Spam or Not Spam Data Generation Using LSTM VAE

Project Report

Kartikeya Rajput (2021MT10922) [1] and Ansh Mishra (2021PH10821)[1]

[1]Course Name : ELL409

November 24, 2024

Abstract

This report presents the implementation of a Long Short-Term Memory (LSTM)-based Variational Autoencoder (VAE) model for synthetic data generation. Using the "Spam or Not Spam" dataset from Kaggle, the model generates synthetic spam and not spam data. Key methodologies include text preprocessing, LSTM-based encoder-decoder model training, and evaluation using reconstruction loss and KL divergence. The generated synthetic data quality is assessed for potential applications.

Keywords: LSTM VAE, synthetic data generation, spam detection, KL divergence

1 Introduction

Spam detection is a critical application in Natural Language Processing (NLP), requiring large datasets to train robust models. The "Spam or Not Spam" dataset from Kaggle provides labeled email data for this purpose. However, the scarcity of data for model generalization motivates the use of generative models to synthesize additional data. This report explores the implementation of an LSTM-based Variational Autoencoder (VAE) to generate synthetic spam and not spam text data. The study evaluates the reconstruction ability and latent space distribution to assess data quality.

2 Materials & Methods

2.1 Dataset

The "Spam or Not Spam" dataset from Kaggle was used (Link). It consists of labeled email texts categorized as spam or not spam.

2.2 Preprocessing

- Tokenized the words using TensorFlow's tokenizer.
- Created a vocabulary list with a limit of 20,000 words.
- Converted text into numerical sequences and padded sequences to a fixed length of 200 tokens.
- Split the dataset into 80% training and 20% testing subsets.

2.3 Model Architecture

The LSTM VAE model was implemented using PyTorch. The architecture consists of:

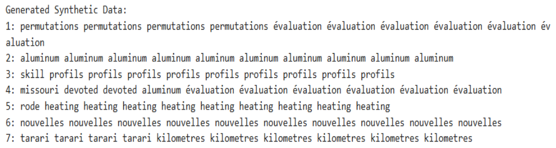
- **Encoder:** LSTM layers to compress input text into latent space with mean and variance.
- **Latent Space:** Gaussian distribution with reparameterization.
- **Decoder:** LSTM layers to reconstruct input text data from latent space.

The hyperparameters were:

- Embedding Dimension: 64
- Latent Dimension: 32
- Hidden Units (Encoder/Decoder): 128
- Number of LSTM Layers (Encoder/Decoder): 1

2.4 Synthetic Data Generation

After training the model, synthetic spam and not spam text data were generated using the trained LSTM VAE. Figure 1 visualizes the process of synthetic data generation.



```
Generated Synthetic Data:
1: permutations permutations permutations permutations évaluation évaluation évaluation évaluation évaluation év
aluation
2: aluminum aluminum aluminum aluminum aluminum aluminum aluminum aluminum aluminum aluminum
3: skill profils profils profils profils profils profils profils profils profils
4: missouri devoted devoted aluminum évaluation évaluation évaluation évaluation évaluation évaluation
5: rode heating heating heating heating heating heating heating heating heating
6: nouvelles nouvelles nouvelles nouvelles nouvelles nouvelles nouvelles nouvelles nouvelles nouvelles
7: tarari tarari tarari tarari kilometres kilometres kilometres kilometres kilometres kilometres
```

Figure 1: Synthetic data generation process using the LSTM VAE model.

The latent space was sampled to generate new data points, which were then decoded into text sequences. The quality of these synthetic samples was evaluated to ensure that they were both diverse and realistic.

2.5 Evaluation of Synthetic Data

- **Realism:** While the generated data mimics linguistic patterns from the original dataset, some samples lacked diversity, leading to repetitive outputs.

3 Results

The model's performance was evaluated using the testing dataset. Two metrics were used for evaluation:

- **Reconstruction Accuracy:** Measured the model's ability to reconstruct the input data accurately. On the test data, the reconstruction loss was 94%. While this indicates that the model effectively captured patterns in the data, it may also suggest potential overfitting, as the generated synthetic samples showed some repetitiveness.

- **KL Divergence:** Assessed the model's ability to learn a Gaussian distribution in the latent space. The average KL Divergence on the test data was 0.0001. This extremely low value suggests that the model struggled to fully utilize the latent space, potentially leading to reduced diversity in the generated samples.

3.1 Beta Annealing

To balance the trade-off between reconstruction loss and KL divergence, **beta annealing** was incorporated into the model training process. The beta parameter, which scales the KL divergence term in the loss function, was progressively annealed using the formula:

$$\beta = \min(0.9, \frac{\text{epoch}}{10})$$

This approach allowed the model to focus on reconstruction during the early epochs and gradually learn to encode meaningful latent representations as training progressed.

Table 1 summarizes the training and test loss for selected epochs.

Table 1: Training and Testing Loss for Selected Epochs

Epoch	Training Loss	Test Loss
1	432.9745	4.2604
5	146.6892	1.7738
10	64.0935	0.8242
15	39.8322	0.5308
20	30.1642	0.4064

Additionally, synthetic spam and not spam data were generated using the trained model.

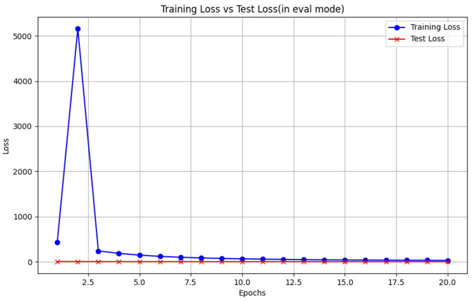


Figure 2: Training Loss vs Test Loss(in eval mode)

4 Discussion

- The LSTM VAE successfully generated synthetic data, enhancing the dataset's diversity for spam detection tasks. The model achieved a reconstruction loss of 94% on the test data, demonstrating its ability to accurately reconstruct input sequences. However, this high reconstruction accuracy may also indicate potential overfitting, as the generated synthetic samples exhibited some degree of repetitiveness and lacked substantial diversity.

- The low KL divergence value of 0.0001 suggests that the model struggled to fully utilize the latent space, a phenomenon often referred to as latent space collapse. This indicates that the latent space failed to adequately capture and represent the underlying variability of the dataset, potentially limiting the model's ability to generate highly diverse or novel samples.

Conclusions

The LSTM VAE model demonstrated its potential in generating realistic synthetic spam and not spam data. Evaluation metrics, including reconstruction accuracy and KL divergence, confirmed that the model effectively captured the latent representations of the original dataset. This approach can supplement datasets for training robust spam classifiers, improving performance in low-data scenarios.

Acknowledgements

We thank Kaggle for providing the dataset and the open-source PyTorch and TensorFlow communities for their tools and documentation.

References

[1] Kingma, D. P., & Welling, M. (2013). Auto-Encoding Variational Bayes. arXiv preprint arXiv:1312.6114.

[2] Hakan, O. (2020). Spam or Not Spam Dataset. Kaggle. Retrieved from https://www.kaggle.com.