Pixel-Level BPE for Auto-Regressive Image Generation

Anton Razzhigaev
Skoltech, AIRI

Anton Voronov
AIRI

Andrey Kaznacheev
MTS AI

Andrey Kusnetsov
AIRI

Denis Dimitrov
AIRI

Alexander Panchenko
Skoltech

Abstract

Pixel-level autoregression with Transformer models (Image GPT or iGPT) is one of the recent approaches to image generation that has not received massive attention and elaboration due to quadratic complexity of attention as it imposes huge memory requirements and thus restricts the resolution of the generated images. In this paper, we propose to tackle this problem by adopting Byte-Pair-Encoding (BPE) originally proposed for text processing to the image domain to drastically reduce the length of the modeled sequence. The obtained results demonstrate that it is possible to decrease the amount of computation required to generate images pixel-by-pixel while preserving their quality and the expressiveness of the features extracted from the model. Our results show that there is room for improvement for iGPT-like models with more thorough research on the way to the optimal sequence encoding techniques for images.

1 Introduction

Modern deep learning includes a broad scope of problems with varying difficulty. To solve these tasks a paradigm of pre-training is widely used in some domains, to the greatest extent in computer vision (CV) and natural language processing (NLP). Whilst unsupervised or self-supervised pre-training is more dominant in the NLP domain, CV models are mainly trained using large amounts of labeled data. Authors of iGPT (Chen et al., 2020) have attempted to prove that given appropriate conditions (namely flexible architecture and significant amount of computation) it is possible to pre-train a model that will reach state-of-the-art performance on several CV downstream tasks even with unlabeled data. They have achieved it using an autoregressive pixel-level image generation as an unsupervised training objective for training a Transformer (Vaswani et al., 2017) model.

The approach of pixel-by-pixel generation exploited in the iGPT paper simply models an image as a continuous sequence of pixels and models the probability distribution of the next pixel conditioned on all previous ones. Flattening images results in sequences of an enormous length, for example, such representation of a 128x128 RGB image will require 49152 tokens, which is infeasible for RNNs as well as for Transformer models where complexity is quadratic with respect to the sequence length.

Despite there being numerous ways of optimizing attention operation in Transformer authors of the iGPT model have deliberately chosen dense attention due to it being domain agnostic and not imposing any additional biases on the data. In our work, we continue research in this direction concentrating on the optimization of the image-to-sequence representation mechanism rather than the attention mechanism or the Transformer architecture itself.

In the presented paper we try to adopt a tokenization approach widely used in the NLP domain: Byte-Pair Encoding (BPE) to the image domain to mitigate the main issue of the original iGPT paper. These methods allow to significantly squeeze input sequences thus reducing the amount of computation required for training and inference. Following the methodology of the original paper, we also test the ability of the Transformer model pre-trained on image generation to be used as a feature extractor that competitively performs on downstream tasks, namely, image classification on CIFAR datasets [1].

The main contributions of this paper are as follows:

- We propose a novel method of image-to-sequence tokenization that allows pre-train image models on a generative objective with lower computational complexity.

[1]The code is available at https://github.com/razzant/bpe-iGPT

Proceedings of the 29th International Conference on Computational Linguistics, pages 26–30
October 12–17, 2022.

- We study the dependence between the size of BPE vocabulary and the amount of computation required for a forward pass.
- We show that pre-training with image-BPE increases the capacity of the model allowing it to learn more meaningful representations.
- We conduct several experiments measuring the model's performance on downstream tasks.

2 Related Work

Autoregressive approaches have proven to be very efficient in the NLP domain both in a pre-training and a variety of natural-language generation tasks (Radford et al. (2019), Raffel et al. (2020)). However, in the CV domain, it has been quite a challenge due to the high dimensionality of the data. One of the effective ways to tractably model a joint distribution of pixels in an image is to cast it as a product of conditional distributions. It was adopted in several models such as fully visible sigmoid belief networks (Neal, 1992) or NADE (Larochelle and Murray, 2011).

Recurrent Neural Networks (RNN) are powerful models that offer a compact, shared parametrization of a series of conditional distributions. Authors of PixelRNN (van den Oord et al., 2016) have applied this architecture to an image domain. The authors suggested two types of convolutional LSTM layers to compute all the states along one of the spatial dimensions (rows or diagonals of the image). Moreover, instead of LSTM blocks a convolutional layer with a mask to avoid seeing the future context was used. This method was called PixelCNN and got further development such as PixelCNN++ (Salimans et al., 2017). A small receptive field was an obvious disadvantage of these approaches that was overcome with the emergence of Transformers.

Transformer-based (Vaswani et al., 2017) models are extremely successful in natural language generation and understanding fields. GPT-2 (Radford et al., 2019) demonstrated human-level performance in text generating and zero-shot tasks via prompt engineering. There were numerous attempts to use GPT architecture for image generation, which can be divided into two groups: discrete feature-based regression (e.g. DALLE Ramesh et al. (2022)) or pixel-level regression (iGPT Chen et al. (2020)). The latter type of model is not fairly popular, as processing the 1D-sequence of flattened RGB-image pixels is too memory-expensive due to the length of the context and attention mechanism. To deal with this problem authors resize images to a low resolution (like $32^2 \times 3$, $48^2 \times 3$, $96^2 \times 3$ or $192^2 \times 3$) with further clustering (R, G, B) pixel values using k-means with $k = 512$ obtaining the resulting context length 32^2 or 48^2. However, the iGPT model demonstrated decent results in low-resolution image generation and downstream tasks over contextualized features. To measure model performance linear probe method was used. The method consists of training multi-class logistic regression on embeddings from a model with frozen weights on an image classification task. During pre-training on ImageNet authors also used VQVAE as a downsampler instead of RGB-clustering to keep the context of 48^2 length.

On the other side, there are numerous methods for sequence length compression in the NLP domain — different tokenization techniques, which exploit the pre-computed merge dictionaries for optimal encoding of words or byte groups. One of the most efficient methods is Byte-Pair-Encoding (Shibata et al., 1999). The idea of this algorithm is to find the most frequent pair of consecutive two-character codes in the text and then substitute an unused code for the occurrences of the pair. This method has become a good trade-off between vocabulary size and the length of the sequence fed to the model. In GPT models special modification of this algorithm is used which works at byte-level (Wang et al., 2020) — this is one more step towards optimal sequence squeezing.

3 iGPT with BPE Image Tokenization

Our BPE-enabled iGPT model relies on the GPT-2 model originally designed for text processing. More specifically we use embedding size $d = 1024$, number of layers $L = 36$ and number of heads in the multi-head attention $m = 8$ resulting in 484 million trainable parameters throughout all experiments. Due to limited computational resources, we have not conducted experiments with the BERT pre-training objective and used only linear probing as an evaluation approach.

In our experiments, we provide results for prompted image generation and linear probe on CIFAR10 and CIFAR100 datasets with pre-training on ImageNet (Deng et al., 2009) dataset. Also we demonstrate unconditional image generation on CelebA dataset (Liu et al., 2015) aligned with

MTCNN framework (Zhang et al., 2016).

3.1 Converting Images to Texts

To train byte-level BPE tokenizer we convert images to text format by assigning each pixel value a corresponding char symbol separating each row of the original image with \n symbol in the resulting text file. Since every pixel has an assigned value from 0 to 255 we can quantize them into 10 discrete buckets using integer division by 26. Now since every pixel has a value from 0 to 9 for the grey-scale setting we can replace each number with the corresponding digit character. However, for RGB images we need to represent values from all three channels in one symbol, that is why we concatenate their values resulting in one number in the range from 0 to 999, and convert this number into a character using the standard chr function.

For example RGB pixel [150, 112, 255] will be converted to a char in the following way:

1. RGB pixel: [150, 112, 255]
2. Quantization: [150, 112, 255] // 26 = [5, 4, 9]
3. Concatenation: [5, 4, 9] $\rightarrow$ 549
4. To char: chr(549) = ζ

3.2 Decoding Images from Tokens

Since an output of the model can have lines of various lengths we bring them to the required fixed resolution by either upsampling or downsampling. Then in the case of grey-scale images, each character is directly translated to the corresponding quantized pixel value while for the RGB scenario we use the python ord function, inverse to the chr method used during encoding.

3.3 Encoding Efficiency

To evaluate the sequence squeezing effect of BPE for images we calculate the squeezing factor — an average ratio of the pixel-sequence length of an image to the length of tokenized pixel sequence. It can be seen from Figure 1 that the squeezing factor grows logarithmically with the size of the BPE vocabulary.

While larger vocabularies produce shorter input sequences they also increase the number of trainable parameters and the size of modeling distribution thus hindering the generation. Figure 2 shows that the vocabulary size of 30 000 tokens gives an optimal trade-off between the input squeezing and computational efficiency of the model. The selected vocabulary allows us to reduce the length of pixel sequences roughly by 9 times for grey-scale images and by 4 times for RGB images, i.e. the 112x112 image can be represented by a sequence of approximately a thousand tokens.

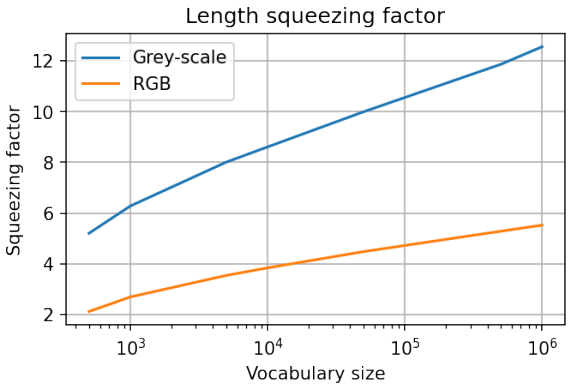

Figure 1: **Compression ratio.** The dependency of sequence squeezing factor from BPE vocabulary size for RGB and grey-scale 112x112 images. The more tokens contains BPE dictionary the shorter the sequences used to represent an image.

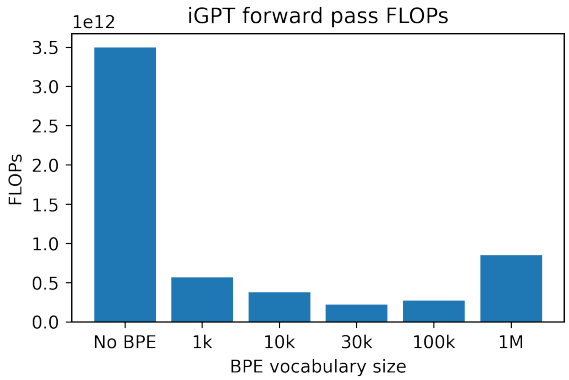

Figure 2: **Computational efficiency evaluation**. iGPT forward pass FLOPs for a 64x64 RGB image with different tokenization strategies: BPE and original pixel-level (No BPE).

4 Experiments

4.1 Examples of Generated Images

Faces generated by our BPE-iGPT model in 112x112 resolution are presented in the Figure 3 . It is worth noting that the authors of the original iGPT provided only examples generated by their largest model iGPT-XL (6.4 billion parameters) in 32x32 resolution, however visual fidelity of our samples remains on the same level. This supports our statement that image-BPE tokenization

allows for pre-train Transformer models on the data of higher dimensionality with less computational overhead.

Figure 3: RGB generated faces 112×112.

We have also tested the ability of our model to image-conditional generation. We show examples of image completion in Figure 4. Even though we have not used any advanced sampling techniques such as nucleus sampling, tuning for the temperature, or beam-search all of the generated images contain clearly recognizable objects.

Figure 4: Image completions (64×64). Top row: prompt fed to model, middle: the result of the generation, bottom: ground truth image.

4.2 Image Representations for Downstream Tasks

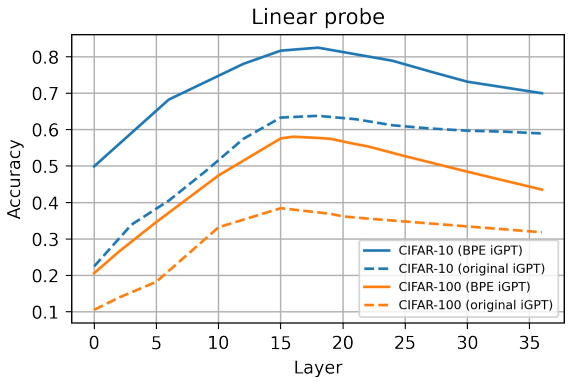

Figure 5: **Effectiveness evaluation.** Linear probe evaluation on CIFAR-10 for features extracted from every layer.

One of the common means to evaluate the representations learned by pre-trained models is linear probing on downstream tasks. To do so we train a logistic regression model over the features extracted from the trained network and compare the classification accuracy of the model pre-trained using image BPE against raw pixel sequences. Following the approach presented in the iGPT paper, we evaluate features extracted from every layer of the network.

Figure 5 shows the results of classification on CIFAR-10 and CIFAR-100 datasets. As can be seen from the plot our findings are in the agreement with the original paper: the best layers to be used as feature extractors are situated around the central layer. Another interesting finding is that even the first layer of the model trained on BPE-image contains representative features in contrast to the model trained on pixel sequence where first results better than random are obtained after several layers. One of the possible explanations for this is that some BPE-tokens represent the most common sequences of pixels which means that they already contain some semantic information in contrast to raw pixel sequences.

Our finding is in the accordance with similar research in the NLP domain. Authors of (Kharitonov et al., 2021) show that the ability of Transformer models to memorize training data is highly dependent on the size of BPE vocabulary. In combination with our results, this suggests that BPE tokenization increases the capacity of models allowing them to learn more information about the data from every layer.

5 Conclusion

In this paper, we explored the use of the BPE technique originally proposed for textual data in the image domain. It allows significantly squeeze the tokenized image sequence length mitigating the limitations of the original iGPT model. We quantitatively show that this method reduces the amount of required computation by an order of magnitude and qualitatively verify that it does not affect the quality of generated images. Moreover, applying BPE tokenization improves the representative ability of the models trained on unlabeled data. Our results suggest that the potential of image-to-sequence squeezing is not fully unleashed yet and that there is room for improvement of iGPT-like models.

References

Mark Chen, Alec Radford, Rewon Child, Jeffrey Wu, Heewoo Jun, David Luan, and Ilya Sutskever. 2020. Generative pretraining from pixels. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 1691–1703. PMLR.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009), 20-25 June 2009, Miami, Florida, USA*, pages 248–255. IEEE Computer Society.

Eugene Kharitonov, Marco Baroni, and Dieuwke Hupkes. 2021. How BPE affects memorization in transformers. *CoRR*, abs/2110.02782.

Hugo Larochelle and Iain Murray. 2011. The neural autoregressive distribution estimator. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics, AISTATS 2011, Fort Lauderdale, USA, April 11-13, 2011*, volume 15 of *JMLR Proceedings*, pages 29–37. JMLR.org.

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. 2015. Deep learning face attributes in the wild. In *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, pages 3730–3738. IEEE Computer Society.

Radford M. Neal. 1992. Connectionist learning of belief networks. *Artif. Intell.*, 56(1):71–113.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21:140:1–140:67.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. 2022. Hierarchical text-conditional image generation with CLIP latents. *CoRR*, abs/2204.06125.

Tim Salimans, Andrej Karpathy, Xi Chen, and Diederik P. Kingma. 2017. Pixelcnn++: Improving the pixelcnn with discretized logistic mixture likelihood and other modifications. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.

Yusuke Shibata, Takuya Kida, Shuichi Fukamachi, Masayuki Takeda, Ayumi Shinohara, and Takeshi Shinohara. 1999. Byte pair encoding: A text compression scheme that accelerates pattern matching.

Aäron van den Oord, Nal Kalchbrenner, and Koray Kavukcuoglu. 2016. Pixel recurrent neural networks. In *Proceedings of the 33nd International Conference on Machine Learning, ICML 2016, New York City, NY, USA, June 19-24, 2016*, volume 48 of *JMLR Workshop and Conference Proceedings*, pages 1747–1756. JMLR.org.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.

Changhan Wang, Kyunghyun Cho, and Jiatao Gu. 2020. Neural machine translation with byte-level subwords. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 9154–9160. AAAI Press.

Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. 2016. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Process. Lett.*, 23(10):1499–1503.