Medical Expertise Style Transfer using Denoising Autoencoder

Mohammad Sabik Irbaz
ID : 160041004
sabikirbaz@iut-dhaka.edu

Abir Azad
ID : 160041024
abirazad@iut-dhaka.edu

Anika Tasnim Preoty
ID : 160041044
anikatasnim@iut-dhaka.edu

Tani Barkat Shalanyuy
ID : 160041083
tanibarkat@iut-dhaka.edu

Supervisor

Md. Kamrul Hasan, Ph.D.
Professor, Department of Computer Science and Engineering
hasank@iut-dhaka.edu

System and Software Lab (SSL), IUT

Introduction

- ***Text style transfer*** is one of the most important NLG (Natural Language Generation) tasks, which aims to control certain attributes that people may expect the texts to possess, such as sentiment, tense, emotion, political position, etc.

- ***Expertise Style Transfer*** (*) aims to transfer the expert text to layman text.

- Our main focus of the research is Expertise Style Transfer in the ***Medical Domain***.

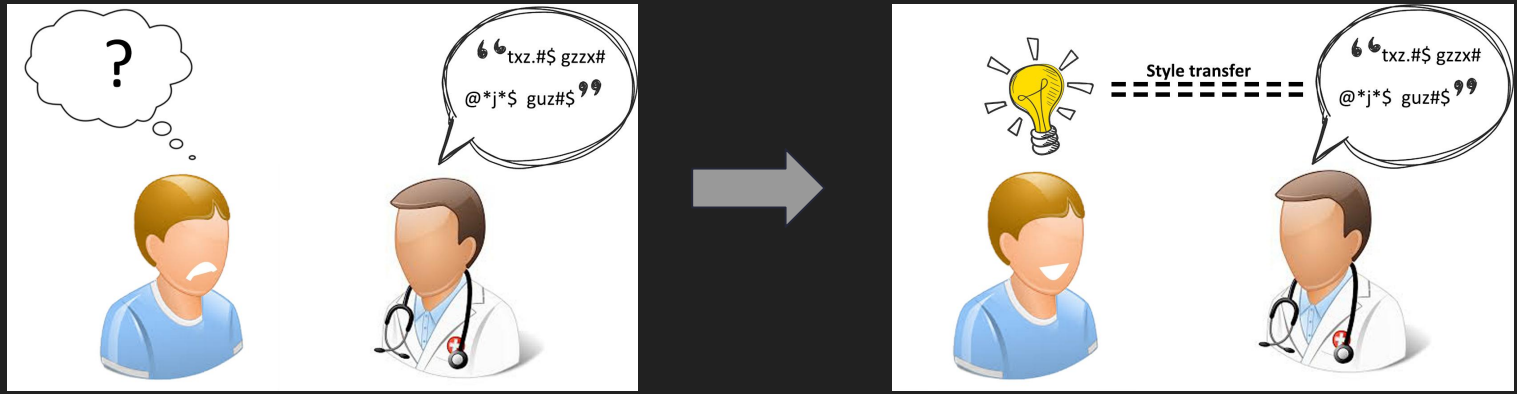
- We aim to formulate an approach to transfer expertised text efficiently while preserving the content.

(*) recently introduced by Yixin Cao, et. al., in ACL 2020 [1]

Medical Expertise Style Transfer

In medical domain, doctors as domain-experts have a different style of interaction and they talk using an expertised vocabulary.

This creates a huge communication gap between doctors and patients.



Motivation

- Expertise style transfer is one of the most recently introduced tasks in the Text Style Transfer Domain.

- Contributing to this research will help in decreasing the communication gap between experts and laymen in real life.

Problem Statement

Given an expert style text
can the machine learn to transfer it into
layman style preserving the content?

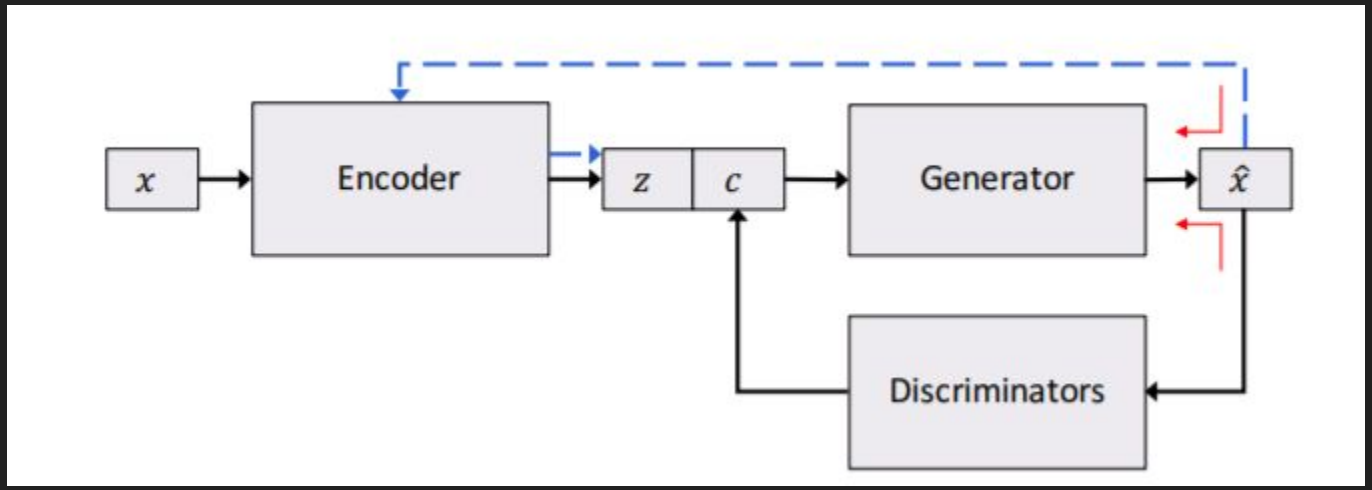
Research Challenges

- Non parallel Dataset
- Unstable evaluation metrics
- Unavailability of hardware like GPU

Background Study

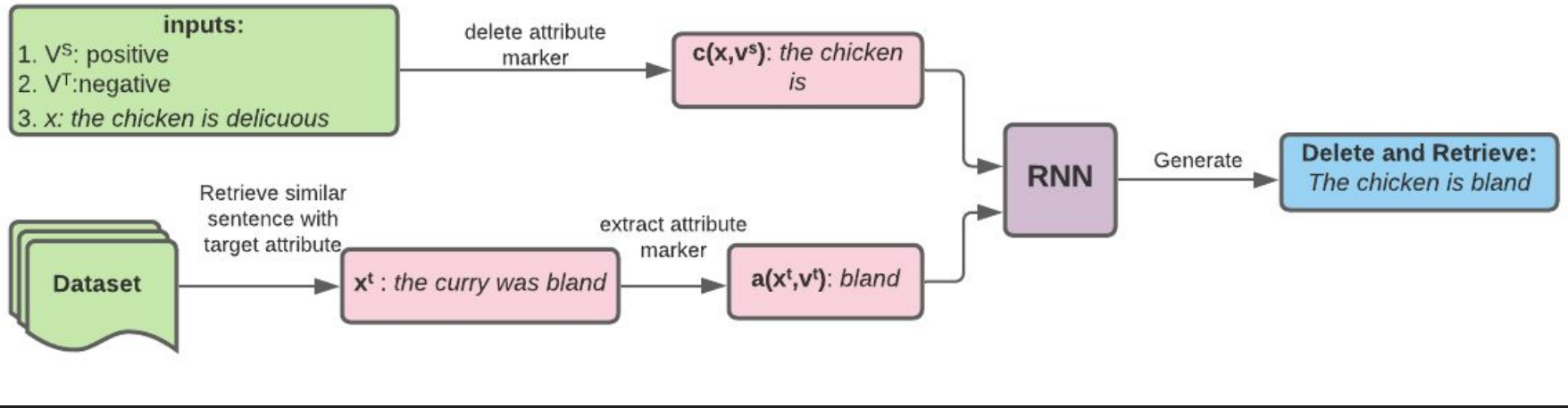
Existing approaches and Literature Review

Controlled Generation [Hu et. al., 2017]



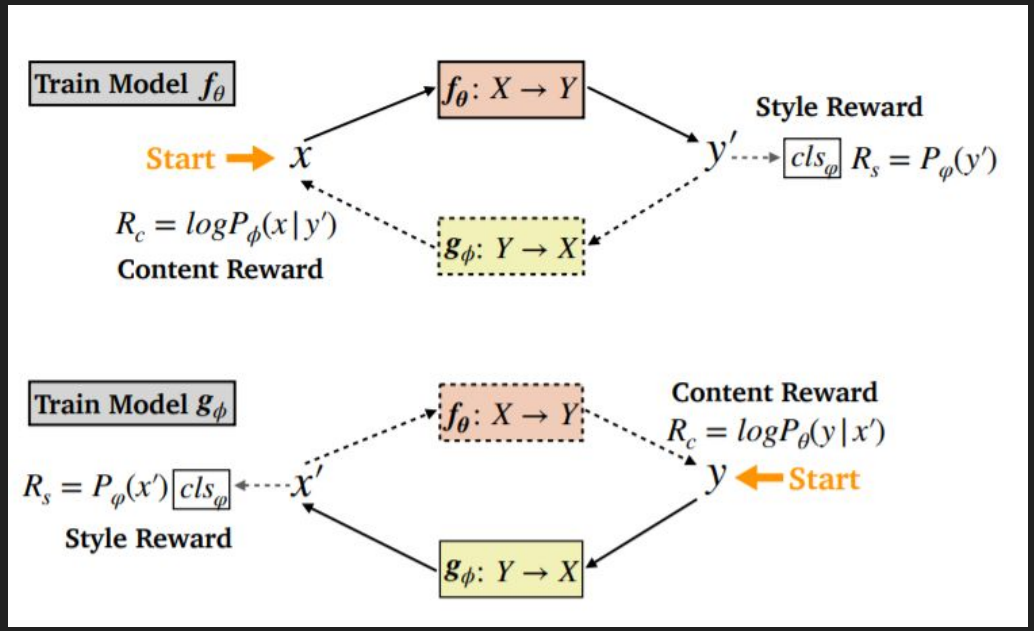
- A style classifier works as discriminator.
- z is the latent space where the texts are encoded
- c is the decoder space need for generation
- a noise factor is returned to encoder after each epoch

Delete And Retrieve [Li, et. al. 2018]



- **Delete**
- **Retrieve**
- **Generate**

Dual Reinforcement Learning (RL) [Luo et. al. 2019]



- Model ‘f’ is responsible for transferring style
- Model ‘g’ is responsible for preserving the content

Shortcomings of the existing models

Controlled-Gen

Performs very well in terms of **content similarity**; Poorly in terms of **style accuracy**.

Delete & Retrieve

Performs very well in terms of **style accuracy**; Poorly in terms of **content similarity**.

Dual RL

Takes too much time to train and test and does not perform robustly in terms of style accuracy and content similarity in a different dataset.

Analogy

Expert Sentence

Sequence of words

Analogy

Expert Sentence

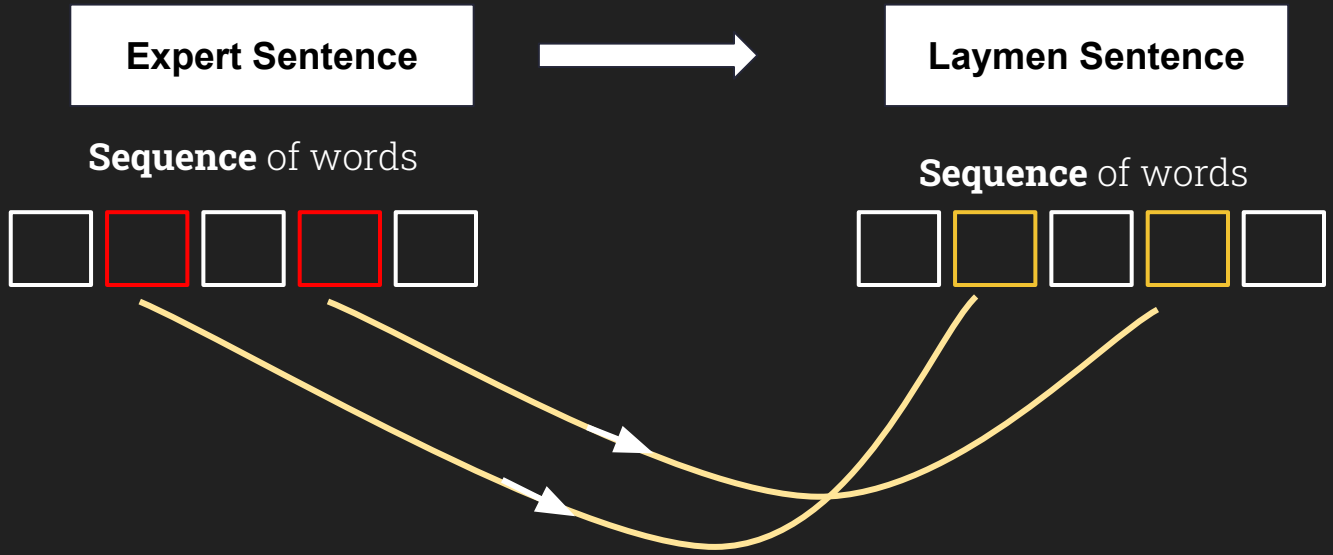
Sequence of words

Analogy

Expert Sentence → **Laymen Sentence**

Sequence of words

Analogy

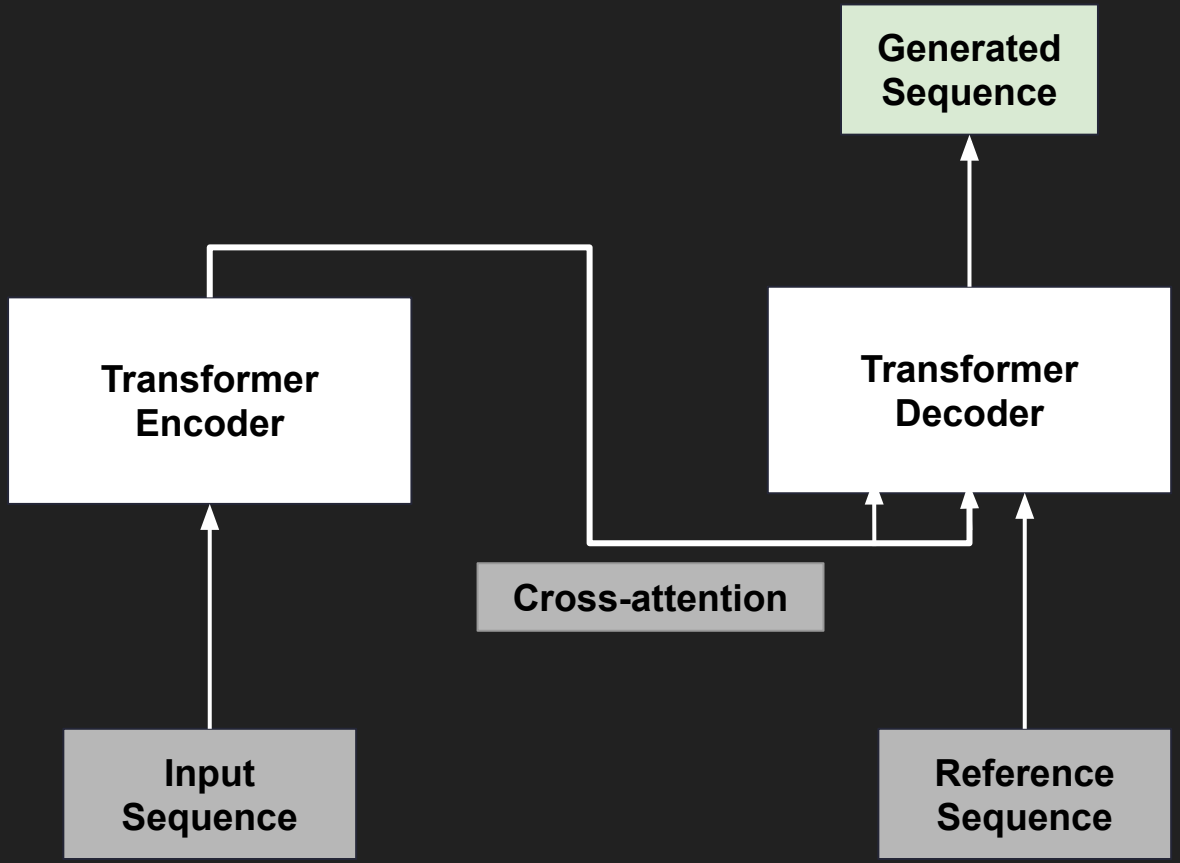

Expert Sentence
Laymen Sentence
Sequence of words
Sequence of words

Transformers [Vaswani et. al., 2017]

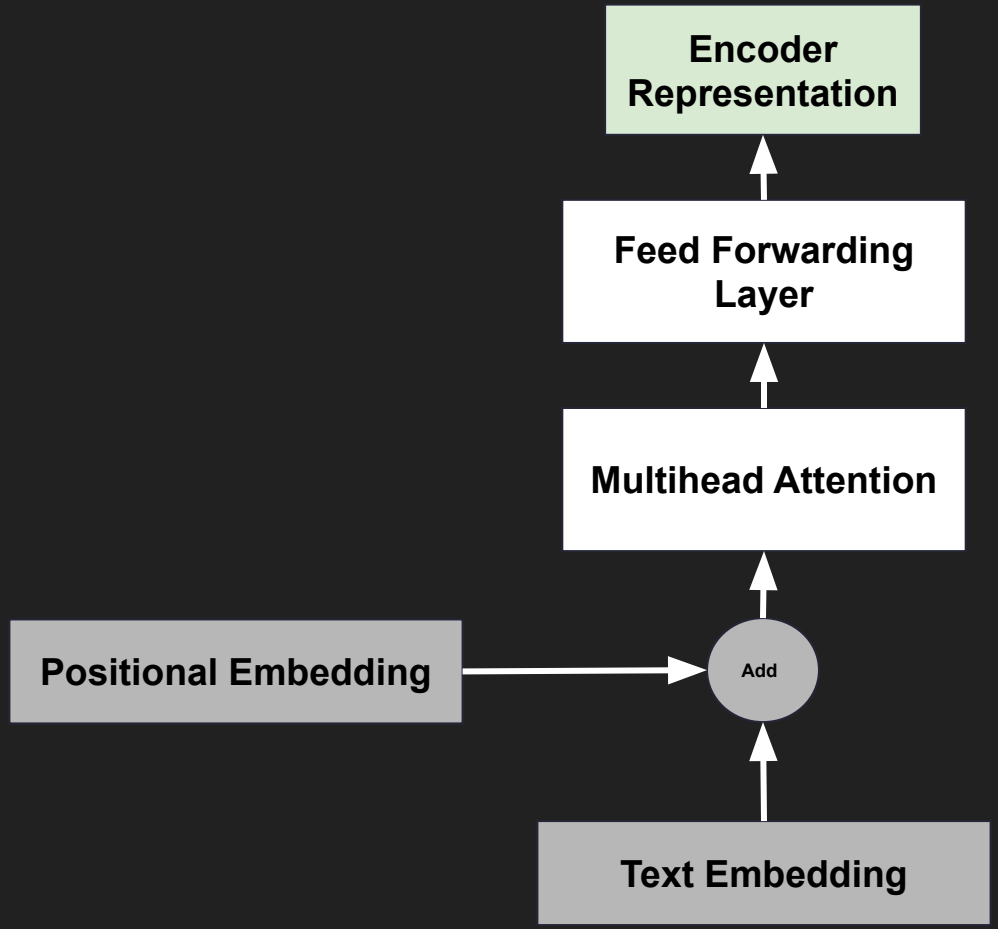
Proposed in the paper “Attention is all you need”

- ➔ Cross Attention
- ➔ Multihead Attention
- ➔ Positional Encoding
- ➔ Feed Forwarding Layer

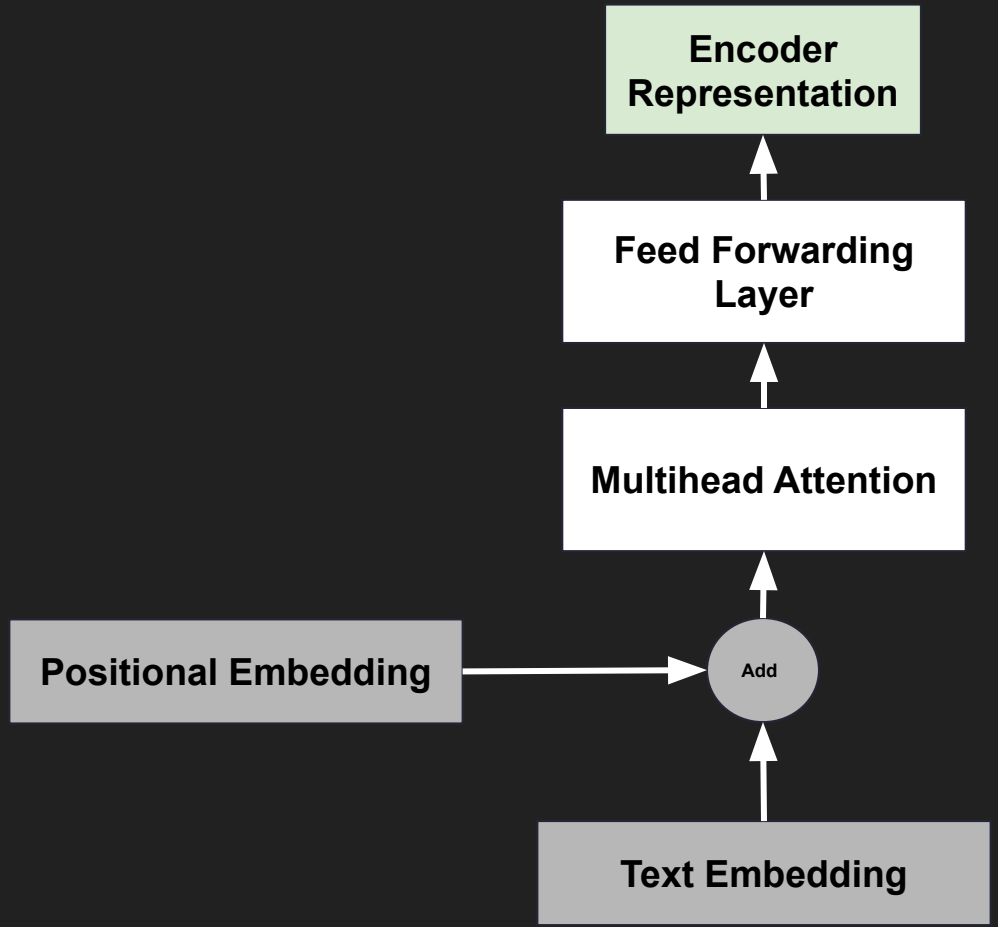
Transformer Architecture



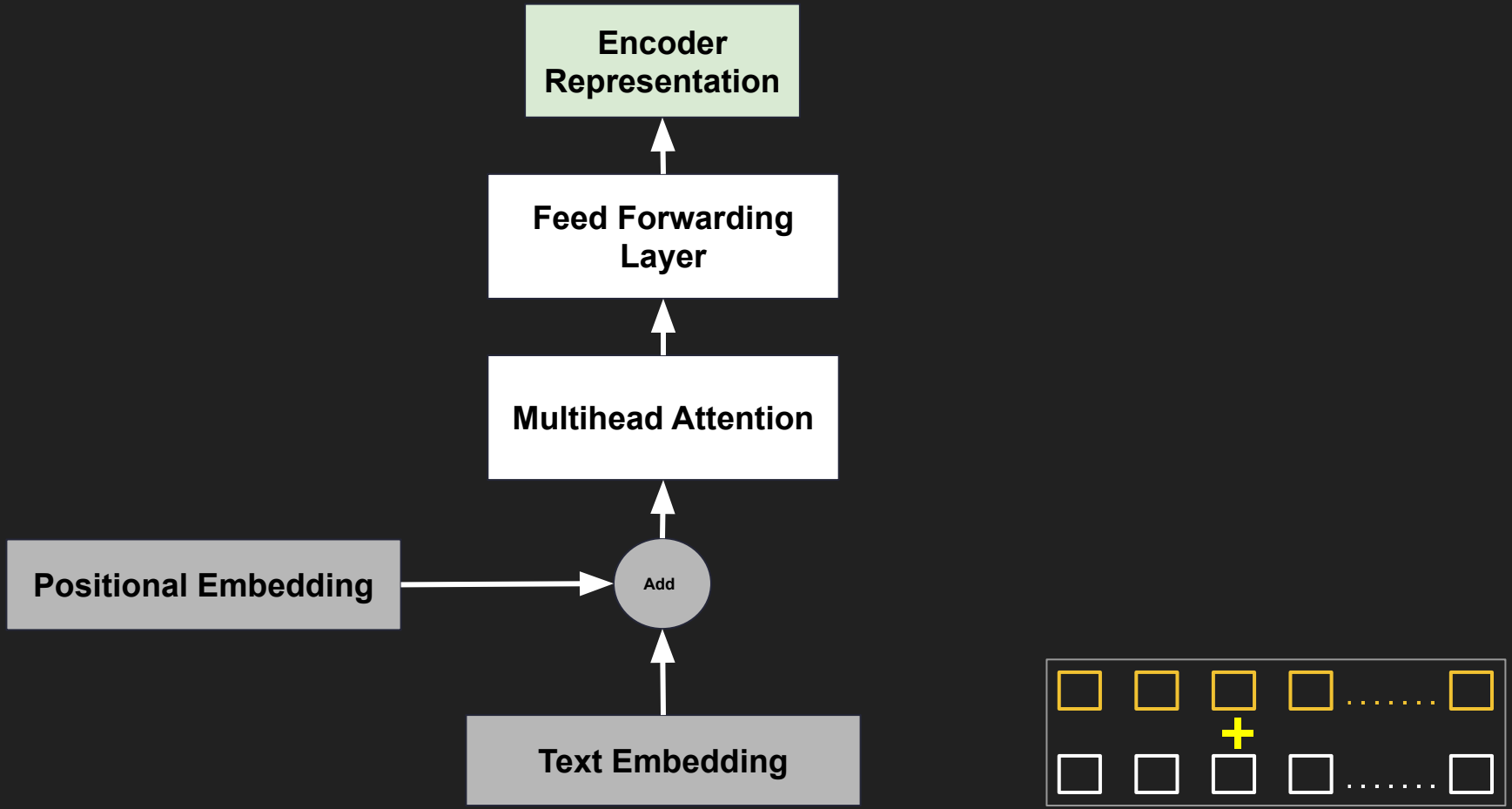
Transformer Encoder



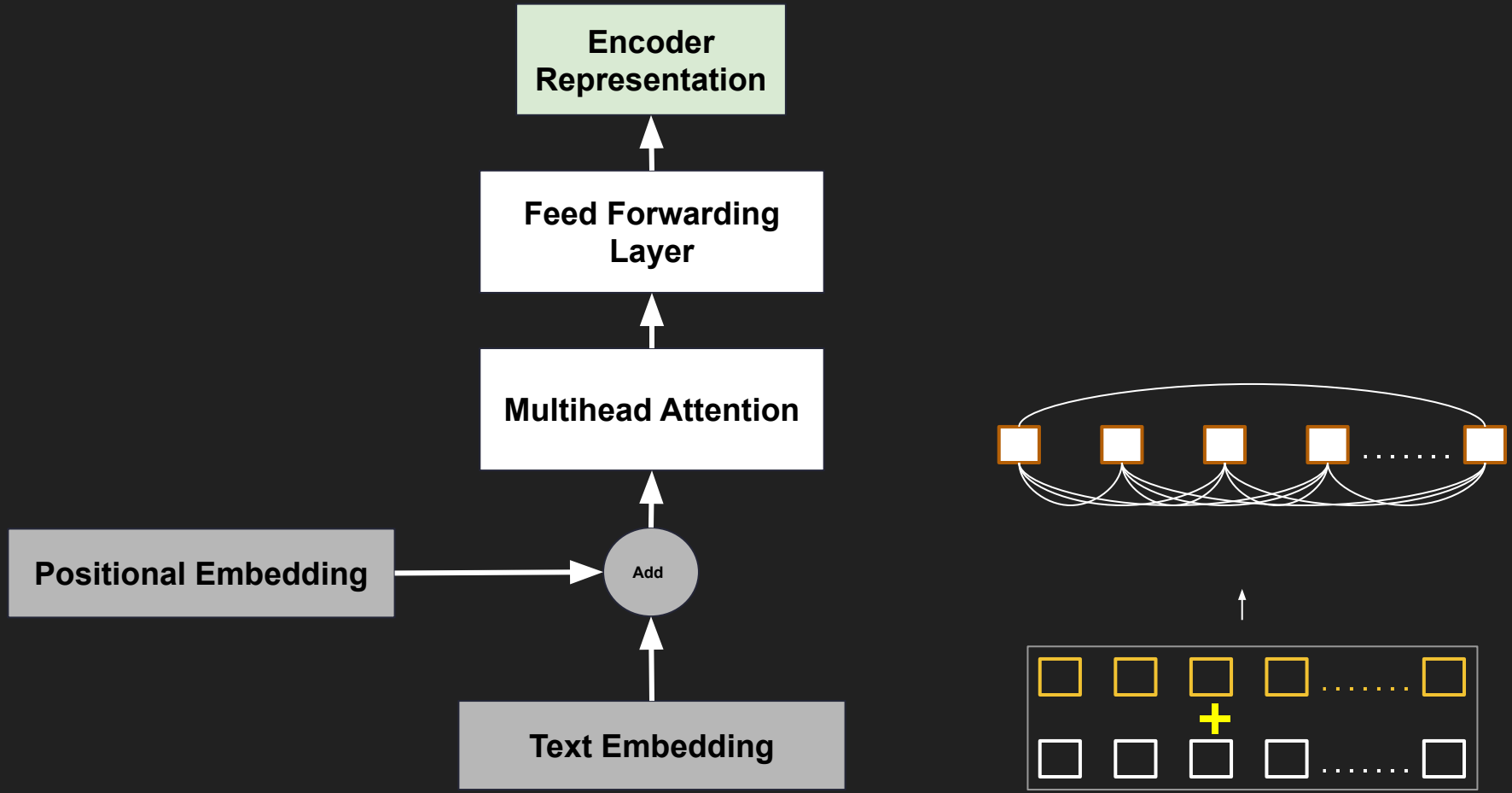
Transformer Encoder



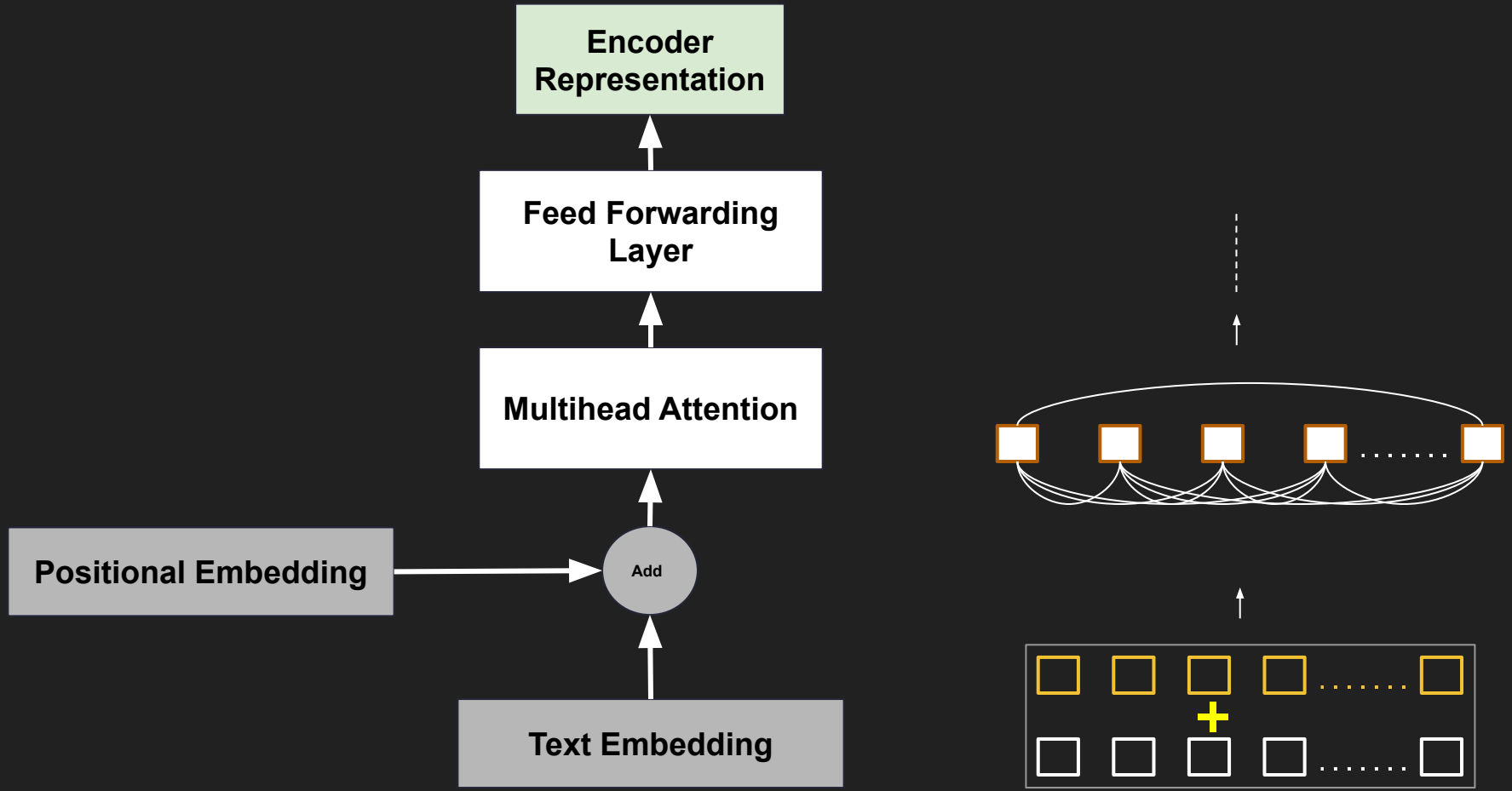
Transformer Encoder



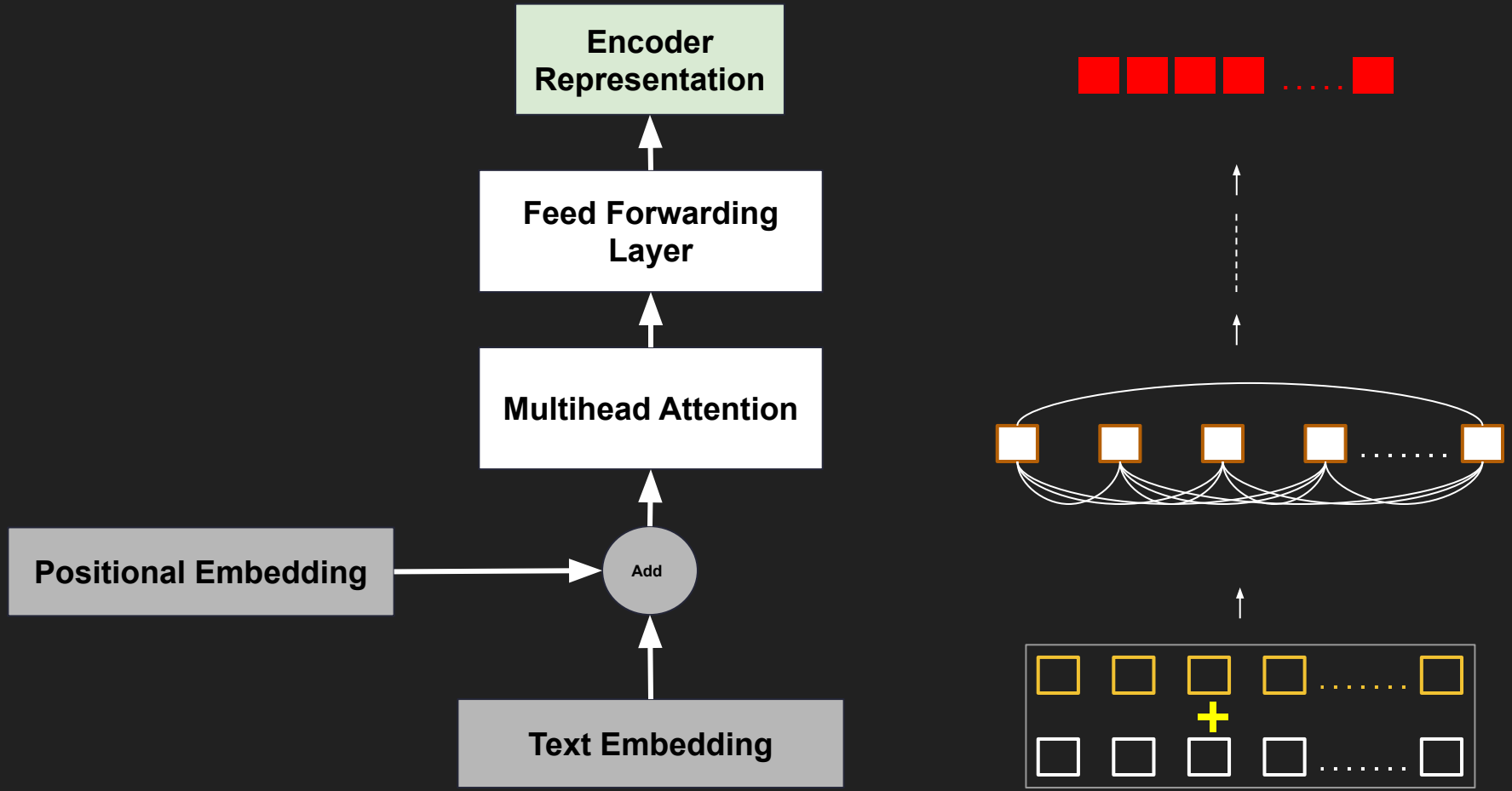
Transformer Encoder



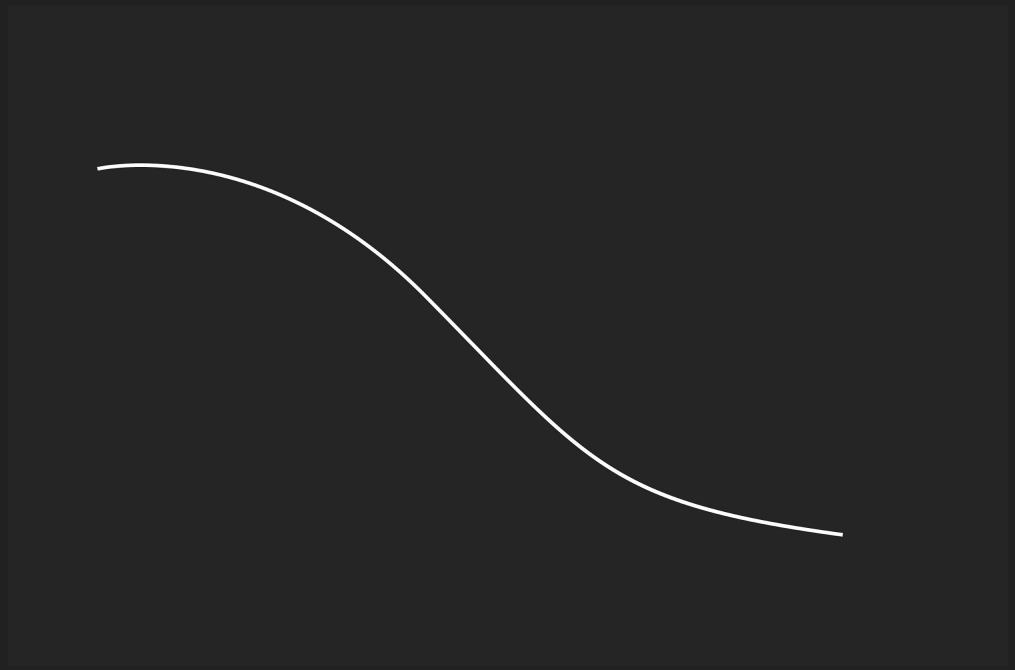
Transformer Encoder



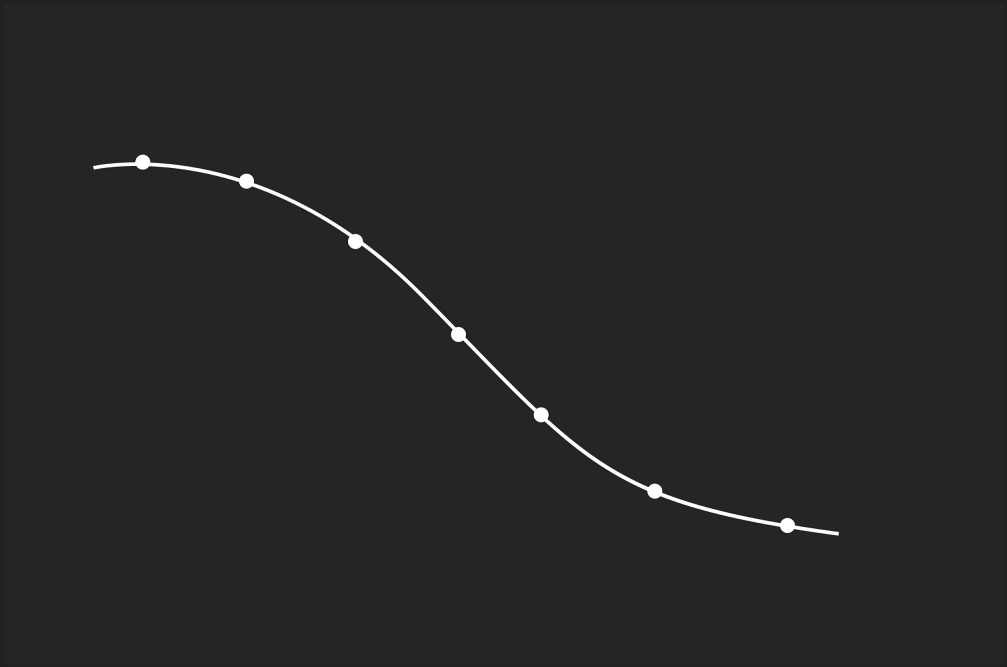
Transformer Encoder



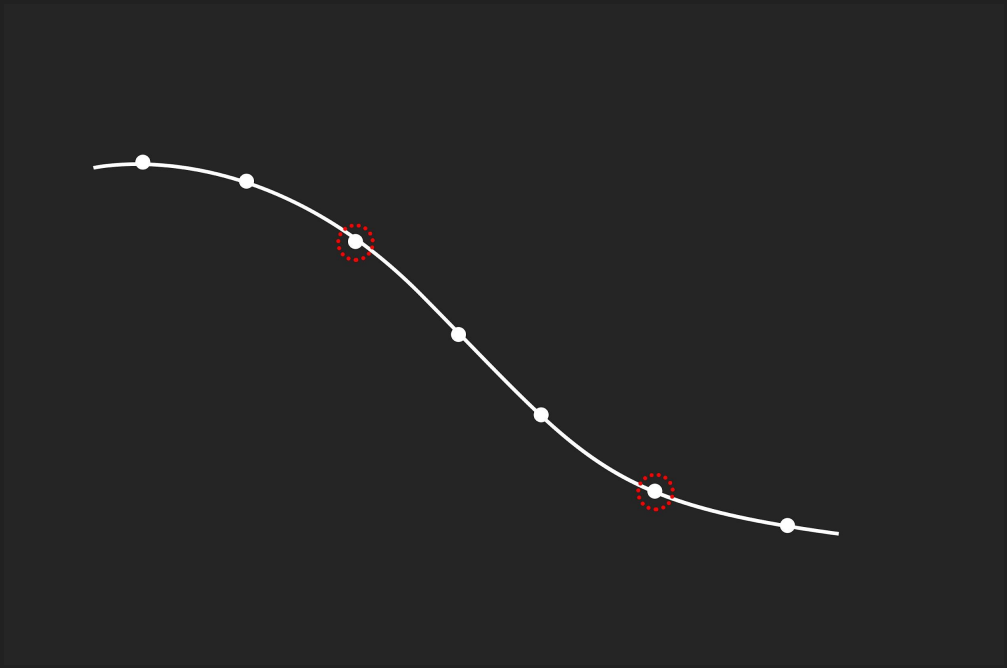
Denoising Autoencoder [Vincent et. al., 2008]



Denoising Autoencoder [Vincent et. al., 2008]

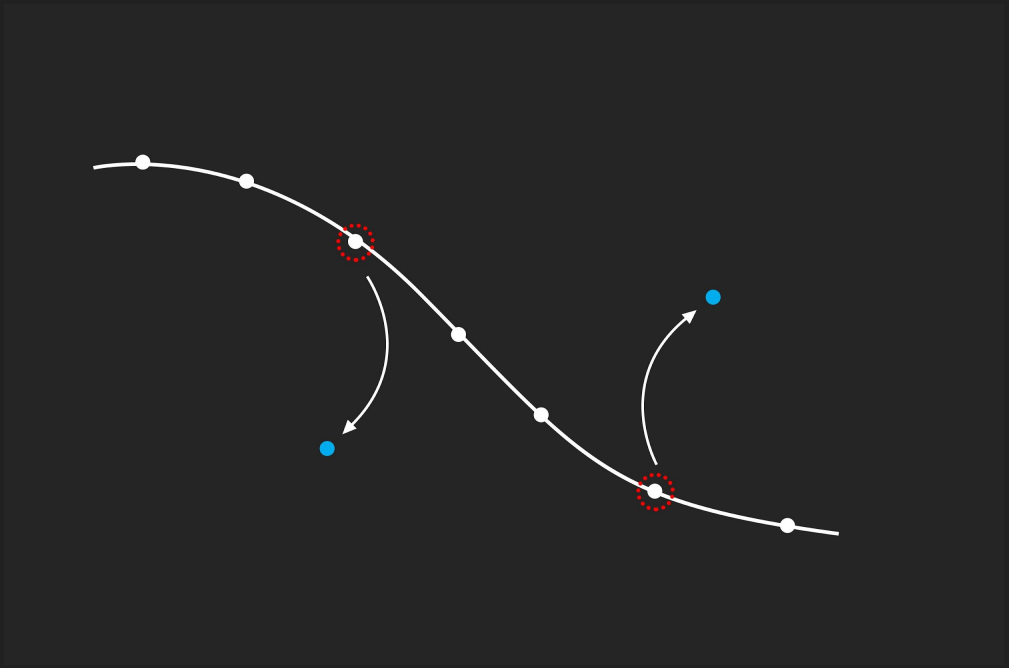


Denoising Autoencoder [Vincent et. al., 2008]



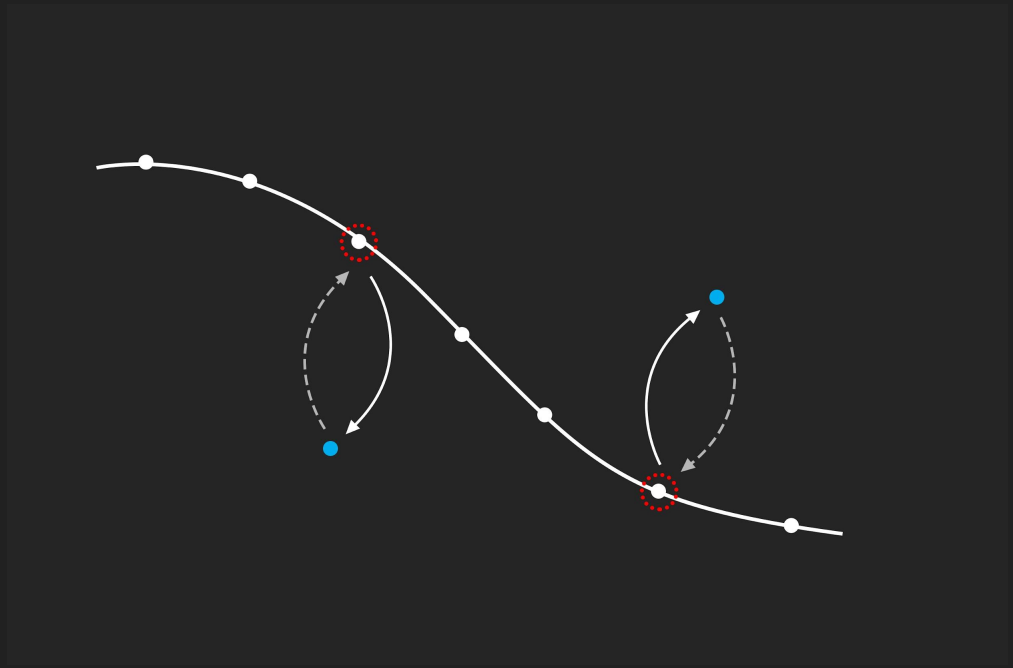
➔ Randomly choose some points

Denoising Autoencoder [Vincent et. al., 2008]



- ➔ Randomly choose some points
- ➔ Add noise to corrupt the signal

Denoising Autoencoder [Vincent et. al., 2008]



- ➔ Randomly choose some points
- ➔ Add noise to corrupt the signal
- ➔ Reproduce a clean repaired signal

Denoising Autoencoder for Text [Devlin et. al., 2018]

Denoising Autoencoder for Text [Devlin et. al., 2018]

Denoising Autoencoder for Text [Devlin et. al., 2018]

Denoising Autoencoder for Text [Devlin et. al., 2018]

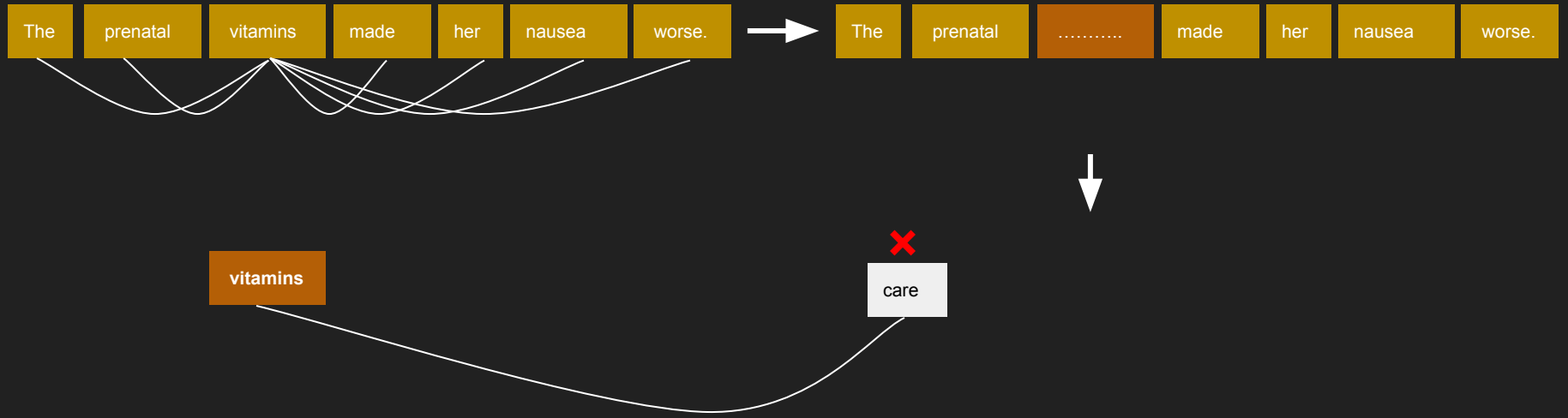
Denoising Autoencoder for Text [Devlin et. al., 2018]

Denoising Autoencoder for Text [Devlin et. al., 2018]

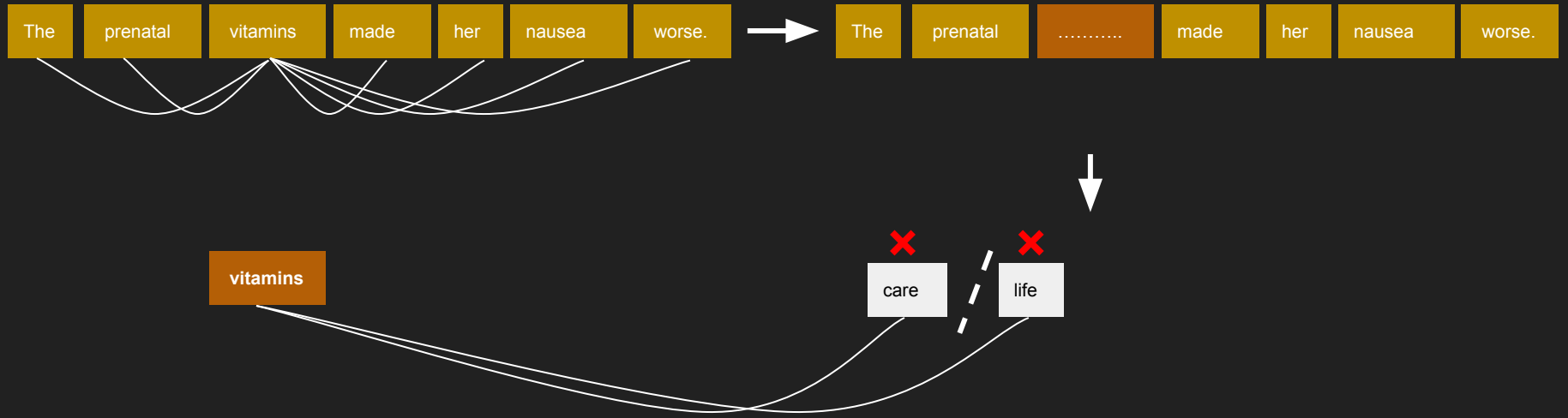
Denoising Autoencoder for Text [Devlin et. al., 2018]

Denoising Autoencoder for Text [Devlin et. al., 2018]

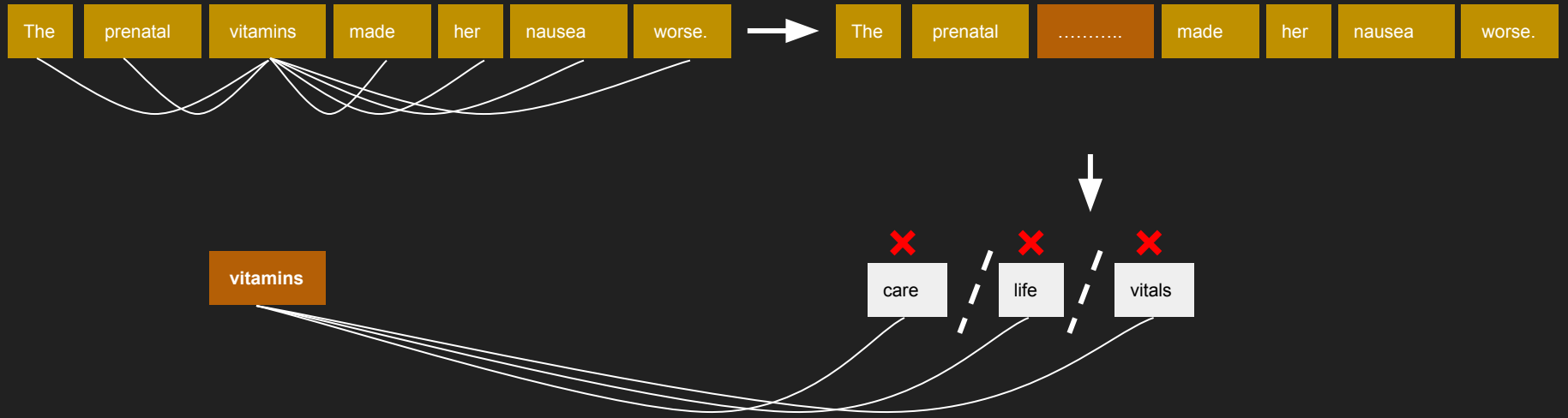
Denoising Autoencoder for Text [Devlin et. al., 2018]



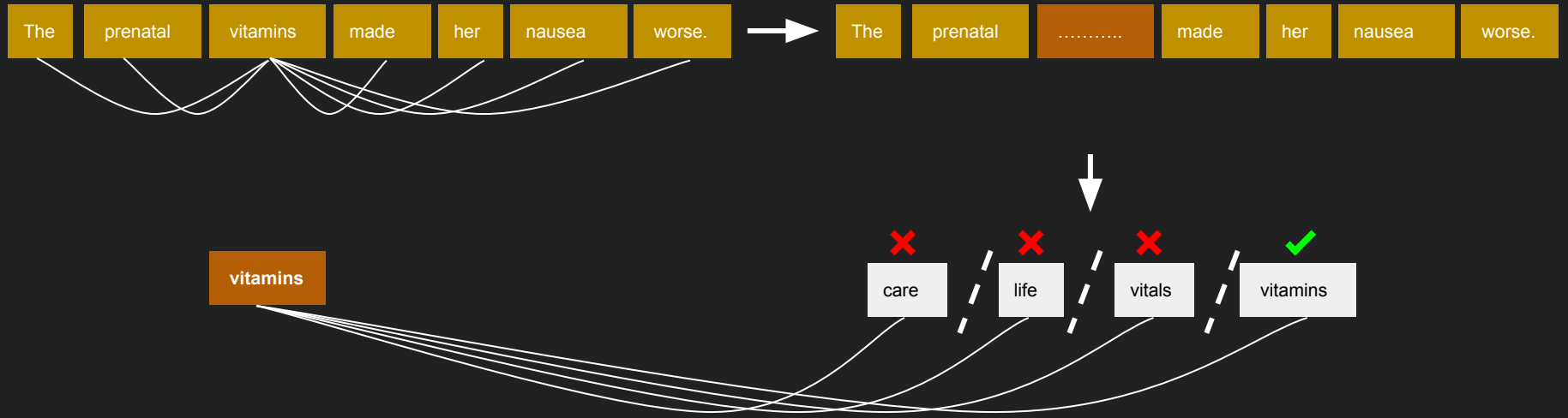
Denoising Autoencoder for Text [Devlin et. al., 2018]



Denoising Autoencoder for Text [Devlin et. al., 2018]



Denoising Autoencoder for Text [Devlin et. al., 2018]



Dataset

MSD Dataset [Cao et. al., 2020]

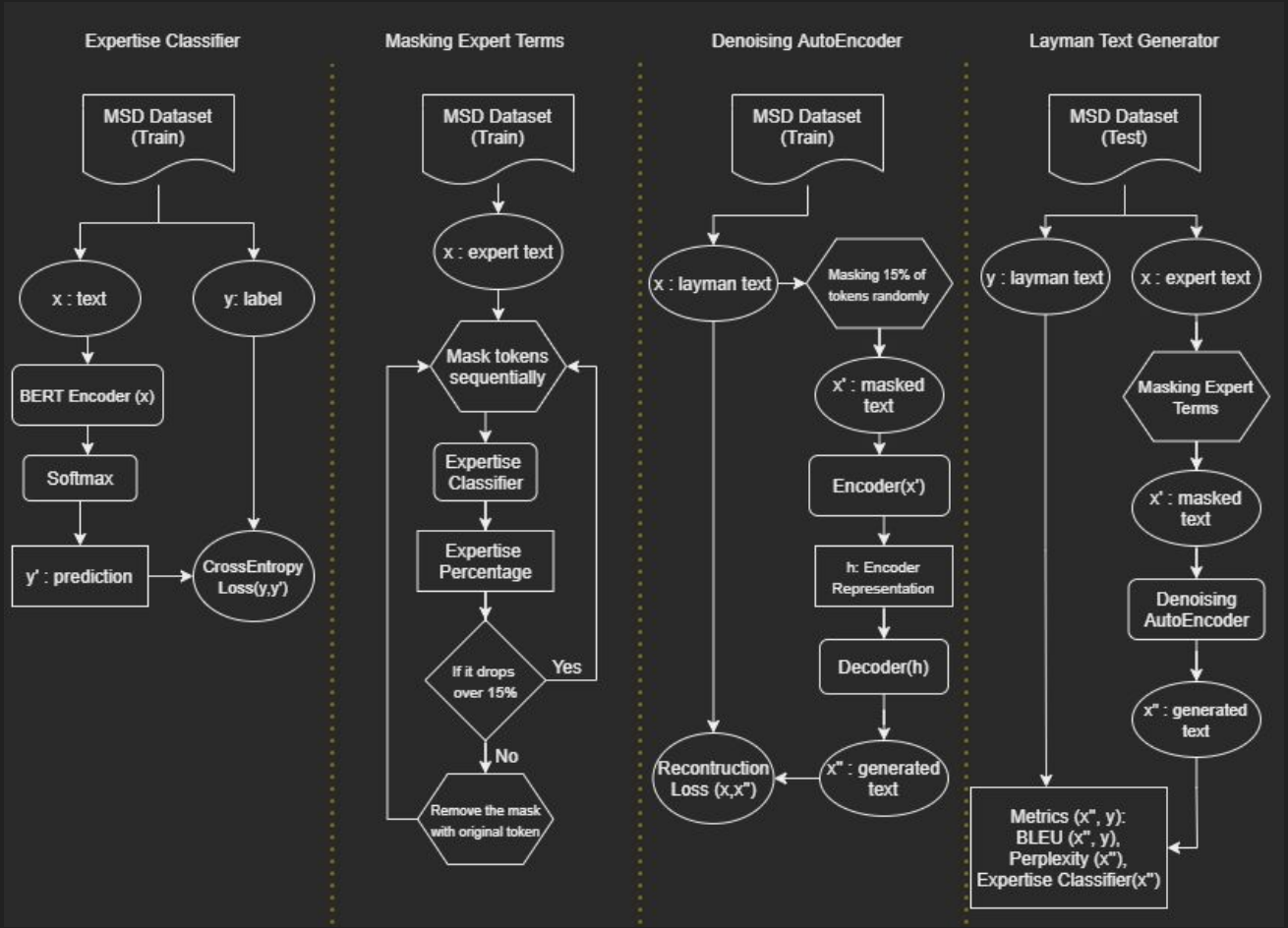
- MSD stands for Merck, Sharp & Dohme, founders of Merck & Co. , one of the largest pharmaceutical companies in the world
- Collected from Merck Manuals (for doctors and consumers)
- Train data was annotated by 3 doctors (domain expert) and parallel sentences of the test data was provided by another doctor.
- 245023 non-parallel sentences in expert and layman styles.
- 1450 parallel sentences in expert and layman styles.

MSD Dataset (cont.)

Text	Style	Concepts
Myocardial fibrosis , left ventricular hypertrophy , and cardiomyopathy can develop .	Expert	[{"range": [0, 2], "term": "myocardial fibrosis", "cui": ["C0151654"]}, {"range": [3, 6], "term": "left ventricular hypertrophy", "cui": ["C0232306"}, {"range": [8, 9], "term": "cardiomyopathy", "cui": ["C0878544"]}]
Chronic use can also damage the heart , causing scarring and thickening of the heart muscle and eventually leading to heart failure .	Laymen	[{"range": [0, 1], "term": "chronic", "cui": ["C1555457"]}, {"range": [6, 7], "term": "heart", "cui": ["C0018787"]}, {"range": [9, 10], "term": "scarring", "cui": ["C0008767"]}, {"range": [14, 15], "term": "heart", "cui": ["C0018787"]}, {"range": [15, 16], "term": "muscle", "cui": ["C4083049"]}, {"range": [20, 22], "term": "heart failure", "cui": ["C0018802"]}
For confirmation , selected noninvasive and invasive cardiac tests are usually done.	Expert	{"range": [1, 2], "term": "confirmation", "cui": ["C1611825"]}, {"range": [6, 7], "term": "invasive", "cui": ["C1334278"]}, {"range": [7, 9], "term": "cardiac test", "cui": ["C4529960"]}, {"range": [10, 11], "term": "usually", "cui": ["C3888388"]}

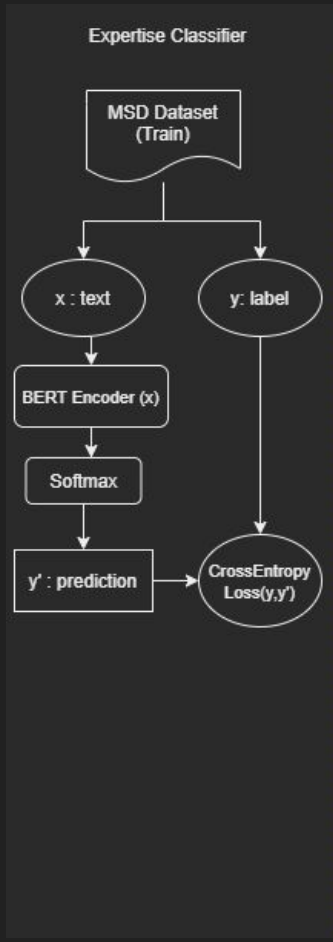
Proposed Approach

Proposed Methodology

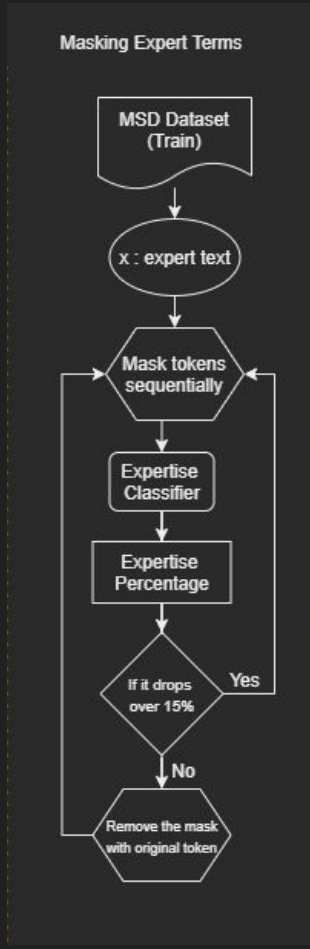


- ➔ BERT is pre-trained on Wikipedia corpus.
- ➔ We fine-tune BERT encoder model for text classification.
- ➔ We take the texts as input and styles as labels from the MSD Dataset.
- ➔ As loss function, we use Cross Entropy Loss.

$$\text{Binary Cross Entropy Loss} = -(y_i * log(p_i) + (1 - y_i) * log(1 - p_i))$$

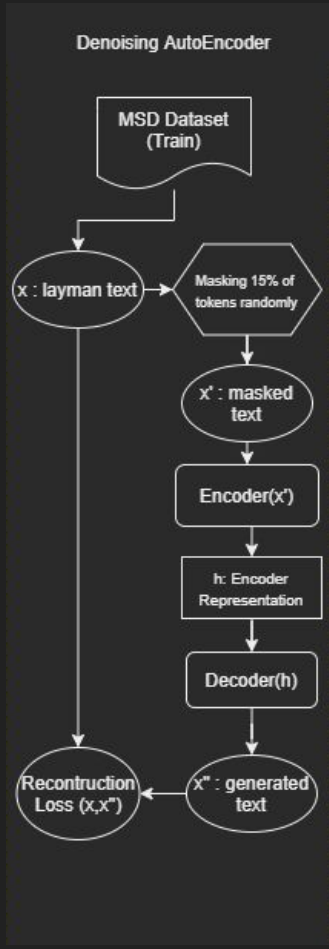


1. Get an expert text.
2. Mask tokens one by one.
3. Get the expertise percentage using expertise classifier.
4. Does the percentage drop more than 15% ?
    a. If yes, keep the mask and go to step 2.
    b. If no, replace the mask with original term and go to step 2.
5. Stop when all the terms are checked.



- We train a denoising autoencoder (DAE) on layman text corpus.
- While training, we
  - Mask out 15% of the terms
  - Get the masked text (x') to denoising autoencoder as input
  - Get the generated text (x'')
  - Keep updating the weights with respect to Reconstruction Loss between the original layman (x) and generated layman (x'').

$$\text{Recontruction Loss} = \frac{1}{2} * ||x - x''||^2$$

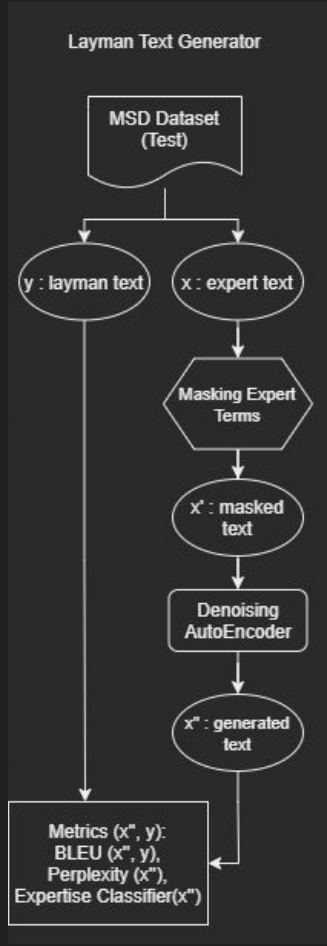


01. Take an expert text (x)

02. Mask the expert terms and get masked text (x’)

03. Use the trained DAE to predict the masked terms and generate layman text (x’’)

04. We use BLEU, Perplexity and Style Accuracy as metric to evaluate our work while x’’ is the generated layman text and y is the reference layman text.



Experimental Analysis

Examples of Generated Layman

Success

Expert

Masked

Generated

Reference

Urea breath testing and stool antigen testing .

Examples of Generated Layman

Success

Expert

Urea breath testing and stool antigen testing .

Masked

[MASK] breath testing and stool [MASK] testing .

Generated

Reference

Examples of Generated Layman

Success

Expert	Urea breath testing and stool antigen testing .
Masked	[MASK] breath testing and stool [MASK] testing .
Generated	Breath testing and stool testing .
Reference	

Examples of Generated Layman

Success

Expert	Urea breath testing and stool antigen testing .
Masked	[MASK] breath testing and stool [MASK] testing .
Generated	Breath testing and stool testing .
Reference	Testing of breath and stool

Examples of Generated Layman

Success

Expert	Urea breath testing and stool antigen testing .	Often , there is no visible oro pharyngeal inflammation .
Masked	[MASK] breath testing and stool [MASK] testing .	
Generated	Breath testing and stool testing .	
Reference	Testing of breath and stool	

Examples of Generated Layman

Success

Expert	Urea breath testing and stool antigen testing .	Often , there is no visible oro pharyngeal inflammation .
Masked	[MASK] breath testing and stool [MASK] testing .	Often , there is no visible [MASK] [MASK] inflammation .
Generated	Breath testing and stool testing .	
Reference	Testing of breath and stool	

Examples of Generated Layman

Success

Expert	Urea breath testing and stool antigen testing .	Often , there is no visible oro pharyngeal inflammation .
Masked	[MASK] breath testing and stool [MASK] testing .	Often , there is no visible [MASK] [MASK] inflammation .
Generated	Breath testing and stool testing .	Often , there is no visible throat inflammation .
Reference	Testing of breath and stool	

Examples of Generated Layman

Success

Expert	Urea breath testing and stool antigen testing .	Often , there is no visible oro pharyngeal inflammation .
Masked	[MASK] breath testing and stool [MASK] testing .	Often , there is no visible [MASK] [MASK] inflammation .
Generated	Breath testing and stool testing .	Often , there is no visible throat inflammation .
Reference	Testing of breath and stool	Often , there is no visible inflammation in the throat.

Examples of Generated Layman

Failure

Expert	Night vision may eventually be lost .
Masked	[MASK] [MASK] may eventually be lost .
Generated	The patient may eventually be lost .
Reference	Seeing at night or in the dark or dim light becomes more difficult .

Examples of Generated Layman

Failure

Expert	Night vision may eventually be lost .	Nausea , vomiting , constipation , severe prostration , restlessness , and irritability are common .
Masked	[MASK] [MASK] may eventually be lost .	[MASK] , vomiting , constipation , severe [MASK] , restlessness , and irritability are common .
Generated	The patient may eventually be lost .	Vomiting , constipation , severe fatigue , restlessness, and, irritability are common .
Reference	Seeing at night or in the dark or dim light becomes more difficult .	Nausea , vomiting , constipation , extreme fatigue , irritability , and restlessness are common .

Experimental Setup

- We did all our experiments using Google Colab which is a hosted Jupyter notebook service. We used it because Google Colab provides free GPU for 12 hours a day.
- In our experiments we used Numpy, Pandas, etc. for data processing and PyTorch for training and testing. PyTorch is an open source machine learning framework.
- In the MSD dataset, there were 245023 non-parallel sentences in expert and layman styles as train set and 1450 parallel sentences test set. We chose 10% randomly as validation set from the MSD train dataset.
- Training details
  - Expertise Classifier =\ [Duration : 5 hours] [10 epochs] [110M parameters] [Acc : 90.55]
  - Masking Expert Terms =\ [Duration : 10 minutes]
  - Denoising AutoEncoder =\ [Duration : 11 hours] [10 epochs] [120M parameters]
  - Layman Text Generator =\ [Duration : 30 minutes]

Evaluation Metrics

Style transfer approaches are evaluated by Style Accuracy, Content Similarity (BLEU) and Perplexity

Style Accuracy :

We trained an expertise classifier leveraging MSD dataset which we used to verify if the generated sentence is layman or not.

BLEU (Bi-Lingual Evaluation Understudy) : [Papineni et. al., 2002]

We used tetra-gram (sequence of 4 consecutive words) BLEU score for this evaluation. To measure this, we check how many times all unique overlapping consecutive sequence of 4 words of the generated sentence appeared in target laymen sentence and divide it with number of tetragrams.

$$p_n = \frac{\sum_{S \in C} \sum_{ngram \in S} Count_{matched}(ngram)}{\sum_{S \in C} \sum_{ngram \in S} Count(ngram)}$$

Perplexity : Checks the grammatical correctness of the generated texts.

$$\text{perplexity} = \prod_{t=1}^{T} \left(\frac{1}{P_{\text{LM}}(\boldsymbol{x}^{(t+1)} \mid \boldsymbol{x}^{(t)}, \ldots, \boldsymbol{x}^{(1)})} \right)^{1/T}$$

Result Analysis

Approaches	Style Accuracy	BLEU	Perplexity
Delete And Retrieve	74.67	2.95	3.92
ControlledGen	11.70	**13.13**	5.97
(Ours)	**78.76**	8.17	**3.12**

- ➔ Our approach works better than the previous baseline models in style accuracy and perplexity.
- ➔ Our approach works better than Delete and Retrieve (D&R) but worse than ControlledGen in terms of Content Similarity.
- ➔ **All the experimentations were done using same train-test split of MSD Dataset.**

Conclusion and Future Plans

- We proposed a new approach which performs stable across all evaluation metrics.
- More hyperparameter tuning and training time needed to get to much better result across all metrics.
- In future, we plan to extend the work to transfer laymen style text to expert style text also.
- We also plan to perform further human evaluation after training and testing due to the deficiency of proper evaluation metrics.

References

1. Yixin Cao, Ruihao Shui. “Expertise Style Transfer: A New Task Towards Better Communication between Experts and Laymen”. in ACL 2020
2. ZHIQIANG HU. “Text Style Transfer: A Review and Experiment Evaluation”. preprint 2020
3. Zhiting Hu, Zichao Yang. “Toward Controlled Generation of Text”. in ICML 2017
4. Jaucen Li, Robin Jia. “Delete, Retrieve, Generate: A Simple Approach to Sentiment and Style Transfer”. in NAACL 2018.
5. Fuli Luo, Peng Li. “A Dual Reinforcement Learning Framework for Unsupervised Text Style Transfer”. in IJCAI 2019.
6. Ashish Vaswani, Noam Shazeer. “Attention is all you need”. in NeurIPS 2017
7. Jacob Devlin, Ming-Wei Chang. “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding”. in NAACL-HLT 2018.
8. Olivier Bodenreider, “The Unified Medical Language System (UMLS): Integrating Biomedical Terminology”. in Nucleic Acids Research 32, 2004.