Exercise Sheet 8

Due: 15.12.25 14h00

Christof Gehrig, Gerrit Gerhartz

General Regulations.

- Please hand in your solutions in groups of two (preferably from the same tutorial group).
- Your solutions to theoretical exercises can be either handwritten notes (scanned), or typeset using LaTeX. For scanned handwritten notes, please ensure they are legible and not blurry.
- For the practical exercises, always provide the (commented) code as well as the output, and don't forget to explain/interpret the latter.
- Please hand in a **single PDF** that includes both the exported notebook and your solutions to the theoretical exercises. Submit the PDF to the Übungsgruppenverwaltung once per group, making sure to include the names of both group members in the submission.
- You can find all the data in the GitHub Repository.

1 Variational autoencoder

In the first part of this exercise, we will discuss some theoretical results to train a variational autoencoder (VAE).

(a) Let x be an observed variable (data) and z be a latent variable. We wish to approximate the true intractable posterior $p(z|x)$ using a simpler variational family $q_\phi(z|x)$. Starting from the Kullback-Leibler (KL) divergence $D_{\mathrm{KL}}(q_\phi(z|x)\|p(z|x))$, show that the log-likelihood of the data (the "evidence") can be decomposed as:

$$\log p(x) = \mathcal{L}(\phi, x) + D_{\mathrm{KL}}(q_\phi(z|x)\|p(z|x))$$

Where $\mathcal{L}(\phi, x)$ is the Evidence Lower Bound (ELBO). (3 pts)

(b) To train the VAE via backpropagation, we need to compute gradients of the expectation with respect to the parameters ϕ. However, we cannot backpropagate through a random sampling operation $z \sim q_\phi(z|x)$. Explain how the Reparameterization Trick allows us to bypass this problem. Specifically, express the random variable $z \sim \mathcal{N}(\mu, \sigma^2)$ as a deterministic function of μ, σ, and an auxiliary noise variable ϵ. (3 pts)

Now we will train a variational autoencoder on the MNIST handwritten digits dataset. This dataset is typically used to test new machine learning methods. The dataset contains 70000 samples of handwritten digits from 0-9.

(c) Use the code given to you in the Jupyter notebook to train a VAE on the MNIST dataset. What architecture is used for the encoder and decoder networks? How many latent dimensions are used? What learning hyperparameters are chosen? (1 pt)

(d) Plot 5 random original images from the test set alongside their reconstructed outputs from the VAE. What differences do you observe between the original and reconstructed images? How well does the VAE capture the details of the digits? (2 pts)

(e) Create a scatter plot of the encoded latent vectors z of the test set. Color the points based on their true digit label (0-9). Discuss if the clusters are distinct or overlapping and relate this to the properties of the KL-divergence term. (3 pts)

(f) Define a grid of points in the latent space (e.g., ranging from -3 to 3 on both axes) and decode them. Plot the resulting grid of images. What can you see? (2 pts)

2 SimCLR as special case of the Unified Contrastive Learning framework

In the paper "Understanding Deep Contrastive Learning via Coordinate-wise Optimization" (Tian, NeurIPS 2022[1]), a unified framework for contrastive losses is presented. The general loss function is given by

$$\mathcal{L}_{\phi,\psi}(\boldsymbol{\theta}) := \sum_{i=1}^{N} \phi\left(\sum_{j\neq i} \psi(d_i^2 - d_{ij}^2)\right),$$

where $d_i^2 = \frac{1}{2}\|z_i - z_{i'}\|^2$ is the (halved) squared distance between sample z_i and its augmented view $z_{i'}$, and $d_{ij}^2 = \frac{1}{2}\|z_i - z_j\|^2$ is the (halved) squared distance between samples z_i and z_j.

The NT-Xent (Normalized Temperature-scaled Cross-Entropy) loss used by SimCLR is defined as

$$\mathcal{L}_{\text{NT-Xent}} = -\sum_{i=1}^{N} \log \frac{\exp(\text{sim}(z_i, z_{i'})/\tau)}{\sum_{j\neq i} \exp(\text{sim}(z_i, z_j)/\tau)},$$

where $\text{sim}(z_i, z_j)$ is the cosine similarity and $\tau > 0$ is the temperature parameter.

(a) Show that for the choice $\phi(x) = \log(x)$ and $\psi(x) = \exp(x/\tau)$, the unified loss function $\mathcal{L}_{\phi,\psi}$ reduces to the NT-Xent loss.

Hint: Embedding vectors z_i can be assumed to be normalized, i.e., $\|z_i\|^2 = 1$. Then:

$$d_{ij}^2 = \frac{1}{2}\|z_i - z_j\|^2 = \frac{1}{2}\left(\|z_i\|^2 + \|z_j\|^2 - 2z_i^\top z_j\right) = 1 - z_i^\top z_j = 1 - \text{sim}(z_i, z_j)$$

(4 pts)

(b) Consider the temperature parameter τ. Explain how the loss function behaves, for τ very large, for $\tau = 1$ or τ very small. (2 pts)

[1] https://proceedings.neurips.cc/paper_files/paper/2022/hash/7b5c9cc08960df40615c1d858961eb8b-Abstract-Conference.html