# KEN 4154 Assignment 4

Paper Review

Adrian Sondermann, Abel de Wit

December 8, 2020

## 1 Introduction

For this assignment we are tasked to review a paper on several aspects that are used to review a paper before publication. The paper we chose is called *Generative Pretraining from Pixels*[1] by the authors *Mark Chen, Alec Radford, Rewon Child, Jeff Wu, Heewoo Jun, Prafulla Dhariwal,David Luan, and Ilya Sutskever* and which was published to the *Proceedings of the 37th International Conference on Machine Learning*, Vienna, Austria, 2020.

## 2 Summary

Pre-training used to play a big role in deep learning for computer vision but new techniques in feature encoding and an abundance of labelled data for supervised learning diminished the need for pre-training or unsupervised learning. Pre-training did flourish in Natural Language Processing. With BERT the prediction of corrupted words closely resembled the Denoising Autoencoder that was originally developed for images. Because generative pre-training methods haven't been touched for a decade, the authors deemed it time to investigate how it compares to modern self-supervised methods.

They aim to compare their approach to state of the art self-supervised approaches, which is specifically interesting because their approach does not encode any of the two-dimensional spatial structure of images. The way they encode images is to reduce their resolution and color channels, and then reshaping the pixels into a one-dimensional array that can be seen as a sentence where instead of words, there are pixels. The reduction of their RGB color channels is achieved by using k-means clustering. Then, when they have a sequence of pixels, they use one of two pre-training objectives, auto-regressive next pixel prediction or masked pixel prediction of BERT, to predict pixels in an image instead of language tokens in a sentence. The representations learned by either of the objectives then get evaluated, i.e. with linear probes.

They train their network on the CIFAR-10, CIFAR-100, STL-10, and the well known ImageNet datasets. Applied to these sets, they claim that better

[1] `https://cdn.openai.com/papers/Generative_Pretraining_from_Pixels_V2.pdf`

models with higher log-probability on validation data achieve higher accuracy on linear probes. Moreover, their biggest model *iGPT-L* consisting of $L = 48$ layers, an embedding size of $d = 1536$ with 1.4B parameters was compared to different state-of-the-art models on the mentioned datasets.

Regarding both CIFAR datasets as well as the STL-10 set, a linear classifier fit to the representations of *iGPT-L* outperforms all mentioned end-to-end supervised classifiers. Moreover, applied on ImageNet the model cannot be efficiently trained without adapting it. Thus, after increasing the context length (model resolution) and using a VQ-VAE (Vector Quantized Variational Autoencoder) preprocessing step, their model yields comparable performance to current self-supervised models. Finally they give insight to how they were able to fine-tune *iGPT-L*.

## 3 Relevance

In this paper, Machine Learning approaches from different fields were combined to create models capable of pixel prediction / image generation. It is concerned with generative pre-training methods, which had a substantial impact in Natural Language Processing and their performance compared to state-of-the-art self-supervised approaches. These approaches were re-examined in the paper when applied to higher dimensional and noisier data like images rather than text. As it is interesting to an AI audience, whether approaches perform well in different, initially unforeseen circumstances and how they can be improved further to adapt to different kinds of inputs, this paper gives interesting insight to an audience concerned with image generation. Moreover, some practical results are provided, when applied to different datasets.

## 4 Significance

Some years ago, generative pre-training algorithms were a popular method regarding images in machine learning problems. But as new solutions were introduced and old ones evolved, the dominant approaches for image generation today are mostly self-supervised methods. In addition pre-training methods have substantial impacts only in NLP.

Therefore the authors decided this class of methods was due for a modern re-evaluation and comparison with the recent progress. They claim that generative pre-training learns representations, which significantly aid low-resolution unsupervised learning settings and is competitive with current self-supervised approaches. All these results are supported by experiments with the autoregressive formulation and the BERT formulation of the architecture. Both formulations were used on four different datasets, CIFAR-10, CIFAR-100, STL-10 and ImageNet. Within the first three datasets 10% were split off for validation, while for the latter ImageNet 4% were split off as experimental validation set.

# 5 Novelty

The problem of image generation is something that researchers have been trying to improve ever since the rise of Machine Learning and image generation with it. Over the last years this field has gotten better and better and with the introduction of General Adversarial Networks (GANs) it has gotten even better. This also means that making advancements and improvements on the newest techniques gets harder and harder. However, the approaches that are taken by the authors of the paper can be considered quite novel.

Conventional image generation works with supervised learning, and all of these architectures also encode the two-dimensional structure of the image. The authors approach the generation of images with techniques that stem from a completely different field, Natural Language Processing. By changing the representations from words and sentences to pixels and images, this research is definitely a new take on image generation. This novel approach to a classical problem seems very promising, although it is still computationally expensive as the memory requirements of the transformer decoder scale quadratically with context length.

# 6 Soundness

The hypothesis the researchers want to answer is whether a model that was initially modelled for text generation could be trained to auto-regressively predict pixels in an image. The conclusion that their model learns a strong image representation is based on their use of linear probing, which shows that the model learns features that are good enough to classify the image with simple linear regression. The researchers explore differences between image resolutions that are used to train the model and show how accuracy improves with higher image quality, but performance and amount of parameters to be learned increases. The claims that predicting pixels learns state of the art representations for low resolution datasets and that in high resolution settings, their approach is also competitive with other self-supervised results, are well based in the wide range of tested input variability, differences in model architectures, and the improvements that are made with fine-tuning.

# 7 Evaluation

The authors took a very experimental approach. In the beginning they shortly introduce both later used objectives, the negative log-likelihood for auto-regression and the negative log-likelihood of "masked" elements $x_M$ conditioned on the "unmasked" ones $x_{[1,n]/M}$. Afterwards they continue with detailed experiments and their corresponding results.

Firstly the layers with the highest representational quality of the data are determined, which in contrast to supervised methods lie in the middle of the network, before the negative correlation between validation loss and representation quality is determined. These claims are well-supported by extensive explanations and descriptive figures, which indicate the claimed results. Furthermore, the claims about the results regarding the autoregressive formulation are supported by linear probes on the different datasets and fully fine-tuned architectures. Similarly extensive evaluations are provided for the use of BERT in combination with the *iGPT-L* model. Again, the importance of fine-tuning the models is illustrated within diagrams.

Overall, the authors provide many different experiments and comparisons with other methods to conclude that the claims about the approach being competitive are well-supported. Thus, these results can be replicated, as experimental details regarding hyperparameters and further details are given in the appendices. But they are aware of the possibilities for improvement like increasing the image resolution or lowering the models parameter count.

# 8 Clarity

After a short abstract the paper is divided into six sections. It starts with an introduction to the topic of pre-training and some of its use cases since the mid 2000's before the authors formulate their ideas for pre-training on images as well as the architecture. In the following their methodology is described. Here they explain, which datasets are used for experiments, which parameters different models have and how they are trained later on before all experiments are explained as well as extensively evaluated. Finally, relations to other developments regarding generative models are presented, as many new approaches came up over the last years before drawing a comprehensible conclusion with possible future research tasks.

To summarize, the paper is structured logically and built upon many references to other up-to-date results. Therefore, the authors refer to over 70 other publications. Thus, the paper is well-organized and understandable to readers familiar with the most recent developments. Moreover the division into different sections is plausible and all steps to obtain the final results are clearly indicated.

# 9 Detailed Comments

Overall we were positive of this paper and the way the researchers presented their methodology and results. The research takes a new approach to image generation, and while it might not seem like predicting pixels in a sequence would make sense in a two-dimensional sense, but just like text generation there has to be coherence between every word but also every sentence, even some sentences back. We consider this paper of a high quality and hence don't have much remarks on the paper. Some things could have been elaborated on such as the amount of clusters chosen or the exact definition of a *'GPT-2 Scale Model'*. The results of the paper are reproducible as their code is completely open sourced, although you would need 1024 TPU cores to actually reproduce the results the researchers achieved.

## 9.1 Questions for the authors

- Do you believe that the performance limitations such as the quadratic scaling of the transformer decoder can be reduced in the future?
- How could the model be adapted in order to predict pixels in an image of higher resolution while maintaining a number of parameters comparable to similar performing models?
- Why did you choose k = 512 for k-means to cluster the (R, G, B) pixel values? You just presented that value, but did not motivate why it needs to be fixed for different models and how the performance relates to it.