Multi-Modal Human Cognitive State Recognition during Reading

Nikita Filimonov

Lomonosov Moscow State University
filimonovn160@gmail.com

Abstract. Human cognitive state recognition is an important and challenging task. Various registration technologies can be used to collect physiological data that potentially contain relevant information regarding current cognitive state of a human subject. Oculography (eye-tracking) and electroencephalography (EEG) are most popular and well-researched registration technologies. Both technologies have cheap commercial variants that do not require laboratory equipment or involvement of professional physiologist to collect the data. However, it is still problematic and expensive to obtain large-scale datasets of physiological data of such sort. In this work a review and analysis of available open source physiological data is provided. A task of natural reading is considered since work of eyes and human brain during reading are of great interest for cognitive science in combination with machine-learning. A multi-modal approach that involves combining EEG and eye-tracking data in jointly trained artificial neural network is proposed. Intermediate results are presented regarding encoding EEG signals with Variational Auto-Encoder (VAE).

Keywords: Eye tracking · Electroencephalography · Artificial neural networks.

1 Introduction

This work[1] is aimed at neural network architecture development for assessing the cognitive state of a person while working with a text. Term cognitive states, in this work, does not mean the emotional states of a person while performing a certain work, but which parts of the brain and how they respond to certain stimulus, in this case, to the information contained in the text. The first part of this work is devoted to the overview of feature selection approaches and the study of methods of data preprocessing. In this work, the following data will be used: Electroencephalograms (EEG), eye-tracking and vectorized text sets. An EEG is a collection of electrical signals registered on the brain. Generally speaking, EEG is a method for studying the functional state of the brain. Mathematically,

[1] The work is performed as a master thesis at the master program "Big Data: Infrastructures and Methods for Problem Solving", Faculty of Computational Mathematics and Cybernetics, Lomonosov Moscow State University.

Copyright © 2021 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR-WS.org/Vol-3036/paper11.pdf

EEG data is represented by an $M \times N$ matrix, where M is the number of channels in the device that was used to record signals, and N is the length of the record. Eye-Tracking data is the result of registration process of the patient's eye movement while reading text, for example. A dataset is a set of coordinates where a person's eyes were focused at a fixed frequency. The next stage of the work is architecture development. Its development will directly depend on what features will be selected and how they will be grouped in the dataset. Feature selection is an important topic because the area of the research is an intersection of neurophysiology and data science, which leads to certain difficulties with feature engineering. In this research not only with mathematical quality metrics, but also with medical ones will be taken in count. Medical metrics mean that at the stages of data processing and preparation, the correspondence of the selected metrics to reality will be checked, and the features themselves will be extracted in accordance with the observations of experts. For example, EEG data have been studied for a long time, and a lot of patterns, how the brain reacts to certain stimulus, are already known and how these patterns look after certain transformations. Later, according to the obtained data, it will be possible to build hypotheses and conclusions, and most importantly, correctly perform batch sampling during training neural-networks. A significant part of research in related topics is focused only on single type of physiological data, either eye-tracking features or electroencephalography features. This work proposes an approach that combines types of features together with semantic information extracted from text. Future development of this work will include training and evaluation of proposed approach on various datasets, including open-source datasets. Apart from deep neural network architecture the work includes the following steps: data preprocessing, feature extraction and feature selection steps. As a final step, a quality metric of the neural network in cognitive state recognition task will be developed. This metric could be presented as hypothesis which requires further test and designing on experiments.

The rest of this paper is organized as follows: Section 2 discusses related works, Section 3 describes the dataset and data preprocessing issues, the approach is proposed in Section 4, and preliminary results are described in Section 5.

2 Related Work

A significant amount of work had been dedicated to feature extraction, selection and preprocessing. Usually, to solve classification, regression and all other problems with EEG or eye tracking data word-level characteristics. The following features are usually extracted from raw word-level data – number of fixations, mean fixation duration, gaze duration, number of fixations on word, the sum of all fixations on the current word in the first-pass reading before the eye moves out of the word, total reading time (TRT), the sum of all fixation durations on the current word, first fixation duration (FFD), the duration of the first fixation

on the prevailing word, go-past time (GPT), the sum of all fixations on the right of the current word.

According to [1], both monolingual and multilingual models achieve high accuracy in predicting a range of eye tracking features across four languages. Comparison of performance of language-specific and multilingual pretrained transformer models in regression task was provided. Main task of this work was to predict eye-tracking features in an experiment where participants were reading texts on Dutch, English, German, and Russian languages. According to results, Bidirectional Encoder Representations from Transformers (BERT) [2] and Cross-Lingual Language Model (XLM) [3] models show the best performance on restoring eye-tracking data, meanwhile XLM models require less data to fine-tune. In [4] used EEG features to supervise machine attention. Also, it was shown that cropping data with random forest-splits won't reduce the model accuracy but will considerably reduces the number of dimensions of the EEG data. They used Bidirectional Long-Short Term Memory (BiLSTM) model with attention mechanism. Not only EEG features can be used to tune attention weights in neural-networks, but eye-tracking features also. In [4] human attention derived from eye-tracking data were used to regularize attention functions in recurrent neural networks. First note that the baseline models only attend to one or two coherent text parts. These conclusions could be made from this paper – baseline models mainly focus on stop-words, rather than on gaze or fixation information and that the regularization made by human attention, learned from eye-tracking data enables neural-networks (bidirectional LSTM in this case) to learn to better focus on the most relevant aspects of sentences for the target tasks.

3 Dataset

Due to expensive data collection process and necessity of involvement of qualified physiologists, a very limited amount of data is available for research community. This work focuses on freely available ZuCo and ZuCo-2 datasets [5] [6].The first version of the datasets appeared many times in various research works while the second version of the dataset has different semantic specification. In ZuCo2 data had been recorder from 19 participants, but one of them had technical problems with the recording, thus dataset consists of recordings collected from 18 participants.

In the experiment from ZuCo-2 dataset, the participants had to read 739 sentences that were selected from the Wikipedia dataset. The corpus provides semantic annotations of semantically different tasks. Normal Reading task consists of a random set of Wikipedia sentences. Following topics had been selected for the task-specific reading part: political affiliation, education, founder, wife/husband, job title, nationality, employer. The sentences have the same length as ZuCo 1.0, with similar semantic. In Normal reading task, the participants had to read 349 sentences, and 390 sentences in a task-specific reading task. Furthermore, there is also an overlap in the sentences between ZuCo 1.0 and ZuCo 2.0. 100 normal reading and 85 task-specific sentences from ZuCo-2 dataset were already

recorded in ZuCo 1.0. This provides an opportunity to compare different recording procedures (i.e. session-specific effects) and perform studies on larger number of participants (subject-specific effects).

Eye-tracking device captures eye position and pupil size. Records had been made at a sampling rate of 500 Hz with EyeLink 1000 Plus, SR Research device. The eye tracker was calibrated with a 9-point grid at the beginning of the session and re-validated before each block of sentences.

In ZuCo dataset, the following word-level features are extracted from eye tracking data:

1. X, Y coordinates of the fixation
2. Fixation durations
3. Gaze duration
4. Total reading time
5. Number of fixations on this word
6. Pupil size

EEG part of the dataset, contains data from 128-channels of raw data with sampling frequency 500 Hz. After the initial filtering and cleaning, that had been made by dataset maintainers, 23 channels has been removed and 105 has left. In the table below, there is a brief description of both task types in ZuCo-2 dataset.

Table 1. ZuCo-2 dataset description.

	Normal reading	Task-specific reading
Sentences	349	390
Sentence length	Mean: 19.6; range: 5 ... 53	Mean: 21.3; range: 5 ... 53
Total words	6828	8310
Word types	2412	2437
Word length	Mean: 4.9; range: 1 ... 29	Mean: 4.9; range: 1 ... 21

Normal reading (NR): Normal reading was the first task, participants had to read the sentences naturally, without any specific tasks or instructions. Task-specific reading (TSR) reading was the second, and the final task In task-specific reading participants had to read sentences with a clearly defined topic. For example, it could be political, economical, scientific related texts or questions on the same topics. Participants were instructed to search for a specific relation in each sentence they read, from the list of topics. Instead of comprehension questions, the participants had to decide for each sentence whether it contains the relation or not, also they were actively annotating each sentence. All sentences within one block involved the same relation type. The blocks started with a test round, which described the relation and was followed by three sample sentences, so that the participants would be familiar with the respective relation type. Event-Related Potentials were calculated for both tasks, based on fixation timestamps.

In a proposed approach event is a fixation on word, so the term fixation-related-potential (FRP) is used instead of ERP. In EEG, FRP is usually calculated using window from -600 ms before fixation, and 1 second after it [7].

$$\bar{x}(t) = \frac{1}{N}\sum_{k=1}^{N} x(t,k) = s(t) + \frac{1}{N}\sum_{k=1}^{N} n(t,k), \tag{1}$$

where N is the length of interval, k is event number, t is time passed after the k-th event, i.e. length of time interval, $s(t)$ is expected value of the signal, and $n(t,k)$ is noise. Fixation-related potentials are extracted from training data because they contain significant information about cognitive human reaction on each word, and in total the information about each sentence is obtained. In other words, FRP is an averaged incentives in the EEG signal, related to a certain action in real life. Therefore, information about human cognitive state during reading each sentence is contained in FRPs, excluding non-informative data between sentences and words. After extracting FRPs, alpha (8 - 12 Hz), beta (12 - 30 Hz), gamma (30 - 45 Hz), theta (4 - 8 Hz), delta (0.5 доби– 4 Hz) frequency bands are extracted. [7]. Extracting frequency bands is a regular method of EEG data preprocessing, because frequency bands reflect cognitive and memory performance. Furthermore, this approach can be used as a dimension reduction method, because raw EEG data is represented by highly correlated multidimensional time series.

4 Proposed Approach

4.1 Feature Extraction

Eye Tracking Features. Feature selection and preprocessing approaches are well studied topics in analysis of physiological data. Usually, to solve classification, regression and other tasks using EEG or eye tracking modalities word-level characteristics are utilized. The following features are usually extracted from raw word-level data (EEG and eye-tracking features for every word in corpus) [5] [6]:

- number of fixations - moments when eyes do not move, i.e fixated on something,
- mean fixation duration,
- gaze duration - sum of all fixations on the word, before fixation on another one (in seconds),
- number of fixations on word,
- total number of fixations on the current word in the first-pass reading before the eye moves out of the word,
- total reading time (TRT),
- the sum of all fixation durations on the current word,
- first fixation duration (FFD),
- the duration of the first fixation on the prevailing word,

– go-past time (GPT),
– the sum of all fixations on the right of the current word.

Among all described above features that could be extracted from eye-tracking data, gaze features are less researched. In the paper [8] it was studied how it is possible to use and properly preprocess gaze features in sequence labelling and sequence classification tasks. The gaze features can simply be concatenated to word-level features as multidimensional vectors representing each word. Several works [9] [10] showed that word-level averages of gaze features helped better than token-level features, i.e word-level features in vectorized representation. Using word-level gaze features does not require gaze at test time, e.g in test dataset or in experiments. The features can be used in the same way as word embeddings that are usually used and several studies also successfully concatenated type-level gaze features with pretrained word embeddings for a richer representation [11]. In the context of word embeddings — embedding is a vectorized representation of the text, whcih could be obtained using multiple techniques. Concerning the EEG embeddings — embedding mean the contraction mapping from one feature space to another. EEG data contains a lot of extra information which does not make sense in this work. Consequently, generating EEG embeddings is a part of feature reduction step.

Electroencephalography Features. A significant part of proposed work is planned to be focused on EEG data. Due to possibility of simultaneous registration of EEG and eye-tracking data, a variety of word-level brain activity signal feature extraction approaches can be utilized. In work [12] it was demonstrated that EEG of semantic processing can complement and enrich eye-tracking data. The eye tracking data provides millisecond-accurate fixation times for each word. Therefore, it is possible to obtain brain activity representations during each fixations of a word, and then extract event-related potentials on a given word. An event-related potential (ERP) is the response from brain, that is the direct result of a specific event, such as a fixation. More formally, it is any response to a stimulus [13]. In this work, 128 channels of EEG data will be used. Then, ERP's will be extracted from these 128 channels data. The study of the brain in this way provides a noninvasive means of evaluating brain functioning.In [14] used EEG brain activity data in NLP tasks: embeddings were generated on sentence level, sentence data had been padded to the size of maximum sentence. It is a common technique to work with multi-modal data, because EEG and eye-tracking data are represented by vectors and tensors of different size for each word. Also, it was shown that EEG data can improve performance on classification tasks in addition to usage BERT embeddings. As the final step, EEG embeddings will be generated from obtained fixation-related potentials.

EEG data can also be used to fine-tune attention mechanisms in NLP tasks. In [4] EEG features are used to supervise machine attention. Also, it was shown that cropping data with random forest-splits does not reduce model accuracy but considerably reduces the number of dimensions of the EEG data. Authors

used Bidirectional LSTM model with attention mechanism. Not only EEG features can be used to tune attention weights in neural-networks, but eye-tracking features also. In [10] human attention derived from eye-tracking data were used to regularize attention functions in recurrent neural networks.Baseline models mainly focus on stop-words, rather than on gaze or fixation information and that the regularization made by human attention, learned from eye-tracking data enables neural-networks (bidirectional LSTM in this case) to learn to better focus on the most relevant aspects of sentences for the target tasks.

4.2 Architecture

The purpose of this work is to propose the approach that could be applied to large scale datasets that consist of heterogeneous sequential data. This is the typical case of a neural networks application when it is problematic to apply popular statistical methods to achieve acceptable result and the size of dataset is counted in hundreds of gigabytes. Recurrent Neural Networks(RNNs) are especially suited for sequential data, such as eye-tracking or EEG data. The most popular type of RNN is LSTM, and its modification Bidirectional LSTM (BiLSTM), is usually used to work with eye-tracking data.

Powerful, but heavy and large architectures, such as BERT or XLM could be used to make predictions, based on eye-tracking data. Nevertheless, these models must be fine-tuned before using them, while BiLSTM [15] could be trained from scratch using much less data, than BERT and XLM requires [16].

In proposed approach, instead of using word embeddings and EEG features like in [15], embeddings generated from EEG features will be used, as well as word embeddings generated with GPT-2 BERT model and eye-tracking data embeddings. Eye-tracking embeddings will be taken from one of the final layers in the network, that will be used to predict next fixation, based on previous ones. A neural network will be trained to solve binary classification task: recognition of cognitive states related to normal reading and task specific reading. Input pipleline of the neural-network consists of fully-connected layers for each type of input features. These layers are used to perform dimension reduction and transform all data into the same shape for further concatenation. Than, the concatenated matrices are passed into BERT. BERT is a powerful architecture that is able to work well with long sequences, like in this case. Proposed network architecture is illustrated on Fig.1. The final goal could be divided into 3 subtasks:

1. Using the sequence of N previous fixations, word embeddings for these fixations and word embedding for the next word predict the probability of the fixation on next word. Take output layer activations as embeddings for eye-tracking data.
2. Generate embeddings for EEG frequency bands.
3. Using word embeddings, EEG embeddings, eye-tracking embeddings solve various classification tasks on ZuCo-2 corpora.

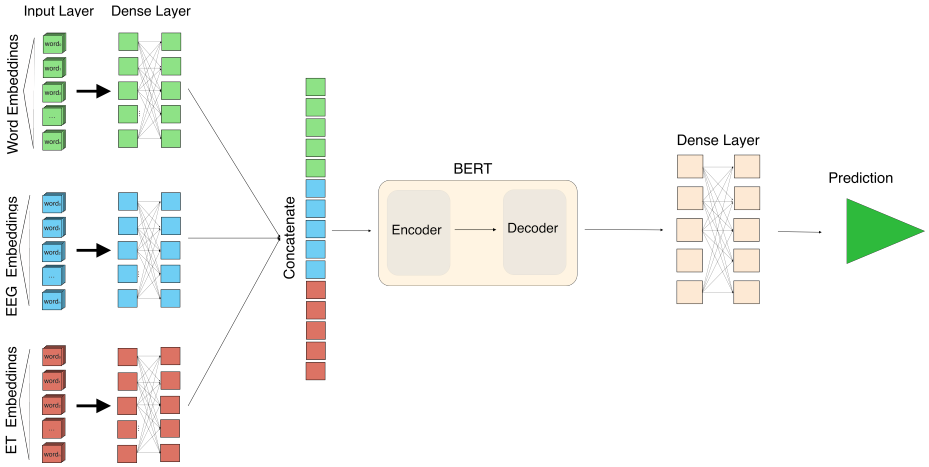

Fig. 1. Task-classification neural-network architecture

In [6] word embeddings and EEG data were used together to classify participant reading tasks from ZuCo dataset. Current paper extends this approach by focusing not only on word embeddings, but using them to generate embeddings from other parts of dataset.

5 Preliminary Results

As it was said above, embeddings of the EEG time series, eye-tracking features and work-level features will be used as an input batch to the neural-network. At present, only EEG input features had been prepared. A popular approach to generate embeddings is using autoencoders. In current work an approach that utilizes Variational Autoencoder (VAE) [17] hidden states as embeddings is proposed for EEG frequency bands data. Variation Autoencoder learns feature distribution parameters, transforms features to the hidden states vector, and then makes reconstruction of the original space. Since the input data is not a random set of time-series it could be described as a large mix of distributions with different parameters, and that is the reason why VAE was chosen. General idea of VAE is presented in the following equation:

$$\log P(x) - KL[Q(Z|X,\theta_1), P(X|Z,\theta_2)] =$$
$$\mathbb{E}_{Z\sim Q}[\log(P(X|Z,\theta_2)] - KL[Q(Z|X,\theta_1), P(Z)],$$

where X is an input data, Z is hidden states, $Q(Z|X,\theta_1)$ and $P(X|Z,\theta_2)$ are arguments of encoder and decoder functions consequently. In in a proposed approach LSTM layers were used in both encoder and decoder parts of VAE. Input for VAE is defined as following vector:

$$X = ((\alpha_0, \beta_0, \gamma_0, \theta_0, \delta_0), \ldots, (\alpha_{105}, \beta_{105}, \gamma_{105}, \theta_{105}, \delta_{105}))^T,$$

where $\alpha, \beta, \gamma, \theta, \delta$ are EEG frequency bands respectively. The main criteria in reconstruction - is to minimize reconstruction loss and make time-series generated by VAE realistic. Difference between original and reconstructed time-series. Even with the minimal batch size, it was impossible to train VAE to reconstruct time-series of the same smoothness as original. VAE had successfully reconstructed all peaks, and learned the structure of the spectre signal. But because reconstructed time series are not as smooth as the original ones, it was impossible to achieve good R^2 score. Scores were computed on full time-series, with 128 channels, while on Fig.2 only first 10 channels are shown. The results are presented in table 2.

Table 2. VAE reconstruction statistics

	mean squared error	mean absolute error	R^2
alpha	0.042	0.015	0.40
beta	0.033	0.011	0.60
gamma	0.037	0.012	0.70
theta	0.027	0.006	0.43
delta	0.031	0.007	0.44

To make sure that generated embeddings captures infromation from the original data well-enough, VAE reconstruction original data. On the charts below,

EEG frequency bands are presented. Left column is the original data, and reconstructed time-series are on the right column.

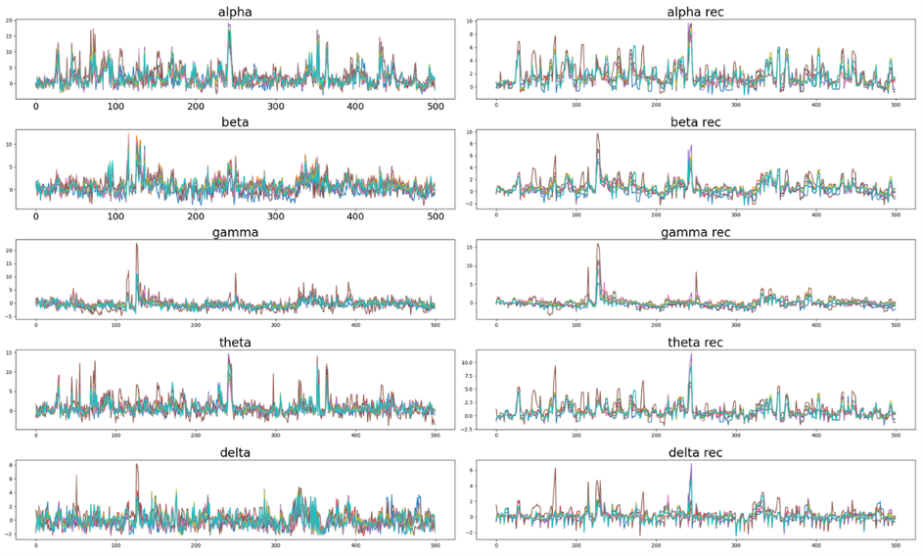

Fig. 2. Reconstruction of EEG frequency bands with Variational Auto-Encoder

6 Conclusions and Future Work

At present only EEG signal emdeddings method had been chosen and testes. Work with eye-tracking features has started, but still in progress. Future development os this work will include generation of embeddings from eye-tracking fetures sequentially predicting next fixation or its probability. Neural network architecture based on the multi-modal embeddings will be trained on available datasets. Main focus will be made on dimensions of multi-modal embeddings to reduce the padding proportion.

Acknowledgement. This work is supervised by Ivan Shanin, Federal Research Center "Computer Science and Control" of Russian Academy of Sciences.

References

1. Hollenstein, Pirovano F., Ce Zhang, Jager L. Beinborn: Multilingual Language Models Predict Human Reading Behavior. arXiv preprint arXiv:2104.05433, (2021)
2. Devlin, J., Chang, M. W., Lee, K., Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805. (2018)

3. Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzman, F., Grave, E., Ott, M., Zettlemoyer, L., Stoyanov, V., Unsupervised Cross-lingual Representation Learning at Scale. arXiv preprint arXiv:1911.02116. (2020)
4. Muttenthaler, L., Hollenstein, N., Barrett, M. Human brain activity for machine attention. arXiv preprint arXiv:2006.05113. (2020)
5. Hollenstein, N., Rotsztejn, J., Troendle, M., Pedroni, A., Zhang, C., Langer, N. ZuCo, a simultaneous EEG and eye-tracking resource for natural sentence reading. Scientific data, 5(1), 1-13. (2018)
6. Hollenstein, N., Troendle, M., Zhang, C., Langer, N. ZuCo 2.0: A dataset of physiological recordings during natural reading and annotation. arXiv preprint arXiv:1912.00903. (2019)
7. Kropotov J.D. Quantitative EEG, Event Related Potentials and Neurotherapy by Juri ISBN: 978-0-12-374512-5 (2009)
8. Barrett, M., Hollenstein, N. Sequence labelling and sequence classification with gaze: Novel uses of eye-tracking data for Natural Language Processing. Language and Linguistics Compass, **14**(11), 1-16 (2020).
9. Culotta, A., McCallum, A., Betz, J. Integrating probabilistic extraction models and data mining to discover relations and patterns in text. In Proceedings of the Human Language Technology Conference of the NAACL, Main Conference (pp. 296-303). (2006)
10. Barrett, M., Bingel, J., Hollenstein, N., Rei, M., Søgaard, A: Sequence classification with human attention. In Proceedings of the 22nd Conference on Computational Natural Language Learning, pp. 302-312. (2018)
11. Barrett, M., Keller, F., Søgaard, A. Cross-lingual transfer of correlations between parts of speech and gaze features. In Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers, pp. 1330-1339. (2016)
12. Barrett, M., Søgaard, A. Reading behavior predicts syntactic categories. In Proceedings of the nineteenth conference on computational natural language learning, pp. 345-349. (2015)
13. Dimigen, O., Sommer, W., Hohlfeld, A., Jacobs, A. M., Kliegl, R. Coregistration of eye movements and EEG in natural reading: analyses and review. Journal of experimental psychology: General, 140(4), 552. (2011)
14. Barrett, M., Bingel, J., Keller, F., Søgaard, A. Weakly supervised part-of-speech tagging using eye-tracking data. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pp. 579-584. (2016)
15. Hollenstein, N. Leveraging Cognitive Processing Signals for Natural Language Understanding (Doctoral dissertation, ETH Zurich). (2021)
16. Hollenstein, N., Renggli, C., Glaus, B., Barrett, M., Troendle, M., Langer, N., Zhang, C. Decoding EEG Brain Activity for Multi-Modal Natural Language Processing. arXiv preprint arXiv:2102.08655. (2021)
17. Kingma, D. P., Welling, M. Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114.(2013)