**TAMING TRANSFORMERS FOR HIGH-RESOLUTION IMAGE SYNTHESIS**

**Abstract**

This method is designed to learn long-range interactions on sequential data and continues to show state-of-the-art results on a wide variety of tasks. In contrast to CNNs, it contains no inductive bias that prioritizes local interactions. This makes it expressive, but also computationally infeasible for long sequences, such as high-resolution images. In this paper we demonstrate how combining the effectiveness of the inductive bias of CNNs with the expressivity of transformers enables them to model and thereby synthesize these high-resolution images that we are trying to achieve. We show how to use CNNs to learn a context-rich vocabulary of image constituents, and in turn utilize transformers to efficiently model their composition within these images. The approach trying to be achieved is readily applied to conditional synthesis tasks, where both non-spatial information, such as object classes, and spatial information, such as segmentations, can control the generated image. We present the first results on semantically guided synthesis of megapixel images with transformers and obtain the state of the art among autoregressive models on class-conditional ImageNet.

**Introduction**

Transformers are on the rise. They are now the de-facto standard architecture for language tasks and are increasingly adapted in other areas such as audio and vision. In contrast to the dominant architecture that came before it, convolutional neural networks (CNNs), the transformer architecture contain no built-in inductive prior on the locality of interactions and is therefore free to learn complex relationships among its inputs. However, this generality also implies that it must learn all relationships, whereas CNNs have been designed to exploit prior knowledge about strong local correlations within images. Observations have shown that transformers tend to learn convolutional structures, thus the question arises: Do we have to re-learn everything we know about the local structure and regularity of images from scratch each time we train a vision model, or can we efficiently encode inductive image biases while still retaining the flexibility of transformers? We hypothesize that low-level image structure is well described by a local connectivity, so by convolutional architecture, but this structural assumption ceases to be effective on higher resolutions. Moreover, CNNs exhibit a strong locality bias and a bias towards spatial invariance by using shared weights across all positions. This makes them ineffective for trying to synthesize a more HD image. Our key insight to obtain an effective and expressive model is that, taken together, convolutional and transformer architectures can model the compositional nature of our visual world. We use a convolutional approach to efficiently learn a codebook of context-rich visual parts and, subsequently, learn a model of their global compositions. The long-range interactions within these compositions require an expressive transformer architecture to model distributions over their constituent visual parts. Furthermore, we utilize an adversarial approach to ensure that the dictionary of local parts captures perceptually important local structure to alleviate the need for modeling low-level statistics with the transformer architecture. Allowing transformers to concentrate on their unique strength (modeling long-range relations) enables them to generate high-resolution images, a feat which previously has been out of reach.

**Related Work**

**The Transformer Family**: The defining characteristic of the transformer architecture is that it models interactions between its inputs solely through attention which enables them to faithfully handle interactions between inputs regardless of their relative position to one another. Originally applied to language tasks, inputs to the transformer were given by tokens, but other signals, such as those obtained from audio or images, can be used. Each layer of the transformer then consists of an attention mechanism, which allows for interaction between inputs at different positions, followed by a position-wise fully connected network, which is applied to all positions independently. More specifically, the (self-)attention mechanism can be described by mapping an intermediate representation with three position-wise linear layers into three representations:

Query Q

Key K

Value V

This computes a single out equation:

Attn (Q, K, V) = softmax \*V

When performing autoregressive maximum-likelihood learning, non-causal entries of Q, i.e., all entries below its diagonal, are set to −∞ and the final output of the transformer is given after a linear, point-wise transformation to predict logits of the next sequence element. Since the attention mechanism relies on the computation of inner products between all pairs of elements in the sequence, its computational complexity increases quadratically with the sequence length. While the ability to consider interactions between all elements is the reason transformers efficiently learn long-range interactions, it is also the reason transformers quickly become infeasible, especially on images, where the sequence length itself scales quadratically with the resolution. Different approaches have been proposed to reduce the computational requirements to make transformers feasible for longer sequences. In some other papers, authors have proposed to restrict the receptive fields of the attention modules, which reduces the expressivity and, especially for high-resolution images, introduces assumptions on the independence of pixels. Other papers propose to retain the full receptive field, but this can reduce costs for a sequence of length n only from to , which makes resolutions beyond 64 pixels still prohibitively expensive. There is still need for improvement, and science is trying to make it happen, but for now we must deal with these complexities.

**Convolutional Approaches:** The two-dimensional structure of images suggests that local interactions are particularly important. CNNs exploit this structure by restricting interactions between input variables to a local neighborhood defined by the kernel size of the convolutional kernel. Applying a kernel thus results in costs that scale linearly with the overall sequence length (the number of pixels in the case of images) and quadratically in the kernel size, which, in modern CNN architectures, is often fixed to a small constant such as 3×3. This inductive bias towards local interactions thus leads to efficient computations, but the wide range of specialized layers which are introduced into CNNs to handle different synthesis tasks suggest that this bias is often too restrictive.

**Two-Stage Approaches:** Closest to ours are two-stage approaches which first learn an encoding of data and afterwards learn, in a second stage, a probabilistic model of this encoding. Bin Dai and David P. Wipf. Demonstrated in their paper both theoretical and empirical evidence on the advantages of first learning a data representation with a Variational Autoencoder (VAE), and then again learning its distribution with a VAE. Patrick Esser, Robin Rombach, and Bjorn Ommer. demonstrate similar gains on their paper when using an unconditional normalizing flow for the second stage, and on another paper, they demonstrate using a conditional normalizing flow.