Text-Conditional Image Generation from Discrete Representations

Aleksandar Shtedritski, supervised by Christian Rupprecht
Department of Engineering Science, University of Oxford
suny@robots.ox.ac.uk

EPSRC Centre for Doctoral Training in Autonomous Intelligent Machines & Systems

Engineering and Physical Sciences Research Council

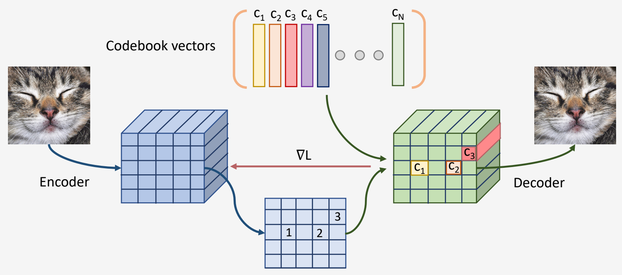
UNIVERSITY OF OXFORD

Text-to-image generation

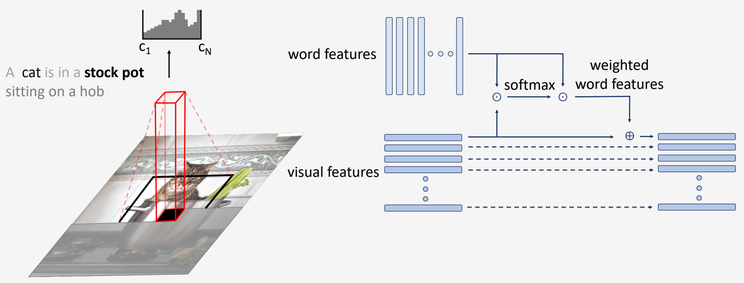
Generating images from text is an important problem with applications in art and media content generation. Traditionally, text-to-image generation has been addressed using GANs, but recently autoregressive models based on transformers such as DALL-E and CogView have shown superior performance. These recent advances draw improvements from the increased model and dataset size. In this work we apply some of the principles that made that possible in a smaller scale and develop autoregressive models for text-to-image generation trainable on a single GPU. In particular, we develop a module that allows us to conditionally generate images from text using a PixelCNN architecture.

Method

We learn an autoencoder that uses discrete rather than continuous representations, following VQ-VAE:



We use a PixelCNN to conditionally generate an image from text, learning a prior over the discrete codebook vectors. We improve on vanilla PixelCNN by introducing cross-modal attention, where each spatial location attends to individual words in the caption. This results in better performance than using a global sentence feature as in prior works, as we are able to capture finer grained details from the caption.



Results

Our method successfully manages to generate images containing all objects in the caption when trained on the complex ClipArt dataset, whereas the baseline PixelCNN that uses global sentence features fails to generate the correct objects.
It also generates visually competitive images with state-of-the-art GANs when trained COCO, while having a less complicated architecture and better sample diversity.

When evaluated on FID and R-Precision, our method falls behind SOTA GANs and recent large autoregressive models. Note however that DALL-E also has lower FID score, while being better visually.

FID score on COCO (↓)

Method	FID
AttnGAN	35.2
DM-GAN	26.0
DALL-E	27.5
CogView	27.1
Ours	40.2

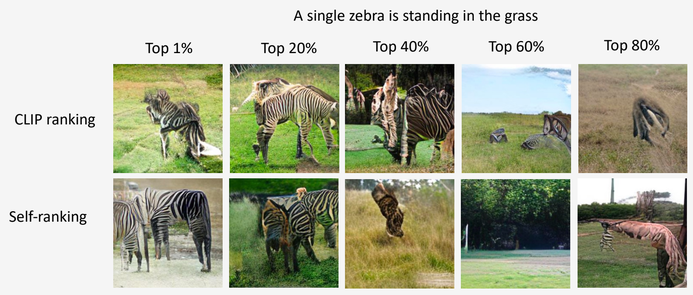
R-Precision score on COCO (↑)

Method	R-Precision
COCO (dataset)	0.76
DM-GAN	0.65
Ours	0.51
Ours w/ seg. masks	0.56

Self-ranking

A drawback of autoregressive text-to-image generation is the need to generate multiple images and rank them, e.g. using CLIP. We use the cross-attention weights to find regions of similarity with words:

We rank the generated images by detecting blobs in the heatmaps and sort by their weighted average area (e.g. using tf-idf):



References

[1] Aditya Ramesh et al. Zero-shot text-to-image generation. *CoRR*, abs/2102.12092, 2021.
[2] Ming Ding et al. Cogview: Mastering text-to-image generation via transformers, 2021.
[3] Minfeng Zhu et al. Dm-gan: Dynamic memory generative adversarial networks for text-to-image synthesis. *CVPR*, 2019.
[4] Tao Xu et al. Attngan: Fine-grained text to image generation with attentional generative adversarial networks. In *CVPR*, 2018.
[5] Aaron van den Oord et al. Neurips. In *Advances in Neural Information Processing Systems*, 2017.