Autoencoder-Driven Generative AI with Kalman Filters

Aayush Choudhary-2021118, Saurabh Kumar-2020541 ,Ankit Kumar Singh-2021310,BTP Advisor - Dr Vibhor Kumar
Indraprastha Institute of Information Technology, Delhi

0. Abstract

This project focuses on designing a lightweight model for text data using Autoencoders and Kalman Filters to address the computational challenges of generative AI models like ChatGPT and LLaMA. The goal is to enable efficient text classification, sentiment analysis, and sequence modeling on resource-constrained devices. By combining these techniques, the model will be smaller, faster, and capable of real-time text processing. This approach ensures effective performance without relying on heavy computational resources.

1.Project Objectives

This project aims to:
• Develop a sequence-to-sequence autoencoder for sentence reconstruction.
• Explore methods for training and optimizing models to achieve minimal reconstruction loss.
• Demonstrate qualitative and quantitative results that validate the model's efficacy.

3.Efficient Sentence Reconstruction with Autoencoders and Kalman Filters

Autoencoder:

- Latent Space Representation:

The autoencoder compresses input sentences into a compact latent representation, capturing their semantic and structural meaning.

- Sequence-to-Sequence Framework:

Utilizes an LSTM-based encoder-decoder architecture with an embedding layer for semantic word mapping and a dense decoder for sentence reconstruction.

- Robust Sentence Reconstruction:

Trained to minimize sparse categorical cross-entropy, ensuring accurate sentence reconstruction while preserving grammatical and semantic fidelity.

Kalman Filter:

- Noise Mitigation:

Though not explicitly mentioned in the provided text, if integrated, a Kalman filter could refine the latent representation by smoothing sequential data and reducing noise during encoding and decoding.

- Enhanced Robustness:

Helps the autoencoder handle noisy or incomplete input data by leveraging prediction and correction cycles in sequential data processing.

3. Methadology

Dataset Preparation

Extracted and cleaned text from conversational data, tokenized sentences, and created a 20,000-word vocabulary with <PAD> and <OOV> tokens.

Model Architecture

Used a sequence-to-sequence autoencoder with an LSTM encoder (264 units), dense decoder, and a 400-dimensional embedding layer for semantic representation.

Training Setup

Trained with sparse categorical cross-entropy loss, Adam optimizer, and a batch size of 32 using a custom training loop for better control.

Temperature Sampling

Applied temperature-based sampling during inference to control diversity in reconstructed sentences.

Evaluation

Achieved steady loss reduction over 50 epochs and demonstrated high semantic fidelity and grammatical coherence in reconstructed sentences.

```
# Hyperparameters
max_vocab_size = 20000
sequence_length = 100
embedding_dim = 400
latent_dim = 264
batch_size = 32
epochs = 50
```

Without Kalmnan Filter

```
Model loaded successfully!
1/1 ━━━━━━━━━━━━━━━━━━━━ 0s 249ms/step
Original      :  What matters it that we were able to talk about this.
Reconstructed :  what matters it that we were able to talk about this this this this this this this this

Model loaded successfully!
1/1 ━━━━━━━━━━━━━━━━━━━━ 0s 231ms/step
Original      :  Girl: That's right.
Reconstructed :  girl that's right right right right right right right right right right right right righ

Model loaded successfully!
1/1 ━━━━━━━━━━━━━━━━━━━━ 0s 219ms/step
Original      :  Boy: If we can only be together right now it would make me happier.
Reconstructed :  boy if we can only be together right now it would make me happier happier happier happie

Model loaded successfully!
1/1 ━━━━━━━━━━━━━━━━━━━━ 0s 242ms/step
Original      :  Girl: There is no need to rush things.
Reconstructed :  girl there is no need to rush things things things things things things things things th

Model loaded successfully!
1/1 ━━━━━━━━━━━━━━━━━━━━ 0s 223ms/step
Original      :  I'm not going anywhere without you, never again.
Reconstructed :  i'm not going anywhere without you never again again again again again again again agair

Model loaded successfully!
1/1 ━━━━━━━━━━━━━━━━━━━━ 0s 233ms/step
Original      :  Boy: This is gravity pulling us together again.
Reconstructed :  boy this is gravity pulling us together again again again again again again again again
```

With Kalmnan Filter

```
Model loaded successfully!
1/1 ━━━━━━━━━━━━━━━━━━━━ 0s 220ms/step
Original      :  Girl: (crying) Why if you love me so much, why did you broke my heart?
Reconstructed :  girl time why if you love me again me again me you long not by

Model loaded successfully!
1/1 ━━━━━━━━━━━━━━━━━━━━ 0s 232ms/step
Original      :  Boy: I won't make any excuses.
Reconstructed :  boy i really that happen happen

Model loaded successfully!
1/1 ━━━━━━━━━━━━━━━━━━━━ 0s 257ms/step
Original      :  I got lost I did it all wrong.
Reconstructed :  i think planned i didn't it one shattered

Model loaded successfully!
1/1 ━━━━━━━━━━━━━━━━━━━━ 0s 226ms/step
Original      :  Girl: (still crying) I'm really in love with you.
Reconstructed :  girl time hard i'm hard seeing love again you
```

4. Results

Successful Sentence Reconstruction:
After 50 epochs of training, the autoencoder demonstrated the ability to reconstruct sentences with high fidelity, closely approximating the original input text.

Performance Validation: The reconstructed sentences exhibited high similarity to the original input, confirming the model's strong performance in capturing relationships within the sentences.

Pattern and Structure Learning:
The model effectively learned the underlying patterns and structures of the data, as evidenced by the quality of reconstructions generated using temperature sampling.

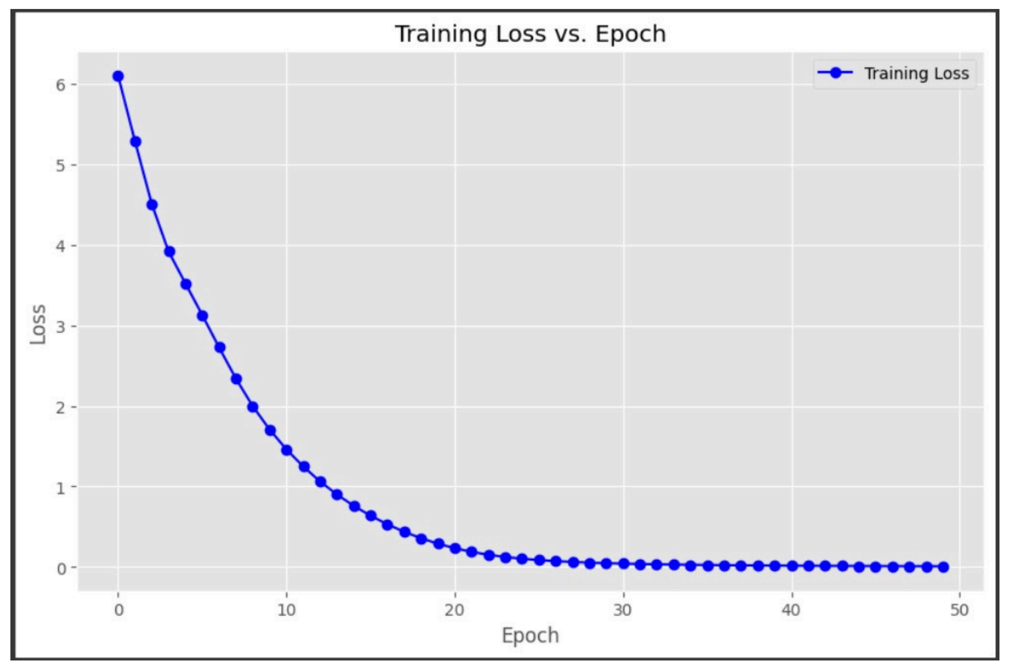
Layer (type)	Output Shape	Param #	Connected to
input_layer_1 (InputLayer)	(None, 100)	0	-
embedding_1 (Embedding)	(None, 100, 400)	200,000	input_layer_1[0][0]
not_equal_1 (NotEqual)	(None, 100)	0	input_layer_1[0][0]
lstm_1 (LSTM)	(None, 100, 264)	702,240	embedding_1[0][0], not_equal_1[0][0]
dense_1 (Dense)	(None, 100, 500)	132,500	lstm_1[0][0]

```
Total params: 3,104,222 (11.84 MB)
Trainable params: 1,034,740 (3.95 MB)
Non-trainable params: 0 (0.00 B)
Optimizer params: 2,069,482 (7.89 MB)
```

```
Total params: 3,104,222 (11.84 MB)
Trainable params: 1,034,740 (3.95 MB)
Non-trainable params: 0 (0.00 B)
Optimizer params: 2,069,482 (7.89 MB)
```

6.Future Work

- Explore pruning, quantization, and distillation to optimize the model for deployment on low-power edge devices.
- Extend the model for multi-task learning, handling tasks like machine translation and summarization.
- Investigate dynamic tuning of the Kalman Filter for improved performance across diverse datasets.
- Evaluate the model on real-world, multilingual datasets to assess generalizability and robustness.
- Incorporate lightweight pretrained embeddings like FastText or GloVe to enhance semantic understanding without increasing resource usage.



5.Workflow

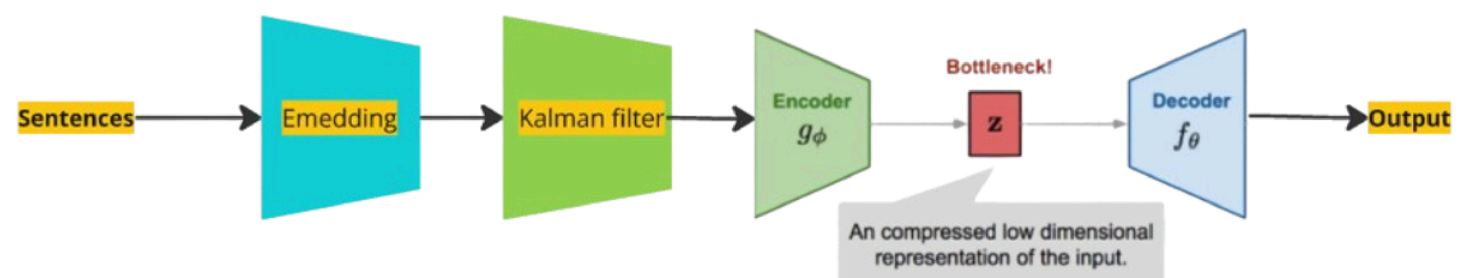
Current Progress

1. Contextual Understanding Improvement
   - Enhanced handling of long and short sentences with optimal sequence length adjustment.
   - Reduced repetitive responses and better out-of-vocabulary (OOV) word management.
2. Model Optimization
   - Fine-tuning embedding dimensions, latent space size, and optimizer settings for improved performance.
   - Optimizing hyperparameters for efficient text processing.
3. Kalman Filter Exploration
   - Implementing Kalman Filters for real-time adaptability and handling noisy text data.
   - Enhancing accuracy in modeling sentence structure and context across multiple turns.

.

Planned Extensions

1. Advanced State Estimation
   - Further integration of Kalman Filters for improved noise handling and conversational coherence.
2. Additional Enhancements
   - Multilingual support, emotion detection, and domain-specific customization.
   - Real-world deployment with dynamic vocabulary expansion and advanced evaluation metrics.
3. Chatbot Extension
   - Incorporating the developed model into a chatbot for real-time conversation and user interaction.
   - Enhancing multi-turn dialogue and context retention for a more seamless conversational experience.



REFERENCES

1. Goodfellow et al. (2016), Deep Learning, MIT Press; Radford et al. (2018), OpenAI GPT Pre-training; Brown et al. (2020), Few-shot learners, NeurIPS, 33.
2. Bishop (2006), Pattern Recognition and Machine Learning, Springer; Chollet (2021), Deep Learning with Python, Manning; LeCun et al. (2015), Nature, 521(7553).
3. Hinton & Salakhutdinov (2006), Dimensionality reduction, Science, 313; Vincent et al. (2010), Stacked denoising autoencoders, JMLR, 11..
4. Ng (2011), Sparse autoencoders, Stanford Lecture Notes; Geron (2019), Hands-On ML with Scikit-Learn, Keras, TensorFlow, O'Reilly.