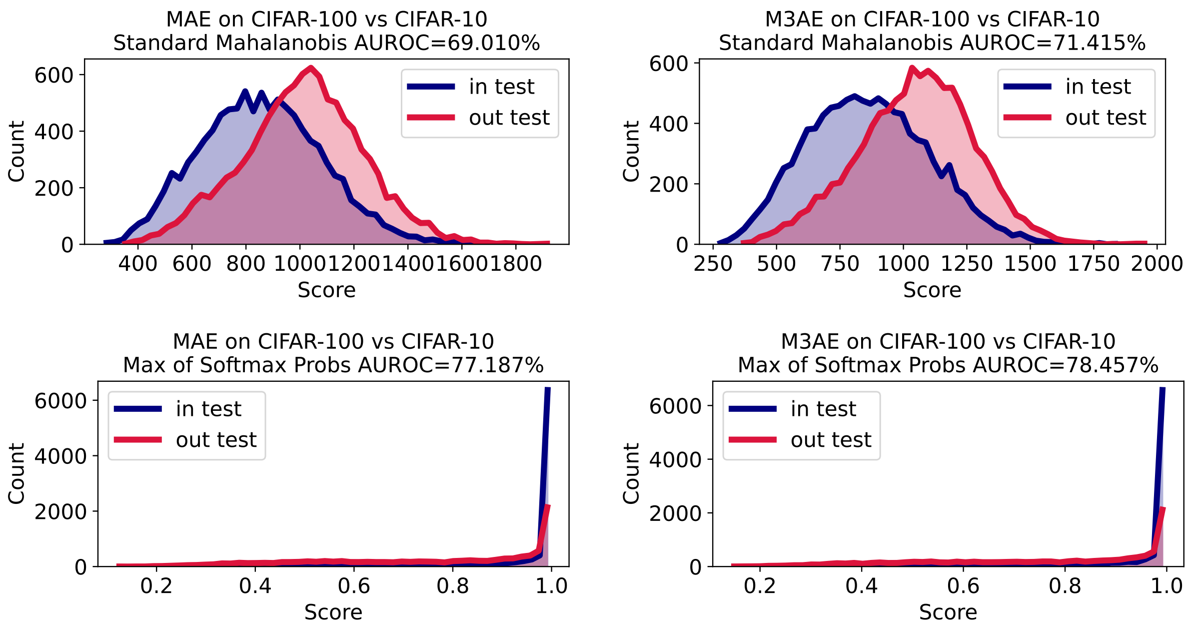


Figure 3. Out-of-distribution detection results on CIFAR-100 (in-distribution) and CIFAR-10 (out-of-distribution). **Upper** shows results based on Mahalanobis outlier score, M3AE achieves *71.4%* which is higher than MAE's *69.0%*. **Lower** shows results based on max over softmax score, M3AE achieves *78.5%* which is also higher than MAE's *77.2%*.

Visualization of cross-modal attention weights. We are interested in what M3AE captures in multimodal attention weights. To do so, we visualize the M3AEencoder attention between a given text token and all image patches, as well as the attention between a given image patch and all text tokens [43] in Figure 5 and Figure 4. M3AE learns to attend relevant concepts in both image and text, showing that our model is able to infer relations between visual and language concepts.

Figure 4. Visualization of attention between a given image patch and all text tokens on CC12M dataset The highlighted rectangle is the image patch for which we visualize the attention. Denser color of the text denotes higher attention. The visualization suggests that M3AE encoder is able to attend to the correct words corresponding to the image patch.

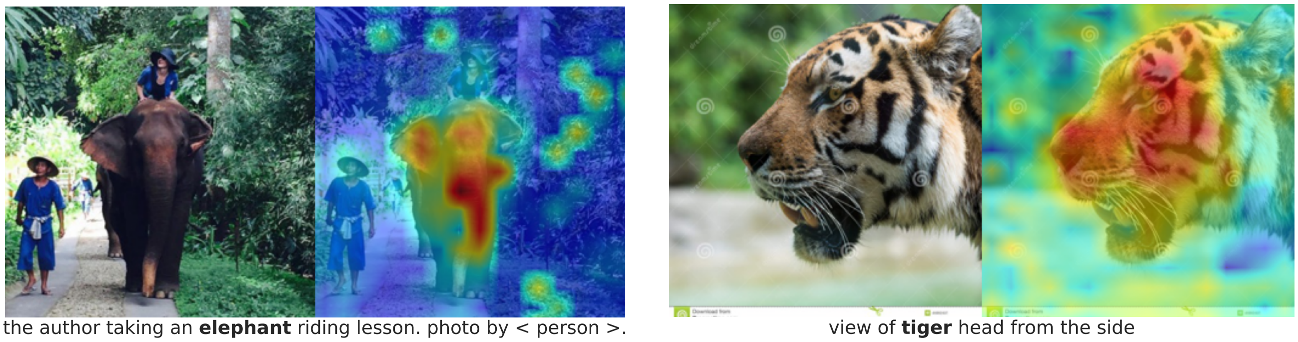


Figure 5. Visualization of attention between a given text token and image patches on CC12M dataset. The text token for which we visualize the attention is bolded. We see that the M3AE encoder is able to attend to the correct objects.

Clustering analysis of representation. We perform t-SNE [40] visualizations of the learned representation of M3AE and MAE for 10 classes on ImageNet validation set in Figure 6. Compared to MAE, M3AE successfully clusters together images that correspond to the same semantic label.

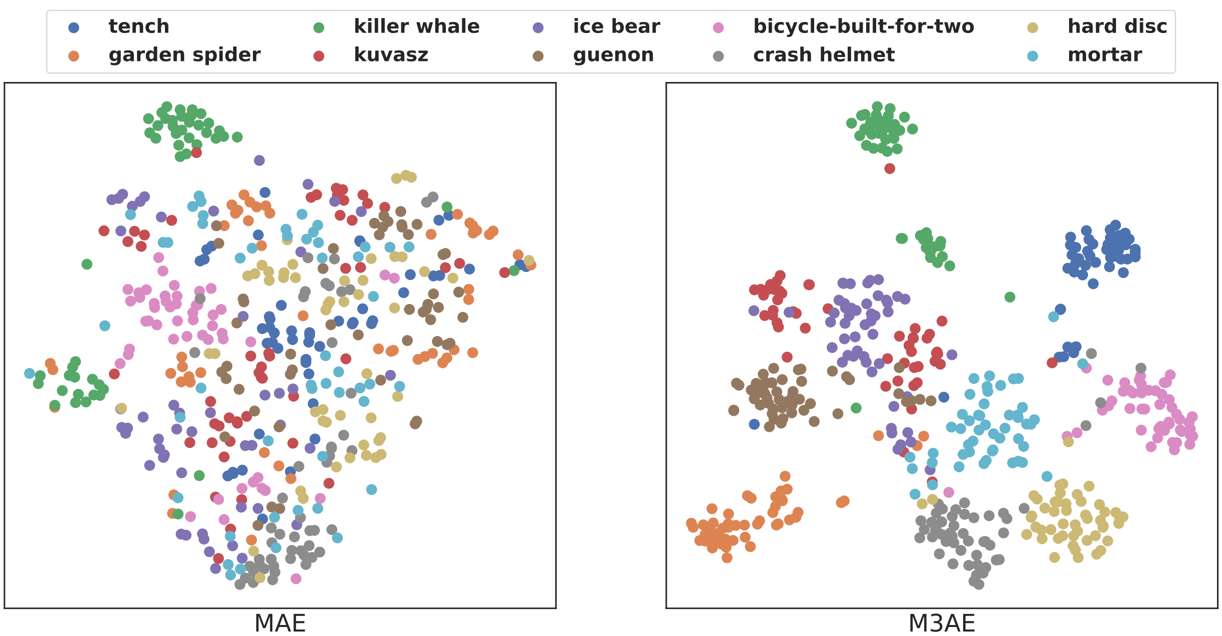


Figure 6. t-SNE visualization for learned representations of 10 classes on ImageNet validation set. Left is MAE and right is M3AE. The representation of M3AE clusters much stronger together with the semantic labels compared to MAE representations.

Reconstruction visualization. We are interested in the reconstruction quality of pretrained M3AE. We randomly sample examples from CC12M and the validation set of ImageNet and show the results in Figure 7. In each reconstructed image, we include original unmasked tokens for better visual quality. We observe that our model infers holistic reconstructions across CC12M and ImageNet datasets, indicating it has learned numerous concepts.

B. Related work

Self-supervised representation learning via reconstruction After the introduction of Transformers [41], self-supervised language modeling has made substantial progress in recent years. After pre-training on a large amount of unlabeled data with reconstruction loss, Large-scale Transformer language models like BERT [10] and GPT [3] are highly successful in learning representations that generalize well to various downstream tasks. Taking inspiration from the success in NLP, research

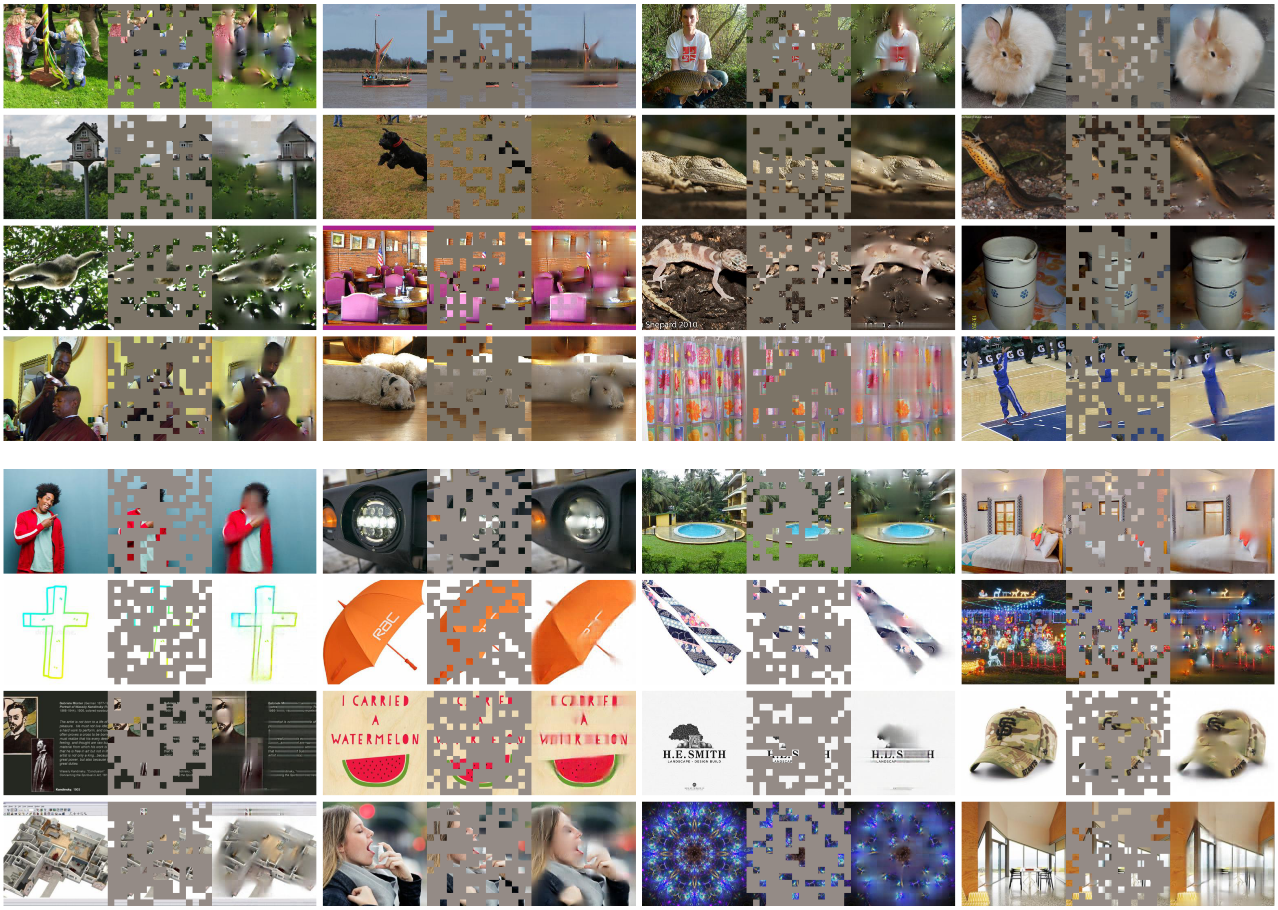


Figure 7. Masked image reconstruction on ImageNet validation images (top) and CC12M (bottom). For each triplet, we show the ground-truth (left), the masked image (mid) and our M3AE reconstruction (right).

have proposed a wide variety of self-supervision method [5, 12, 2, 17]. iGPT [5] that operates on sequences of pixels and reconstruct the unknown pixels. ViT [12] studies masked patch prediction for self-supervised learning. BEiT [2] proposes to predict discrete tokens [38, 31]. MAE [17] proposes to randomly mask patches of the input image and reconstruct the missing pixels. Heavily inspired by MAE and BERT, our M3AE brings together image and language data and learns a shared representation for both modalities by applying a unified masked patch and token prediction objective.

Self-supervised representation learning via contrastive objectives Besides reconstruction, another major paradigm for self-supervised learning is contrastive learning, which models similarity and dissimilarity between two or more views of images or texts [14, 7, 16, 39, 15, 42]. SimCSE [14] proposes constructing positive sentence pair through applying Dropout. SimCLR [6] studies applying random image augmentation for contrastive learning. Contrastive learning often rely heavily on data augmentation and can therefore introduce bias during training. Our M3AE does not rely on contrastive objectives so it can be applied without data augmentation.

Joint learning for language and image Learning representations for a single modality has high importance as it extracts semantic formation from the raw data of modality. Learning a joint representation for several modalities is challenging since it requires alignment between semantic information from different modalities, of which the information contained may vary drastically. Specifically, learning joint representation for vision and language has been a long standing problem in artificial intelligence. Recently, CLIP [30] successfully tackled this challenge by leveraging contrastive learning over a large dataset of aligned text-image pairs. Several works followed this idea, further improving the joint representation. BLIP [25] used noisy web data by bootstrapping the captions with synthetic ones. SLIP [29] learned a joint representation by combining CLIP [30] and SimCLR [6] techniques and leveraging both a paired dataset, and a much larger image-only dataset. DeCLIP [26] utilized more image-text pairs collected from CLIP [30] by adding multiple self-supervised techniques. Inspired by BERT, other methods study cross-modal matching loss [8, 28, 34, 35, 44]. FLAVA [34] employs both contrastive and multimodal training objectives on paired and image-only datasets. Perceiver [21] proposes cross-attention to combining language and image modalities. CoCa [44] combines cross-modal contrastive learning and autoregressive caption prediction. Our M3AE models provides a simple but effective alternative for learning joint representations by processing language tokens and image patches through a shared encoder-decoder architecture. We train our model with masked token reconstruction loss, eliminating the need to handle each modality separately.

C. Implementation Details

C.1. Pre-training datasets

Conceptual 12M (CC12M)[1] [4] contains approximately 12M of image-text pairs, the original dataset images are provided in the form of internet URLs. Note that due to some expired URLs and non-English captions, we did not obtain the complete data in the dataset.

C.2. Downstream datasets

We evaluate the image encoder transferability on ImageNet [33]. We report top-1 validation accuracy of a single 256×256 crop. We evaluate evaluate out-of-distribution detection on CIFAR-100 and CIFAR-10 datasets [23]. Table 3 provides the detailed information of these datasets.

DATASET	Classes	Train size	Test size	Evaluation metric
CIFAR10	10	50,000	10,000	Accuracy
CIFAR100	100	50,000	10,000	Accuracy
ImageNet	1000	$1,281,167$	50,000	Accuracy

Table 3. Details of downstream datasets

C.3. Network architectures

Following MAE, we use ViT [12] as the model architecture and consider three different sizes of ViT for the M3AE image and text encoder. The model consists of a stack of standard Transformer blocks [41], and each Transformer block consists of a multi-head self-attention and an MLP. We use the original ViT-B/16 and ViT-L/16 architectures [12] for our encoder, as well as ViT-S/16 [37] which is comparable to ResNet-50 in FLOPs and parameters. Following MAE [17], our decoder is lightweight and has 8 blocks and a width of 512. As in MAE, since our encoder and decoder have different width, we adopt a linear projection layer after the encoder to match the dimension. For linear probing, we use the auxiliary CLS token for training the classifier as done in MAE.

C.4. Pre-training hyperparameters

For the pre-training of M3AE and MAE, we follow the hyperparameters of the original MAE. We keep the optimizer, learning rate, weight decay the same as the original MAE on ImageNet. The only additional hyperparameters unique to M3AE are text token mask ratio and text token classification loss weight. We provide all the hyperparameters in Table 4, where the same hyperparameters are used to train network of all sizes and epochs. The base learning rate corresponds to the learning rate of 256 batch size, and it is linearly proportionally scaled according to the actual batch size.

C.5. Downstream evaluation hyperparameters

For downstream tasks of linear classification on ImageNet and OOD detection on CIFAR, we use the same hyperparameters for M3AE and MAE. We list the hyperparameters for ImageNet 1K linear classification in Table 5, and OOD detection for CIFAR in Table 6

[1] `https://github.com/google-research-datasets/conceptual-12m`

Hyperparameter	M3AE	MAE
Optimizer	AdamW	
Base learning rate	1.5e-4	
Weight decay	0.05	
Optimizer momentum	$\beta_1 = 0.9, \beta_2 = 0.95$	
Batch size	4096	
Learning rate schedule	cosine decay	
Warmup epochs	5	
Image data augmentation	RandomResizedCrop	
Image patch mask ratio	0.75	
Text token mask ratio	0.75	N/A
Text token cross entropy loss weight	0.5	N/A

Table 4. Hyperparameters for pre-training M3AE and MAE on CC12M

Hyperparameter	M3AE and MAE
Optimizer	LARS
Base learning rate	0.1
Weight decay	0
Optimizer momentum	0.9
Batch size	2048
Learning rate schedule	cosine decay
Epochs	90
Warmup epochs	10
Image data augmentation	RandomResizedCrop

Table 5. Hyperparameters for linear classification on ImageNet 1K

Hyperparameter	M3AE and MAE
Optimizer	AdamW
Base learning rate	0.001
Weight decay	0.05
Optimizer momentum	$\beta_1 = 0.9, \beta_2 = 0.999$
Batch size	1024
Learning rate schedule	cosine decay
Epochs	100
Warmup epochs	10
Image data augmentation	RandAugment

Table 6. Hyperparameters for fine tuning on CIFAR10.

C.6. Computation Resources

All the experiments are performed on the Google Cloud TPU platform. We implement our model using JAX and parallelize the large batch training across many TPUs with data parallelism. For all the pre-training, we use batch size 4096. We report the total amount of compute and the type of resources used in Table 7.

Model	ViT-S	ViT-B	ViT-L
MAE	16.5h (v3-64)	8.5h (v3-128)	11.5h (v3-128)
M3AE	9.5h (v3-128)	5h (v3-256)	10h (v3-256)

Table 7. TPU pod size and compute hours used for training 50 epochs of M3AE and MAE on CC12M.