Notes for Bachelor's Thesis, C.S. & the ISA

Contents

I :: Taming Transformers for High-Resolution Image Synthesis[1]

I.1 :: Key Insights

- Taken together, convolutional and transformer architectures can model the compositional nature of our visual world.
- A powerful first stage, which captures as much context as possible in the learned representation, is critical to enable efficient high-resolution image synthesis with transformers.

I.2 :: Approach

- Instead of representing an image with pixels, they represent it as a composition of perceptually rich image constituents from a codebook.
  - By learning a good code, one can significantly reduce the description length of compositions, which allows for effeciently modelled global interrelations within images with a transformer architecture.

I.2.i :: Learning an Effective Codebook of Image Constituents for Use in Transformers

- Constituents of an image needs to be expressed in the form of a *sequence*.
- Instead of building individual pixels, complexity necessitates an approach that uses a discrete codebook of learned representations, such that any image $x \in \mathbb{R}^{H \times W \times 3}$ can be represented by a spatial collection of codebook entries $z_q \in \mathbb{R}^{h \times w \times n_z}$, where n_z is the dimensionality of codes.
  - An e*quivalent representation is a sequence of $h \cdot w$ indices which specify the respective entries in the learned codebook.
- To effectively learn such a discrete spatial codebook, first one can learn a convolutional model consisting of an encoder E and a decoder G, such that taken together, they learn to represent

images with codes from a learned, discrete codebook $Z = \{z_k\}_{k=1}^K \subset \mathbb{R}^{n_z}$ (refer to [1] p. 3 for a diagram).

- One can approximate a given image x by $\hat{x} = G(z_q)$.
- Obtain z_q using the encoding $\hat{z} = E(x) \in \mathbb{R}^{h \times w \times n_z}$ and a subsequent element-wise quantization $\boldsymbol{q}(\cdot)$ of each spatial code $\hat{z}_{ij} \in \mathbb{R}^{n_z}$ onto its closest codebook entry z_k:

$$z_{\boldsymbol{q}} = \boldsymbol{q}(\hat{z}) := \left(\arg\min_{z_k \in Z} \|\hat{z}_{ij} - z_k\| \right) \in \mathbb{R}^{h \times w \times n_z}. \tag{1.}$$

- The reconstruction $\hat{x} \approx x$ is then given by

$$\hat{x} = G(z_{\boldsymbol{q}}) = G(\boldsymbol{q}(E(x))). \tag{2.}$$

- Backpropagation through the non-differentiable quantization operation in Eq. Equation 2 is achieved by a straight-through gradient estimator, which simply copies the gradients from the decoder to the encoder, such that the model and codebook can be trained end-to-end via the loss function

$$\begin{aligned} \mathcal{L}_{\text{VQ}}(E, G, Z) = \|x - \hat{x}\|^2 + \|\text{sg}[E(x)] - z_{\boldsymbol{q}}\|_2^2 \\ + \|\text{sg}[z_{\boldsymbol{q}}] - E(x)\|_2^2. \end{aligned} \tag{3.}$$

- Here, $\mathcal{L}_{\text{rec}} = \|x - \hat{x}\|^2$ is a reconstruction loss, $\text{sg}[\cdot]$ denotes the stop-gradient operation, and $\|\text{sg}[z_{\boldsymbol{q}}] - E(x)\|_2^2$ is the so-called "commitment loss".

I.2.i.i :: Learning a Perceptually Rich Codebook

- Use a VQ-GAN, a discriminator and perceptual loss to keep good perceptual quality at increased compression rate.
- Apply a single attention layer on the lowest resolution to aggregate context from everywhere.

I.2.ii :: Learning the Composition of Images with Transformers

I.2.ii.i :: Latent Transformers

- We can represent images in terms of the codebook-indices with their encodings with E and G available.
- The quantized encoding of an image x is given by $z_{\boldsymbol{q}} = \boldsymbol{q}(E(x)) \in \mathbb{R}^{h \times w \times n_z}$.
  - This is equivalent to a sequence $s \in \{0, ..., |Z| - 1\}^{h \times w}$ of indices from the codebook, which is obtained by replacing each code by its index in the codebook Z:

$$s_{ij} = k \text{ such that } \left(z_{\boldsymbol{q}}\right)_{ij} = z_k. \tag{4.}$$

- By mapping indices of a sequence s back to their corresponding codebook entries, $z_{\boldsymbol{q}} = \left(z_{s_{ij}}\right)$ is readily recovered and decoded to an image $\hat{x} = G(z_{\boldsymbol{q}})$.
- After choosing some ordering of the indices in s, image-generation can be formulated as autoregressive next-index prediction: Given indices $s_{<i}$, the transformer learns to predict the distribution of possible next indices, i.e., $p(s_i | s_{<i})$ to compute the likelihood of the full representation as $p(s) = \prod_i p(s_i | s_{<i})$.
  - This allows for directly maximizing the log-likehood of the data representation:

$$\mathcal{L}_{\text{transformer}} = \mathbb{E}_{x \sim p(x)}[-\log p(s)]. \tag{5.}$$

I.2.ii.ii :: Conditioned Synthesis

- Often a user demands control over the generation process by providing additional information from which an image shall be synthesized.

    - This information, c, could be a single label describing the overall image class or another image itself.
- The task is then to learn the likelihood of the sequence given this information:

$$p(s|c) = \prod_i p(s_i|s_{<i}, c). \tag{6.}$$

- If the conditioning information c has spatial extent, we first learn another VQ-GAN to obtain again an index-based representation $r \in \{0, ..., |Z_c| - 1\}^{h_c \times w_c}$ with the newly obtained codebook Z_c.
    - Due to the autoregressive structure of the transformer, one can then simply prepend r to s and restrict the computation of the negative log-likelihood to entries $p(s_i|s_{<i}, r)$.

I.2.ii.iii :: Generating High-Resolution Images

- The attention mechanism of the transformer architecture puts limits on the sequence length $h \cdot w$ of its inputs s.
- To generate images in the megapixel regime, we have to work patch-wise and crop images to restrict the length of s to a maximally feasible size during training.
- To sample images, then use a transformer in a sliding-window manner (see [1], p. 5).

II :: GFowNet-EM for Learning Compositional Latent Variable Models[2]

III :: Other Notes

III.1 :: Transformers

- Designed to learn long-range interactions on sequential data.
- Expressive but computationally infeasible for long sequences (such as high-resolution images).
    - Quadratic complexity in the sequence length (as all pairwise interactions are taken into account).
- The (self-)attention mechanism can be described by mapping an intermediate representation with three position-wise linear layers into three representations, query $Q \in \mathbb{R}^{N \times d_k}$, key $K \in \mathbb{R}^{N \times d_k}$, and value $V \in \mathbb{R}^{N \times d_v}$, to compute the output as

$$\text{Attn}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right) V \in \mathbb{R}^{N \times d_v}. \tag{7.}$$

III.2 :: Convolutional Neural Networks (CNN)

- Contain certain (inductive) biases:
    - biases that prioritize local interactions;
    - biases towards spatial invariance through the use of shared weights across all positions.
- These biases make them ineffective if a more holistic understanding of the input is required.

III.3 :: Vartional Autoencoders (VAE)

- Can be used to learn a representation of some data.

III.3.i :: Vector Quantized Variational Autoencoder (VQ-VAE)

- An approach to learn discrete representations of images.

III.4 :: Generative Adversarial Networks (GAN)

III.4.i :: Vector Quantized Generative Adversarial Networks (VQ-GAN)

- We replace the L_2 loss for $\mathcal{L}_{\text{rec}}$ by a perceptual loss and introduce an adversarial training procedure with a patch-based discriminator D that aims to differentiate between real and reconstructed images:

$$\mathcal{L}(\{E, G, Z\}, D) = [\log D(x) + \log(1 - D(\hat{x}))]. \tag{8.}$$

- The complete objective for finding optimal compression model $\mathcal{Q}^* = \{E^*, G^*, Z^*\}$ then reads

$$\mathcal{Q}^* = \arg \min_{E,G,Z} \max_{D} \mathbb{E}_{x \sim p(x)} \Big[\mathcal{L}_{\text{VQ}(E,G,Z)} + \lambda \mathcal{L}_{\text{GAN}(\{E,G,Z\},D)} \Big], \tag{9.}$$

  where we compute the adaptive weight λ according to

$$\lambda = \frac{\nabla_{G_L}[\mathcal{L}_{\text{rec}}]}{\nabla_{G_L}[\mathcal{L}_{\text{GAN}}] + \delta} \tag{10.}$$

  where $\mathcal{L}_{\text{rec}}$ is the perceptual reconstruction loss, $\nabla_{G_L}[\cdot]$ denotes the gradient of its input w.r.t. the last layer L of the decoder, and $\delta = 10^{-6}$ is used for numerical stability.

Bibliography

[1] P. Esser, R. Rombach, and B. Ommer, “Taming Transformers for High-Resolution Image Synthesis.” [Online]. Available: https://arxiv.org/abs/2012.09841

[2] E. J. Hu, N. Malkin, M. Jain, K. Everett, A. Graikos, and Y. Bengio, “GFlowNet-EM for learning compositional latent variable models.” [Online]. Available: https://arxiv.org/abs/2302.06576