1 Previous Work

Masked Autoencoders (MAE) is applied in natural language processing (NLP) tasks. Multiple methods have proven its effectiveness, e.g. BERT[2] ang GPT[1]. These methods mask a portion of the sequence and try to train a model to predict the loss components. The ViT[3] paper studies a new method to predict masked parts for self-supervised learning in the computer vision field.

2 Problem Introduction

This paper[4] explores MAE as a self-supervised learning method for computer vision (CV) tasks. One of the primary challenges in image recognition tasks is the high cost and time required to train models on large datasets. To address this issue, the authors propose an asymmetrical MAE architecture that randomly samples a small portion of image patches for training the encoder. Since CV tasks typically exhibit heavy redundancy in semantics compared to NLP tasks, the authors remove 75% of the image patches to reduce redundancy and save computational resources and memory. In the paper, a random sampling approach with a uniform distribution is employed during the pre-training phase to enhance the accuracy of the model on general image datasets before fine-tuning on downstream tasks. The results of the study demonstrate the efficacy of the proposed approach when applied to large datasets in the context of computer vision.

3 Algorithms and Method

It has been observed that the choice of mask sampling strategy has a significant impact on the training performance of the Masked Autoencoder (MAE) model. The authors of the paper suggest that a masking ratio of 75 percent with uniformly distributed random sampling outperforms other patterns in terms of generalization performance. In our project, we intend to explore various mask sampling strategies, such as different masking ratios with different distributions, to investigate their impact on the performance of the MAE model in various downstream tasks.

4 Dataset

We will do self-supervised pre-training on a subset of the ImageNet dataset. However, we have not decided which one we would use. Here we list some potential subset, like ILSVRC2012[5].

5 Evaluation Criteria

Similar to the paper[4], we will compute the mean square loss (MSE) between the original image and the reconstructed image as our loss function. We will only compute MSE on masked pixels, like it was done in BERT[2].

References

[1] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[2] 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.

[3] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[4] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16000–16009, 2022.

[5] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015.