TABS: Transformer Based Seizure Detection

J. Pedoeem, S. Abittan, G. Bar Yosef and S. Keene

The Cooper Union, School of Engineering, Department of Electrical Engineering, New York, USA
{jonathanped, shifraabittan, guybaryosef}@gmail.com, keene@cooper.edu

Abstract— In this paper, we develop and submit a novel model to the Neureka 2020 challenge[1] that can predict seizures from an EEG signal. This challenge places emphasis on limiting the number of false positives, or reducing the places in which background signal is erroneously labeled as a seizure. As a result, our model is intended to act as a tool to aid neurologists as opposed to supplanting them completely. The goal is to allow a doctor to move more quickly through EEG data, enabling the doctor to treat more patients and devote more time toward treatment. Our model is a transformer based neural network architecture. We call it TABS, after Transformer Based Seizure detector. TABS scored a sensitivity of 9.03% and a false alarm rate of 31.21 per 24 hours when evaluated using the Time-Aligned Event Score.

I. Introduction

Electroencephalography or EEG is a medical examination where electrodes are placed on different parts of the brain. An EEG study can be used for many different applications [1]. For example, it can be used to detect sleep abnormalities or brain trauma. A very common use of EEG is to detect seizures, in order to diagnose a patient with epilepsy. The procedure is noninvasive as the electrodes are placed onto the surface of the skull of the patient using an adhesive. EEG recordings can last between an hour and a whole day, in length. This exam is administered by a medical assistant. After the exam finishes, the assistant may cut out portions of the exam that he or she deems irrelevant. The rest of the file is left for the doctor to look over. Then, a doctor has to sift through hours of files across several different patients [2].

The voltage measured by a particular electrode is called a channel. Often nineteen electrodes are employed. In order to meaningfully examine the signal data, doctors take the differences between various channels and form a collection known as a montage. Montages are useful, because they cancel out noise while emphasizing events of interest. Despite the advantage of using a montage, the process of interpreting an EEG is time consuming for a doctor. Computers can aide in this process as they are good at doing repetitive tasks. Therefore, this problem is ripe for machine learning, particularly deep learning. Deep learning is a subset of machine learning where a high parameter model is trained gradient descent and back propagation [3].

[1] https://neureka-challenge.com/

TABS, Transformer Based Seizure Detection, is a deep learning model that is built around a transformer [4]. A transformer consists of a multi-head self attention mechanism followed by two fully connected layers. Multi-head self attention allows for the model to assign stronger weight to relevant parts of the input signal. We have decided to use a transformer because seizures are temporal in nature.

After building a machine learning model, there are various metrics available to evaluate the model. A false positive is an instance in which a patient is labeled as ill, but in reality the patient does not have the disease. A false negative is an instance in which a patient does have a disease, but is labeled as healthy. Some researchers will argue that a good model should have a low false negative rate: it is better for the model to declare more false positives then miss a potentially lethal positive. This argument definitely has its grounding; however, we argue that it depends on the use case. If the model is being used in conjunction with a doctor, and the model outputs many false positives, the doctor will quickly become frustrated with the system and not use it. The scoring metric for the Neureka 2020 challenge encapsulates this issue by penalizing a model's false positives heavily:

$$\begin{aligned}\text{Score} = \text{Sensitivity} - 2.5 * (\text{False Alarms/ 24 hrs}) \\ - 7.5 * \frac{\text{Avg Number of Channels}}{19}\end{aligned} \tag{1}$$

TABS' contributions are two-fold:

1) We demonstrate that we could achieve comparable results to the state of the art without a comprehensive pre-processing scheme.
2) We use a transformer based neural network architecture on seizure detection. To our knowledge this has not been done before.

II. Related Work

There have been several different attempts to algorithmically detect seizures in EEG data, spanning from signal processing to statistical analysis [5–13]. Deep learning has recently achieved state of the art in image and pattern recognition [14, 15], and natural language processing [16–18] making it a great choice for this

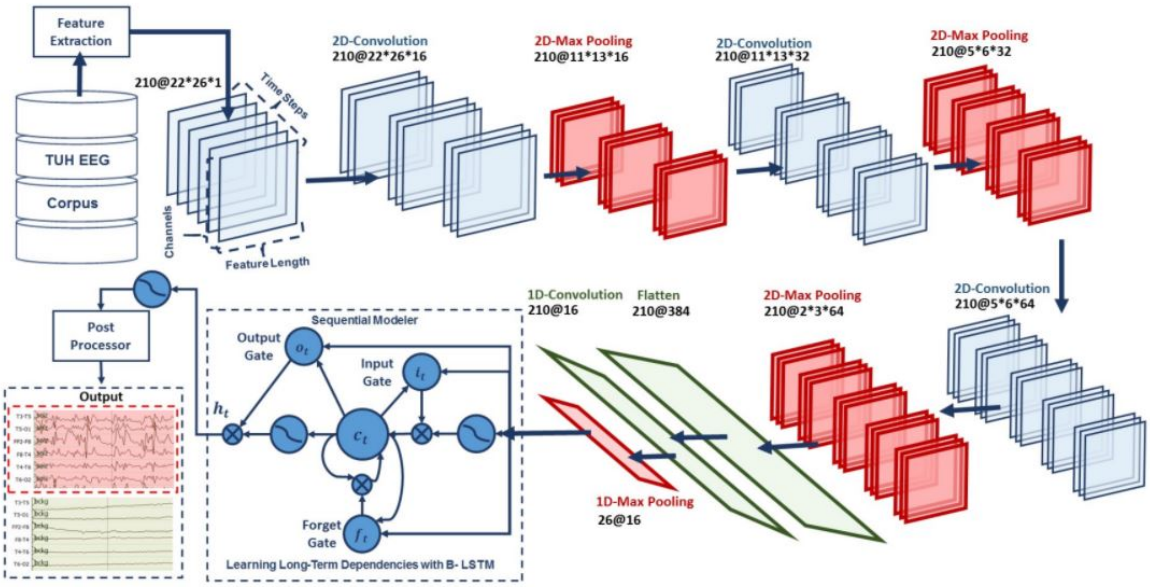

Figure 1. SOTA Model Architecture

problem.

However, deep learning requires a tremendous amount of data to build reliable models. Historically, there was not enough labeled data available to apply these techniques to seizure detection. The Neural Engineering Data Consortium at Temple University set out to solve this data problem. They collected and compiled approximately fourteen years of EEG data from patients at Temple Hospital and curated a corpus for research.

Version 1.5.1 of the corpus, released in March 2020, contains 642 subjects with a total of 1,423 sessions [19]. 447 of these sessions contain seizures. There are a total of 922 hours of data. Seizure events comprise about 63 hours, or 6.8% of the annotated data. The data set also includes metadata, in the form of physician's notes. These notes include patient demographics and medication.

One great difficulty in building a seizure detection system, even if limitless amounts of data are available, is that seizures do not have a precisely defined waveform. Even while hand-labeling the data, the annotators often debate whether or not a particular signal qualifies as a seizure. Seizures often lack discrete start and stop times.

The current state-of-the-art model achieves a false positive rate of 6 per 24 hours with sensitivity of 30.83% and a specificity of 97.10% [2]. This model is made up a time-distributed convolutional neural network (CNN) and a long short-term memory network[20] (LSTM) Figure 3. They also applied pre-processing and post-processing stages, before and after the deep learning.

The pre-processing stage takes the raw EEG files and extracts features from them, while keeping the EEG channels separate. To construct each feature data point, 9 time samples are utilized, each 0.1 seconds long. The features extracted include linear frequency cepstral coefficients, differential energy terms and first and second derivative terms. The model input was a matrix of size 22 channels x 26 features.

Each model input consisted of 210 time samples. Every time sample is then passed through a convolutional neural network in a time distributed manner. These values were then recombined before being passed into the LSTM.

The next stage was a CNN, followed by a bidirectional LSTM. The LSTM is helpful when working with sequential data. Finally, post-processing was applied, including a regression model, thresholding and filtering.

In our architecture search, we wanted to see how heavily we can rely on deep learning. For that, we decided to reduce the amount of pre-processing used.

When looking for the best architecture unit we decided to go with a transformer. Transformers allow for a model to view the context of a signal, similar to the recurrent aspects of the SOTA model.

The benefit of a transformer when compared to an LSTM or other recurrent architectures is that it can learn to place more attention on certain portions of the input.

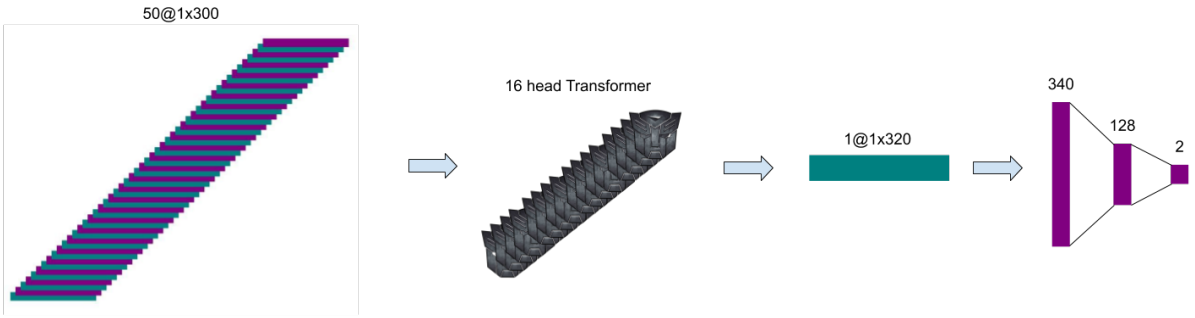

Figure 2. Our TABS Model Architecture

III. Motivation

In this paper, we set out to improve Temple's model (Figure 3) in two main areas. Firstly, Temple's SOTA model does not operate directly on the raw data, instead they first apply a series of signal processing feature extraction techniques. We saw this as an area for possible optimization, as using a pre-processing technique that is not learned runs the risk of losing information. We believe the neural network should learn the best preprocessing itself. Therefore, our neural network accepts the raw channels normalized to a uniform sampling rate (Section V-C).

Secondly, the machine learning community has been relying on LSTMs in their models since 1995. However, recent SOTA models in Natural Language Processing [16, 18] have begun using Transformers (see Section IV-B) with great success. In a similar way to how an LSTM [20] "remembers" previous inputs, Transformers also incorporate recurrency by allowing the model to access a large time window of the signal at once. The main differences between the two is that Transformers can be run in parallel while an LSTM looks at the input serially. The recent success of Transformers, outperforming all other deep learning architectures including LSTMs [16, 18], motivated us to use them instead of Temple's LSTM.

IV. Model Architecture

Our model consists of four parts: a Convolutional Neural Network layer, a Transformer layer, a CNN layer and Fully connected layers.

IV-A. Convolutional Layers

The CNN layer consists of a 1D convolution with an input size of 19 channels and output size of 50 channels, a kernel of size 5 and padding of 12. Following the 1D convolution we do 1D batch normalization, a dropout with probability 0.7 and ReLU layer. The CNN layer allows us to create a latent representation of the channels and have the model learn contextual information around each sample of the EEG.

IV-B. Transformer Layer

Following the CNN layer we have a 16 head Transformer with a hidden dimension of 20. Through experimentation we found that increasing the hidden dimension layer causes the model to overfit.

A Transformer is a neural network layer architecture that is composed of multi-head self attention (Equation 2) and some fully connected layers. The multi-head self attention allows for the model to look at a full signal and weigh (or attend to) different parts of the signal based on their relevance:

$$\begin{aligned} \text{MultiHead(Q,Q,Q)} = & Concat(head_1, ..., head_h)W^o \\ & \text{where } head_i = \\ & Attention(QW_i^Q, QW_i^k, QW_i^v) \\ & \text{and } Attention(Q,K,V) = \\ & softmax(\frac{QK^T}{\sqrt{d_k}})V \end{aligned} \tag{2}$$

where Q is the input into the transformer. In TABS it is a learned representation of a 300 sample window of an EEG signal. The W matrices are learned projection matrices. During training the model learns the optimal projections of the data.

IV-C. Fully Connected Layers

After this, we use another 1D convolution to form a linear combination of the 50 channels into 1 channel. The kernel is of size 1, dilation of 2, and padding of 10. We then pass this through batch normalization, dropout with probability 0.7 and finally a ReLU layer. The final fully connected (FC) layers are used to bring the model into the correct dimensions for evaluation. The first FC layer goes from 340 to 128 and is followed by batch

normalization, a dropout of 0.7, and a Relu, while the second and final FC layer goes from 128 to 2.

V. Discussion

V-A. Data

As previously mentioned, we used the TUH EEG Corpus to train our data. This corpus is the first collection of EEG data that is large enough to train robust deep learning models. However different EEG files in the corpus have different sampling rates as well as a different combination of channels. Additionally, even if the same channels appear in different files, they could be named differently.

Our pre-processing regime addresses these aspects of the data. It consisted of resampling all the files to a uniform 250hz. We additionally grouped together 19 channels that are common to all the files. We purposely made our pre-processing as minimal as possible. We operated directly on the raw EEG signals, without computing feature vectors.

Analyzing the data, we found an undersampling of the majority class, with only 6.8% of the samples in the corpus being seizures. Therefore, we only used a subset of the data, thereby achieving an approximately a 50/50 split between seizure and background data. Although this means we are not utilizing the full amount of data for training, we found that this prevents our model from over-fitting and labeling everything as background noise.

One of the techniques we used for regularization is called mixup [21]. Mixup takes a convex combination of two training samples and truth values. This allows for a more robust training. The coefficient for the convex combination is taken from a beta distribution with its alpha parameter set to 0.6. This essentially keeps beta around 1 or 0.

V-B. Pipeline

In order to help facilitate an experimentation we have developed a framework where we can easily plug in and experiment with a model. This includes data loading, model training, and model evaluation.

We developed a fast data loader built specifically for our pre-processed EEG data. The data loader allows us to choose whether we want our data sequentially (for validation and testing) or randomly (for training). We are also able to easily test out different hyper-parameters with this framework and save/load checkpoints based on training and validation accuracy, sensitivity, and specificity values. The data loader works keeping in memory a constant number of files and sampling from them until a threshold is reached. Once a threshold is reached the file is swapped out for a new one. This data loader is more than an order of magnitude faster than our original brute-force data loader implementation.

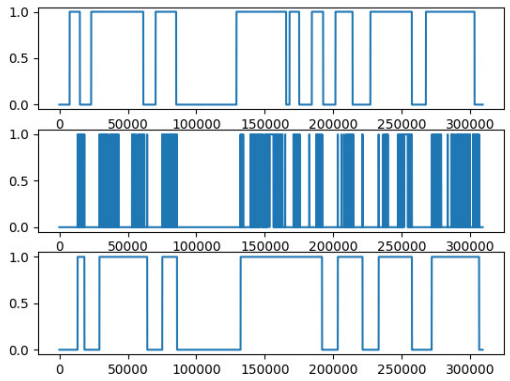

Figure 3. Example of Vote Based Smoothing (top: ground truth, middle: model output, bottom: smoothed result).

For testing, we built a model agnostic script that accepts any model and associated checkpoint, running the post-processing scripts on the model outputs and finally running Temple University's scoring script. This scoring script computes several different scoring metrics, including the one used in the Neureka contest, Time-aligned Event Scoring (TAES). More information on the scoring metrics can be found below.

V-C. Post-Processing Techniques

Due to the large penalty attributed to the false alarm rate, we experimented with several different post-processing heuristics. We experimented with different combinations of thresholding, vote-based smoothing, moving averages, and smoothing polynomial filters. We achieved our best results when pairing a Savitzky Golay filter [22], a smoothing filter which works by fitting polynomials to the data, with a post thresholding vote. We argue that smoothing before the thresholding gets rid of the jitters that was giving us a lot of false positives in the output.

V-D. Scoring Metrics

Because each data point is either a seizure or not, the types of errors are clear. However due to the fact that seizures come in sequences as apposed to single time samples, finding a scoring metric is challenging. Often, the seizures identified in the ground truth and the predicted sequence overlap. How must such patterns be labeled? There is an array of different evaluation metrics available. The two that we focused on are OVLP and TAES [23].

OVLP, is the metric we initially used because it is more lenient. This metric is term-based and not frame-based.

This means that each individual event, or seizure, is what is counted and not a comparison of the label at each individual time sample. In OVLP, a true positive is counted anytime the hypothesis overlaps in any way with the ground truth seizure annotation. A false positive is counted anytime the prediction does not overlap at all with a ground truth seizure. The length of a seizure or overlap is ignored and irrelevant in scoring. As such, OVLP is considered a somewhat permissive scoring metric.

TAES, or Time-Aligned Event Scoring, was the metric used in the Neureka 2020 Challenge. This metric considers the percentage and overlap between events in the ground truth and prediction sequences and weights the error. The true positive count is the total duration of a detected seizure divided by the total duration of the ground truth seizure. The false negative score is the fraction of the time that the ground truth seizure was missed divided by the total duration of the ground truth seizure. The false positive score is the total duration of the incorrect seizure in the predicted sequence divided by the total amount of time this seizure was incorrect according to the ground truth annotation. As compared to OVLP, TAES is quite strict.

VI. Results

We ran our model outputs through an evaluation metric script developed by The Neural Engineering Data Consortium of Temple University. This evaluation script provides several of the more common evaluation metrics used in EEG related research, including the any-overlap metric, time-aligned event scoring, epoch-based sampling, and dynamic-programming alignment. Temple presents a full description for each type of metric in [23].

Our goal in this paper was to optimize the any-overlap (OVLP) metric. This was chosen due to [2] describing it as the most popular choice in the neuroengineering community.

On the official test set we scored a sensitivity of 9.03% False Alarm rate of 31.21 per 24 hours and 19 channels giving us a score of -76.50. This placed us ninth place out of fourteen contestants in the the Neureka competition. The first place winners achieved a score of 12.37% sensitivity and a False Alarm rate of 1.44 per 24 hours.

These results show that our model is not as accurate as Temple's. We attribute several reasons for this. The first is that the signal-to-noise ratio(SNR) of seizures in EEGs is too low for deep learning to learn a strong feedback signal. The SNR was further lowered because of our choice to use the raw channel data. Secondly, due to time and computation constrains we were not able to fully tune our hyper-parameters.

VII. Conclusion

In this paper we present TABS, a novel model for EEG-based seizure detection. The design specification that was most important when developing TABS was achieving a very small false positive rate. The model architecture draws from cutting-edge, contemporary deep learning research. We built a hybrid architecture of convolutional layers, fully connected layers, and a transformer. Importantly, the only data preprocessing we use is grouping the data into uniform channels and resampling the time steps to a uniform sampling rate. This is noteworthy as it is significantly less preprocessing than what appears in Temple University's state-of-the-art model.

Our results are comparable to the SOTA and therefore suggest that much of the preprocessing used by Temple and others can be delegated to a more comprehensive deep learning model.

VIII. Future Work

There are a few possible adjustments and additions that could possibly improve the accuracy of TABS. A general bottleneck in our development was training time, as we did not have access to an unlimited number of GPUs. Due to a lack of time and resources, we were not able to fully explore these possibilities.

Firstly, initial values impact the stability of the model. By retraining the model many times, we can search the initial value space to find the most stable and fruitful set of values. Secondly, hyperparameter tuning is in order. Proper hyperparameters in deep learning can often improve results significantly. Although we did do a significant amount of hyperparameter adjustment, there may be room for improvement in this area [24]. Finally, we would have liked to incorporate data from the doctor's notes, such as patient medication, weight and gender. This information dictates the shape of the patient's brain waves and may help the model distinguish between seizures and background and could be used as a *multi-modal* approach [25]. For example, a patient who is already taking several medications may exhibit relatively subdued brain waves.

References

[1] C. P. Panayiotopoulos, "CURRENT PRACTICE OF CLINICAL ELECTROENCEPHALOGRAPHY, 3rd edition," *Brain*, vol. 127, no. 1, pp. 236–237, 01 2004. (available at: *https://doi.org/10.1093/brain/awh025*).

[2] M. Golmohammadi, V. Shah, I. Obeid, and J. Picone, *Deep Learning Approaches for Automated Seizure Detection from Scalp Electroencephalograms*. Cham: Springer International Publishing, 2020, pp. 235–276. (available at: *https://doi.org/10.1007/978-3-030-36844-9_8*).

[3] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. The MIT Press, 2016.

[4] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in Neural Information Processing Systems 30*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds. Curran Associates, Inc., 2017, pp. 5998–6008. (available at: *http://papers.nips.cc/paper/7181-attention-is-all-you-need.pdf*).

[5] B. Boashash, "Time-frequency signal analysis and processing: A comprehensive reference," 01 2003.

[6] M. Mera, D. M. López, R. Vargas, and M. Miño, "Automatic detection of epileptic spike in eegs of children using matched filter," *Brain Informatics*, S. Wang, V. Yamamoto, J. Su, Y. Yang, E. Jones, L. Iasemidis, and T. Mitchell, Eds. Cham: Springer International Publishing, 2018, pp. 392–402.

[7] P. Li, X. Wang, F. Li, R. Zhang, T. Ma, Y. Peng, X. Lei, Y. Tian, D. Guo, T. Liu, D. Yao, and P. Xu, "Autoregressive model in the lp norm space for eeg analysis," *Journal of neuroscience methods*, vol. 240, 11 2014.

[8] Y. Li, M.-L. Luo, and K. Li, "A multiwavelet-based time-varying model identification approach for time-frequency analysis of eeg signals," *Neurocomput.*, vol. 193, no. C, p. 106–114, Jun. 2016. (available at: *https://doi.org/10.1016/j.neucom.2016.01.062*).

[9] G. Rodríguez-Bermúdez and P. García Laencina, "Analysis of eeg signals using nonlinear dynamics and chaos: A review," *Applied Mathematics Information Sciences*, vol. 9, pp. 1–13, 01 2015.

[10] M. Eichler, R. Dahlhaus, and J. Dueck, "Graphical modeling for multivariate hawkes processes with nonparametric link functions," 2016.

[11] A. Schad, K. Schindler, B. Schelter, T. Maiwald, A. Brandt, J. Timmer, and A. Schulze-Bonhage, "Application of a multivariate seizure detection and prediction method to non-invasive and intracranial long-term eeg recordings," *Clinical neurophysiology : official journal of the International Federation of Clinical Neurophysiology*, vol. 119, pp. 197–211, 02 2008.

[12] K. Schindler, R. Wiest, M. Kollar, and F. Donati, "Eeg analysis with simulated neuronal cell models helps to detect pre-seizure changes," *Clinical neurophysiology : official journal of the International Federation of Clinical Neurophysiology*, vol. 113, pp. 604–14, 05 2002.

[13] P. Cherian, M. Vos, R. Swarte, J. Blok, G. Visser, P. Govaert, and S. Huffel, "Automated neonatal seizure detection mimicking a human observer reading eeg," *Clinical neurophysiology : official journal of the International Federation of Clinical Neurophysiology*, vol. 119, pp. 2447–54, 10 2008.

[14] J. Zbontar, F. Knoll, A. Sriram, M. J. Muckley, M. Bruno, A. Defazio, M. Parente, K. J. Geras, J. Katsnelson, H. Chandarana, Z. Zhang, M. Drozdzal, A. Romero, M. Rabbat, P. Vincent, J. Pinkerton, D. Wang, N. Yakubova, E. Owens, C. L. Zitnick, M. P. Recht, D. K. Sodickson, and Y. W. Lui, "fastmri: An open dataset and benchmarks for accelerated MRI," *CoRR*, vol. abs/1811.08839, 2018. (available at: *http://arxiv.org/abs/1811.08839*).

[15] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. S. Bernstein, A. C. Berg, and F. Li, "Imagenet large scale visual recognition challenge," *CoRR*, vol. abs/1409.0575, 2014. (available at: *http://arxiv.org/abs/1409.0575*).

[16] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: pre-training of deep bidirectional transformers for language understanding," *CoRR*, vol. abs/1810.04805, 2018. (available at: *http://arxiv.org/abs/1810.04805*).

[17] A. Wang, A. Singh, J. Michael, F. Hill, O. Levy, and S. R. Bowman, "GLUE: A multi-task benchmark and analysis platform for natural language understanding," *CoRR*, vol. abs/1804.07461, 2018. (available at: *http://arxiv.org/abs/1804.07461*).

[18] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei, "Language models are few-shot learners," 2020.

[19] I. Obeid and J. Picone, "The temple university hospital eeg data corpus," *Frontiers in Neuroscience*, vol. 10, 05 2016.

[20] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, p. 1735–1780, Nov. 1997. (available at: *https://doi.org/10.1162/neco.1997.9.8.1735*).

[21] H. Zhang, M. Cissé, Y. N. Dauphin, and D. Lopez-Paz, "mixup: Beyond empirical risk minimization," *CoRR*, vol. abs/1710.09412, 2017. (available at: *http://arxiv.org/abs/1710.09412*).

[22] A. Savitzky and M. J. E. Golay, "Smoothing and differentiation of data by simplified least squares procedures." *Analytical Chemistry*, vol. 36, no. 8, pp. 1627–1639, 1964. (available at: *https://doi.org/10.1021/ac60214a047*).

[23] S. Ziyabari, V. Shah, M. Golmohammadi, I. Obeid, and J. Picone, "Objective evaluation metrics for automatic classification of EEG events," *CoRR*, vol. abs/1712.10107, 2017. (available at: *http://arxiv.org/abs/1712.10107*).

[24] J. Frankle and M. Carbin, "The lottery ticket hypothesis: Training pruned neural networks," *CoRR*, vol. abs/1803.03635, 2018. (available at: *http://arxiv.org/abs/1803.03635*).

[25] X. Liu, P. He, W. Chen, and J. Gao, "Multi-task deep neural networks for natural language understanding," *CoRR*, vol. abs/1901.11504, 2019. (available at: *http://arxiv.org/abs/1901.11504*).