Assignment 5

Ehsan Merrikhi 400101967
github repository

Deep Learning

Dr. Fatemizadeh

January 21, 2025

Deep Learning

Assignment 5

Ehsan Merrikhi 400101967
github repository

Contents

Contents **1**

Question 1 (20 Points) **2**
Part 1. . . . 2
Part 2. . . . 2
Part 3. . . . 3
Part 4. . . . 3

Question 2 (30 Points) **4**
Part 1. . . . 4
Part 2. . . . 4
Part 3. . . . 5
1. . . . 5
2. . . . 6

Question 3 (20 Points) **7**

Question 4 (30 Points) **8**
Part 1. . . . 8
Part 2. . . . 8

Question 1 (20 Points)

In this question, we intend to discuss the differences between VAE and AE.

Part 1.

Suppose we want to generate data similar to a dataset using a typical AE. We train the AE and select a random point (with a uniform distribution) in the latent space, then input it into the trained decoder module. In your opinion, is the likelihood higher that the decoder's output will resemble the dataset, or is it more likely that a randomly chosen point in the data space will resemble the dataset? Why? (5 points)

Soloution

The likelihood that the decoder's output will resemble the dataset is **lower** compared to a randomly chosen point in the data space. This is because:

1. **Latent Space of AE:** A typical AE does not enforce a structured latent space. Random points in the latent space are unlikely to align with regions learned from the data, leading to meaningless outputs.
2. **Data Space Distribution:** The data space inherently contains real data points, making a randomly chosen point more likely to resemble the dataset.
3. **Decoder Behavior:** The decoder maps specific latent representations to the data space, but random latent points are not guaranteed to produce valid data.

Part 2.

List at least three shortcomings of the method in part (a) for generating data similar to the dataset, and explain how **VAE** addresses these problems. (5 points)

Soloution

Shortcomings of the method in part (a):

1. **Unstructured Latent Space:** The latent space of a typical AE lacks structure, so random samples often result in meaningless outputs.
2. **Low Coverage of Data Space:** Latent representations only correspond to training data regions, making random points unlikely to produce valid results.
3. **No Diversity in Outputs:** AEs do not enforce diversity, limiting the ability to generate varied samples.

How VAE Addresses These Problems:

1. **Structured Latent Space:** VAE learns a probabilistic prior (e.g., Gaussian) that ensures meaningful outputs from random samples.
2. **Higher Coverage:** By aligning the latent space with the data distribution, VAE generates samples resembling the dataset.
3. **Encourages Diversity:** Sampling from a prior distribution allows VAE to produce diverse outputs reflective of the original data.

Part 3.

Suppose that during the training process of an AE, Gaussian noise with zero mean and variance $0.05 \times R$ is added to the output of the encoder. Here, R represents the mean squared distance of points in the latent space from their center, which is updated in each training step. Will the decoder trained using this method perform better than the decoder of a typical AE? Specifically, if a random point is selected from the latent space, which decoder is more likely to produce an output resembling the dataset? (5 points)

Soloution

Adding Gaussian noise to the output of the encoder improves the decoder's performance compared to a typical AE because:

1. **Smoother Representations:** Noise helps the decoder learn robust latent representations, improving generalization.
2. **Better Random Sampling:** The decoder becomes better at handling random points in the latent space due to reduced overfitting.
3. **Regularization Effect:** The noise acts as a regularizer, encouraging the learning of meaningful features.

Part 4.

Does VAE have an advantage over the method mentioned in (3)? What is the key difference between these two methods? (5 points)

Soloution

Advantages of VAE over the noise-augmented AE method:

1. **Structured Latent Space:** VAE explicitly learns a structured probabilistic distribution (e.g., Gaussian), enabling better sampling.
2. **Probabilistic Framework:** VAE uses KL divergence to align the latent space with a prior distribution, ensuring meaningful outputs from random points.
3. **Higher Validity:** Random samples from VAE's latent space are more likely to produce outputs resembling the dataset.

Key Difference: VAE explicitly models a structured and continuous latent space, while the noise-augmented AE only indirectly regularizes the latent space with added noise.

Question 2 (30 Points)

In this exercise, we aim to deepen our understanding of Maximum Likelihood Estimation (MLE) and its relationship with VAE.

Part 1.

Suppose our dataset is given as $D = \{x_1, x_2, \ldots, x_n\}$. Study the concept of maximum likelihood estimation and explain why the parameters of the distribution should be such that the following relation is maximized:

$$\sum_{i=1}^{n} \log(p_\theta(x_i))$$

Note that $p_\theta(x_i)$ represents the probability of observing x_i as a function of the parameters θ. (5 points)

Soloution

Maximum Likelihood Estimation (MLE) is a method for estimating the parameters of a statistical model by maximizing the likelihood function. The goal of MLE is to find the parameters θ that make the observed data $D = \{x_1, x_2, \ldots, x_n\}$ most probable.
Reason for maximizing the relation:

1. **Likelihood Function:** The likelihood function for the dataset is given by:

   $$L(\theta) = \prod_{i=1}^{n} p_\theta(x_i)$$

   This function represents the joint probability of observing all the data points under the given parameters θ.

2. **Log-Likelihood:** To simplify computation, the log of the likelihood function is taken, which turns the product into a sum:

   $$\log L(\theta) = \sum_{i=1}^{n} \log(p_\theta(x_i))$$

   Maximizing this log-likelihood is equivalent to maximizing the likelihood, as the logarithm is a monotonic function.

3. **Parameter Optimization:** By maximizing $\sum_{i=1}^{n} \log(p_\theta(x_i))$, we are adjusting θ to find the parameters that best explain the observed data, ensuring that the model aligns with the data distribution as closely as possible.

The parameters θ are chosen to maximize the log-likelihood because it leads to the highest probability of observing the given dataset under the model, ensuring that the model is the best fit for the data.

Part 2.

Show the equivalence between minimizing cross-entropy loss and Maximum Likelihood Estimation (MLE). (5 points)

Soloution

Equivalence between Cross-Entropy Loss and MLE:

1. **Cross-Entropy Loss:** For a dataset $D = \{x_1, x_2, \ldots, x_n\}$ with true labels y_i, the cross-entropy loss is defined as:

$$\text{Loss} = -\frac{1}{n}\sum_{i=1}^{n}\sum_{c=1}^{C} y_i^c \log(p_\theta(x_i^c)),$$

   where $p_\theta(x_i^c)$ is the predicted probability for class c and y_i^c is 1 if x_i belongs to class c, and 0 otherwise.

2. **Simplification for True Labels:** Since $y_i^c = 1$ only for the true class of x_i, this simplifies to:

$$\text{Loss} = -\frac{1}{n}\sum_{i=1}^{n} \log(p_\theta(x_i)).$$

3. **Maximum Likelihood Estimation:** The likelihood of the dataset under the model is given by:

$$L(\theta) = \prod_{i=1}^{n} p_\theta(x_i).$$

   Taking the logarithm of the likelihood (log-likelihood):

$$\log L(\theta) = \sum_{i=1}^{n} \log(p_\theta(x_i)).$$

   Maximizing the log-likelihood is equivalent to minimizing the log-likelihood

Part 3.

We know that the ultimate goal of a **VAE** is to have a generative model whose output distribution resembles the dataset's distribution. Similar to conventional neural networks, **VAE** uses **stochastic gradient descent**. Thus, during the learning process, instead of seeing the entire dataset at once and maximizing:

$$\sum_{i=1}^{n} \log(p_\theta(x_i)),$$

we aim to update the model parameters after each input to slightly increase $\log(p_\theta(x_i))$. This means that for each input, the goal is to slightly increase the likelihood of generating an output resembling that input. As discussed in the course, this logarithmic probability aligns with the **ELBO** equation, which is shown below.

$$\log p_\theta(x_i) - D_{\text{KL}}[q_\phi(z|x_i) \parallel p_\theta(z|x_i)] = \mathbb{E}_z[\log p_\theta(x_i|z)] - D_{\text{KL}}[q_\phi(z|x_i) \parallel p_\theta(z)] \tag{1}$$

In equation (1), θ represents the parameters of the decoder, and ϕ represents the parameters of the encoder.

1.

Prove that the KL divergence D_{KL} is non-negative and thereby derive a lower bound for the logarithm of the likelihood. (5 points)

Soloution

Proof of Non-Negativity of KL Divergence: The KL divergence between two distributions $q(z)$ and $p(z)$ is defined as:

$$D_{\mathrm{KL}}[q(z) \parallel p(z)] = \int q(z) \log \frac{q(z)}{p(z)} \, dz.$$

Using Jensen's inequality, which states that $\mathbb{E}[\log(X)] \leq \log(\mathbb{E}[X])$, we have:

$$\int q(z) \log \frac{q(z)}{p(z)} \, dz \geq 0.$$

This proves that the KL divergence is always non-negative, with equality only when $q(z) = p(z)$ almost everywhere.
Lower Bound (ELBO): From the equation:

$$\log p_\theta(x_i) = \mathbb{E}_z[\log p_\theta(x_i|z)] - D_{\mathrm{KL}}[q_\phi(z|x_i) \parallel p_\theta(z)],$$

since $D_{\mathrm{KL}} \geq 0$, we derive:

$$\log p_\theta(x_i) \geq \mathbb{E}_z[\log p_\theta(x_i|z)].$$

The **Evidence Lower Bound (ELBO)** is:

$$\mathrm{ELBO} = \mathbb{E}_z[\log p_\theta(x_i|z)] - D_{\mathrm{KL}}[q_\phi(z|x_i) \parallel p_\theta(z)].$$

▬ *2.*

Justify why, in many implementations of **VAE**, the term $\mathbb{E}_z[\log p_\theta(x_i|z)]$ corresponds to minimizing the cross-entropy loss between the dataset and the decoder's output. (15 points)

Soloution

Relation to Cross-Entropy: - The term $\mathbb{E}_z[\log p_\theta(x_i|z)]$ represents the expected log-likelihood of the data given latent z. - For Gaussian or Bernoulli distributions, minimizing $-\mathbb{E}_z[\log p_\theta(x_i|z)]$ corresponds directly to cross-entropy loss:

$$\text{Cross-Entropy Loss} = -\frac{1}{n} \sum_{i=1}^{n} x_i \log(\hat{x}_i) + (1 - x_i) \log(1 - \hat{x}_i).$$

Justification:

1. Cross-entropy measures divergence between true data and decoder outputs.
2. Minimizing cross-entropy aligns with maximizing $\mathbb{E}_z[\log p_\theta(x_i|z)]$.
3. It ensures the decoder produces outputs resembling the dataset.

The VAE's term $\mathbb{E}_z[\log p_\theta(x_i|z)]$ directly relates to minimizing cross-entropy loss, ensuring decoder outputs resemble the true data distribution.

Question 3 (20 Points)

Why is it assumed in VAE that the latent space distribution is Gaussian? (Refer to reasons beyond simplification of computations.) Investigate whether distributions other than Gaussian are also used in practice.

Soloution

Reasons for Assuming Gaussian Distribution in VAE:

1. **Smoothness and Continuity:** A Gaussian distribution ensures that the latent space is continuous and smooth, allowing small changes in the latent variables to correspond to meaningful variations in the generated data.
2. **Representational Power:** The Gaussian distribution is highly flexible and can approximate many types of data distributions when transformed by a deep neural network.
3. **Interpretability:** Gaussian distributions are well-understood in statistics and provide interpretable properties, such as mean and variance, which can represent central tendencies and variations in the data.
4. **Regularization Effect:** Using a Gaussian prior encourages the encoder to distribute latent variables in a way that avoids overfitting, improving generalization in generative tasks.

Other Distributions Used in Practice: While Gaussian priors are common, other distributions are also used depending on the application:

1. **Uniform Distribution:** For cases where a uniform latent space is desired, such as in some adversarial autoencoders.
2. **Mixture Models:** Mixture-of-Gaussians or categorical priors can be used when the data naturally clusters into distinct groups.
3. **Non-Euclidean Spaces:** Distributions on manifolds (e.g., hyperspheres or hyperbolic spaces) are used in specific applications like text or graph generation.
4. **Heavy-Tailed Distributions:** Distributions such as Laplace or Student-t can handle outliers and heavy-tailed data better than Gaussian.

Question 4 (30 Points)

Study the paper on **β-VAE** and answer the following questions:

Part 1.

Briefly explain the idea of β-VAE and state how it differs from VAE. (15 points)

Soloution

β-VAE extends the traditional Variational Autoencoder (VAE) by introducing a hyper-parameter $\beta(> 1)$ to the Kullback–Leibler (KL) divergence term in the loss function. Specifically, the loss is modified to:

$$\mathcal{L}_{\beta\text{-VAE}}(\theta, \phi; x) = -\mathbb{E}_{q_\phi(z|x)}[\log p_\theta(x|z)] + \beta\, D_{\mathrm{KL}}(q_\phi(z|x) \,\|\, p(z)).$$

When $\beta = 1$, this reduces to the standard VAE objective. For $\beta > 1$, the KL term is emphasized, encouraging the model to learn more disentangled latent factors. Typically, each latent dimension in a disentangled representation corresponds to a distinct data-generating factor. Hence, β-VAE trades off a bit of reconstruction fidelity (reconstruction loss) for more factorized or "disentangled" latent representations.

Part 2.

Based on the information in Section 2 of the paper, describe the importance and function of the disentanglement metric. (15 points)

Soloution

Section 2 of the paper introduces a disentanglement metric that quantifies how well the learned latent variables capture separate underlying factors of variation. Concretely, one ground-truth factor is perturbed at a time while the others are kept fixed, and the model's latent outputs are observed. A high disentanglement score indicates that varying one factor primarily affects a single latent dimension, demonstrating that each latent dimension encodes an independent factor. This is critical because well-disentangled representations are more interpretable and improve generalization in downstream tasks.