Learning Compositional and Causual Inductive Biases for Character Generation from Sketches

Kevin Mueller, Gordon Erlebacher

Department of Scientific Computing, Florida State University

Abstract

In recent years, the performance of handwriting recognition algorithms has increased dramatically, enabling many artificial intelligence (AI) applications that are now increasingly becoming part of our everyday lives. However, despite their impressive performance, current algorithms neglect to utilize the causal processes that underlie handwritten character generation, forcing them to learn every character from scratch. One way to better capture these processes is by building generative models that combine sketch and image data. Here, we show preliminary steps towards this goal by applying a state of the art sketch generative model to the Omniglot dataset, a popular dataset for handwritten character generation, and show that generative models that utilize sketches have more expressive latent spaces than models that exclusively use images. Furthermore, we apply this model to individual strokes to explore how well current generative models for sketches capture the degrees of variation across strokes.

Introduction

It is hypothesized that understanding handwritten characters, and all of their parts, exemplifies most of the fundamental problems of artificial intelligence. Specifically, in order for an AI system to fully understand handwritten characters, it needs to be able to: i) classify previously unseen characters after only seeing them once, ii) generate new characters from previously unseen alphabets, iii) parse the characters into their strokes and relationships, and iv) unconditionally generate new examples of characters. While the vast majority of previous work is concerned with problems (i,ii,iv), there is significantly less research into solving problem (iii). In the current work, we seek to bridge this gap by utilizing strokes, in the form of sketches, to explore sketch based methods for character generation.

Omniglot

Omniglot is a crowdsourced dataset of hand written characters from fictional and real alphabets form around the world. Omniglot is commonly referred to as the transpose of MNIST, as it contains a similar total number of images but with fewer examples for a larger number of classes. As a quick summary, the dataset includes:

- 1623 total character classes
- 50 alphabets (30 Training + 20 Testing)
- 20 images per character
- Size: 105 x 105
- Strokes (as xy-coordinates)

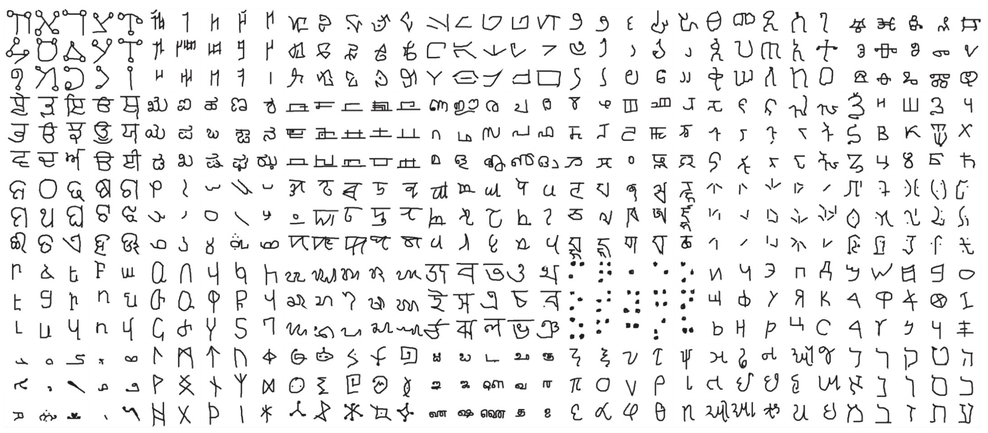

Figure 1: *Examples of characters from the omniglot dataset, displayed as blocks of 15 characters from 35 alphabets. (Blake et. al, 2015)*

Preprocessing

Due to a small amount of training data available in the omniglot dataset, it is necessary to perform a prepossessing step to reduce the total number of points for each stroke. In order to solve this task, we apply the Ramer–Douglas–Peucker (RDP) algorithm, a sub-sampling algorithm that removes points based on distance metric between the original and simplified curve. The parameter ϵ is used as a threshold parameter to decide whether a point should be kept. For the Omniglot dataset, setting $\epsilon = 2$ was deemed sufficient.

Model

We apply the sketch-rnn model from google to train a generative model for the omniglot dataset. We define an input sketch X from our dataset $\mathcal{D} = \{X\}_{i=1}^{N}$ as $X_i = \{\Delta x, \Delta y, p_1, p_2, p_3\}_{j=1}^{M}$ where Δ_x and Δ_y are the offset distance in the x and y directions. The pen states (p_1,p_2,p_3) are one-hot vectors that indicate if the pen is currently touching, if the pen will be lifted, or if the drawing has ended. The hyper-parameter M indicates the maximum stroke length allowed.

Sketch-RNN consists of two main parts:

- **RNN-Encoder**: A bi-directional RNN that maps an input sketch X to the variational parameters z_μ, z_{σ^2}. The latent random variable z is sampled using the reparemerization trick as $z = \sqrt{z_{\sigma^2}}\epsilon + z_\mu$, where $\epsilon \sim \mathcal{N}(0, 1)$.
- **RNN-Decoder**: A bi-directional RNN that maps the latent variable z back to X, by utilizing mixture density networks. Specifically, the decoder outputs the following Gaussian mixture parameters:

$$y_i = \left[\{\mu_{x,k}, \mu_{y,k}, \sigma_{x,k}, \sigma_{y,k}, \rho_{xy,k}, \Pi_k\}_{k=1}^{K}, (q_1, q_2, q_3)\right],$$

  where$\mu_{x,k}, \mu_{y,k}$ are the means of the kth Gaussian $(\sigma_{x,k}, \sigma_{y,k}, \rho_{xy,k})$ represent the covariance matrix in the x and y directions, Π_k are the mixture coefficients, and (q_1, q_2, q_3) represent the logits needed to calculate the pen states (p_1, p_2, p_3). Additionally, in order to assure the parameters are valid for a GMM, the $\exp$ and $\tanh$ operations are applied to the variance and co-variance terms respectively. The softmax operation is applied to categorical distribution parameters (q_1, q_2, q_3) and Π, to assure they sum to one.

With the above parameters we can calculate the joint probability of the GMM as

$$p(x, y) = \sum_k \Pi_k \mathcal{N}\left(\Delta x, \Delta y \mid \mu_{x,j}, \mu_{y,j}, \sigma_{x,j}, \sigma_{y,k}, \rho_{xy,j}\right) \text{ where } \sum_k \Pi_k = 1.$$

The reconstruction loss is calculated as the sum of the log-likelihood of $p(x, y)$ and cross-entropy of pen states. As is standard for VAEs, a KL-divergence regularizer is applied to the variational parameters (z_μ, z_{σ^2}).

Results

We train sketch-rnn on the Omniglot dataset and evaluate its performance by analyzing the encoded latent space z of characters and individual strokes.

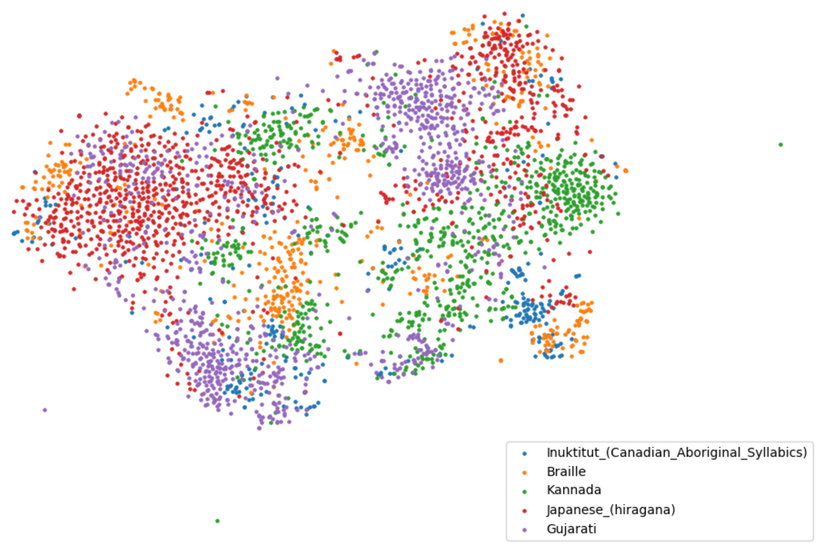

Figure 2: *A t-SNE projection (with perplexity = 40.0) of the latent space of character sketches from 5 randomly sampled alphabets.*

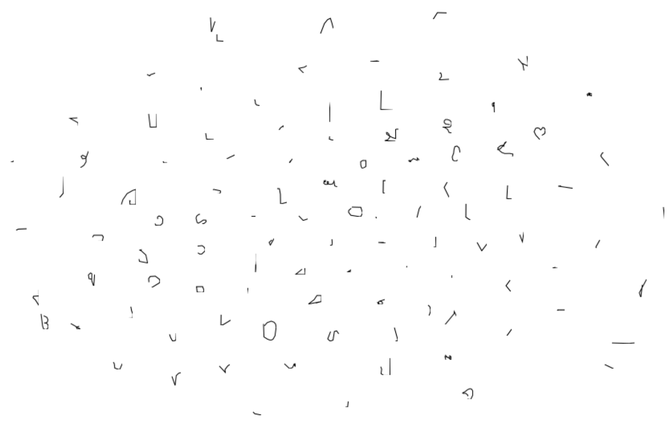

Figure 3: *A t-SNE projection (with perplexity = 40.0) of the latent space for 100 single strokes randomly sampled from the omniglot dataset. The projection illustrates that the model is able to capture properties (i.e sharpness, roundness, length) of strokes that repeatedly appear in different characters across alphabets.*

Discussion and Future Work

- This work provides strong evidence for modeling sketches in character generation
- Machine learning models for sketches allow for generalization to individual strokes,
- For future work, we would like to directly map images to sketches by utilizing a similar model with a CNN encoder.
- It may be possible to utilize the latent space of individual strokes to perform transfer learning across alphabets.

References

[1] D. P. Kingma and M. Welling Auto-Encoding Variational Bayes. *arxiv*, e-prints, Dec. 2013.

[2] David Ha and Douglas Eck. A Neural Representation of Sketch Drawings *arxiv*, 1704.03477, 2017.

[3] Lake BM, Salakhutdinov R, Tenenbaum JB: Human-level concept learning through probabilistic program induction. Science 2015, 350:1332-1338.