RECOGNITION OF SILENTLY SPOKEN WORD FROM EEG SIGNALS USING DENSE ATTENTION NETWORK (DAN).

Sahil Datta[⋆] *Akuha Aondoakaa*[⋆] *Jorunn Jo Holmberg*[†] *Elena Antonova*[†]
[⋆]Department of Electrical and Electronic Engineering, College of Engineering, Design, and Physical Sciences
[†]Division of Psychology, Department of Life Sciences, College of Health, Medicine and Life Sciences
Brunel University London

ABSTRACT

In this paper, we propose a method for recognizing silently spoken words from electroencephalogram (EEG) signals using a Dense Attention Network (DAN). The proposed network learns features from the EEG data by applying the self-attention mechanism on temporal, spectral, and spatial (electrodes) dimensions. We examined the effectiveness of the proposed network in extracting spatio-spectro-temporal information from EEG signals and provide a network for recognition of silently spoken words. The DAN achieved a recognition rate of 80.7% in leave-trials-out (LTO) and 75.1% in leave-subject-out (LSO) cross validation methods. In a direct comparison with other methods, the DAN outperformed other existing techniques in recognition of silently spoken words.

Index Terms— Brain Computer Interface (BCI), Silently Spoken Speech, Attention Mechanism, Electroencephalogram (EEG)

1. INTRODUCTION

Electroencephalogram (EEG) enables non-invasive recording of brain activity, which can be used to design a communication based brain computer interface (BCI) for people with motor disabilities. However, EEG signals have poor spatial resolution and low signal-to-noise ratio (SNR) [1]. In addition, EEG signals suffer from inter-trial and inter-subject variations [1]. Traditional feature extraction methods such as common spatial patterns (CSP), Fast Fourier Transform (FFT), and discrete wavelet transform (DTW) cannot easily adapt to the variations in the EEG signals [2].

On the other hand, deep learning algorithms have been used to extracting more robust representations than achieved by traditional feature engineering [3]. Panachakel [4] used a multi-layer perceptron for recognition of silently spoken speech from EEG signals. Further, Kumar [3] proposed a convolutional long short memory (CNN-LSTM) network to learn spatio-temporal features for recognition of silent speech. Similarly, Datta [5] recognized silently spoken words using convolutional attention network.

In addition, attention mechanism [6] in deep learning have made it feasible to address desirable features in multiple dimensions of the input [7], [8], [9]. For instance, *Squeeze-and-Excitation* (SE) module [9] and *Convolution Block Attention Module* (CBAM) [8] have been proposed to highlight important channel-wise features in the CNNs. However, the convolutional operation misses the global information focusing on local neighbourhood and it is computationally more expensive [10]. On the other hand, the self-attention mechanism can focus on long rage dependencies, enhance desirable features, and suppresses background noise [11], [7]. Further, the attention based networks are faster to train and require less computational resources compared to CNNs [12].

In this work, we propose the Dense Attention Network (DAN), an architecture designed entirely using fully connected (dense) layers and self-attention mechanism. The network learns spatial, spectral, and temporal patterns from EEG spectrograms of silently spoken words. The DAN exploits inter-dependencies of spectral features from different brain regions by applying attention across electrodes (channels). Further, the DAN learns spatio-spectral representations for each time point in the spectrogram separately, treating it as a time varying input.

We assessed the performance of the proposed network using two different experimental protocols, one of which included testing the network on data from an independent participant, i.e., not used for training the DAN. Further, we compared the performance of our network with existing techniques, where our network outperformed previously proposed methods in recognition of silently spoken words.

2. DATA ACQUISITION AND PRE-PROCESSING

2.1. Data Collection

EEG signals were recorded from 12 participants (Mean age 37, range 21-71); none of the participants had any neurological or speech-related disorders. A Neuroscan 64 channel Quik cap (electrodes) was used, with a sampling rate of 1kHz. EEG data collection took place in an EEG lab, where each participant sat in a chair in front of a computer screen at

a distance of 1 meter. Participants were instructed to remain immobile throughout the recording. Each recording session included the words "Apple" and "Write." To avoid systematic presentation order effects, words were presented on the screen in a random order [13]. Each participant completed 10 trials of each word. Recording of the EEG signals was time-locked for the accurate timing of the stimulus presentation. A blank screen was displayed for 1 second prior to the commencement of the stimulus, followed by word presentation for 2 seconds. Participants were instructed to mentally read the word as soon as it appeared on the screen. The presentation of the word was followed by a blank screen for 1 second. To circumvent overlapping inter-trial EEG activity three seconds (-0.5 s to 2.5 s) of the trial were used for the analysis. This work has been approved by the College of Engineering, Design, and Physical Sciences Research Ethics Committee, Brunel University London, reference number 7361-LR-Sep/2017-8301-1.

2.2. Pre-processing

The recorded EEG signals were high-pass filtered at 0.1Hz to remove noise or artifacts due slow voltage shifts at frequencies below 0.1Hz. Similarly, a notch filter was used to remove the harmonics of the 50Hz line noise. Noise at the higher frequencies, such as noise due to muscle movement, was eliminated by means of EMG electrode. Artifacts due to eye blinks were corrected by measuring the peak-to-peak voltage of the VEOG signal along with the threshold voltage of $\pm 200\,\mu V$ [14]. Baseline of the raw data was corrected in real time and during offline processing, using Neuroscan Curry 8 software.

3. DENSE ATTENTION NETWORK

3.1. Input Feature Map

We used spectrograms of EEG signals to enables more precise analysis. For Short Time Fourier Transform (STFT), we used Hann window of length 256 with a temporal overlap of 87% between two consecutive windows. A shorter window length was used to enhance the temporal resolution. In our analysis included frequencies between $5Hz$ and $330Hz$. Input to the network was a multi-dimensional feature map, $I_m \in R^{T \times F \times C}$, where T is the time, F is the frequency, and C is the number of channels (electrodes).

3.2. Network Architecture

Our proposed network learns spectro-spatial features at each time point separately. DAN uses self-attention and dense (MLP) layers to learn representations across channels and frequency information. The attention mechanism across channel dimension enables the network to select the most task discriminative electrodes. In addition, spectral attention helps the network capture interactions between different frequencies. Furthermore, the network uses self-attention mechanism to model the temporal dynamics of EEG spectrograms.

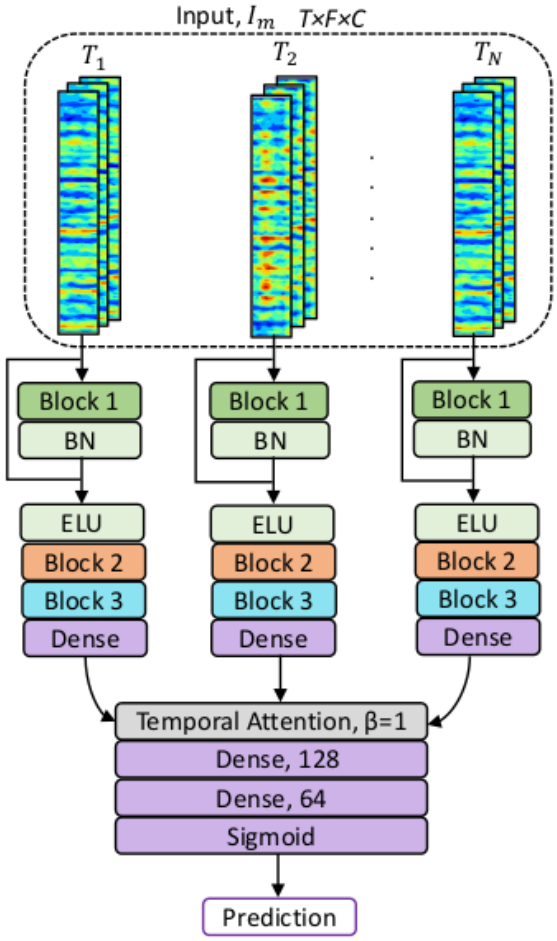

Fig. 1: DAN's Architecture. The network processes each time point separately using three blocks which extract spectral and spatial features. This is followed by dimensionality reduction using a dense layer with 256 hidden units and modelling temporal dynamics using self-attention (temporal attention). The dense layers in block 2 had 128 hidden units, and 50 hidden units in block 3. BN: batch-normalization.

The DAN's architecture is shown in Figure 1. The network consists of three blocks, block 1 shown in Figure 2, apply attention mechanism along the channel and spectral dimensions of the input. In each block, we used a residual connection [15], followed by batch normalisation [16]. Block 2 and 3 learn features using the multi-layer perception (MLP) with single hidden layer and the exponential linear unit (ELU) activation [17]. The MLP is applied first across channel dimension and then across spectral dimension. The architecture of blocks 2 and 3 is shown in Figure 3. After third block, the features across channel and frequency axis

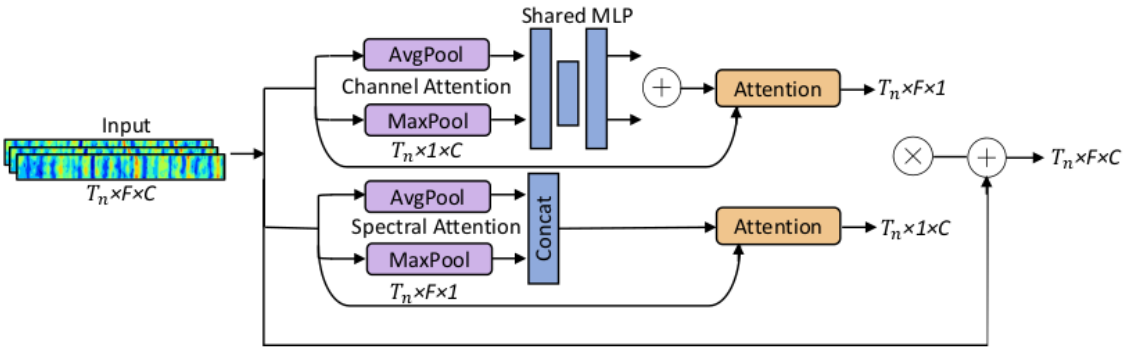

Fig. 2: Block 1 architecture, used for implementing channel and spectral attention in the network. The figure shows processing of n^{th} time point T_n. Concat refers to concatenating of two vectors obtained from average and max-pooling operations.

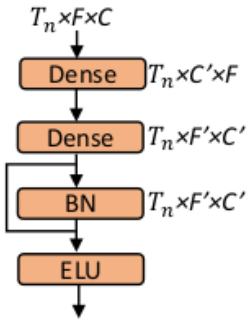

Fig. 3: Block 2 and 3 architecture. The first dense layer is applied along the channel dimension C, this is followed by swapping the axes, with the second dense layer applied across the spectral dimension F. C' and F' refers to the transformed features, the dimensions of the transformed features depends on the number of hidden units in the dense layers.

are transformed to a vector and subjected to dimensionality reduction using a dense layer. This is followed by temporal attention which provide most discriminative time points for recognition of silently spoken words.

The attention across the channel dimension is estimated by aggregating spectral information. The spectral information is aggregated using max-pooling and average-pooling operation, generating two feature descriptors; I^c_{avg}, and I^c_{max}, which are fed to a shared network composed of the multi-layer perception (MLP). The MLP contains single hidden layer with $\frac{C}{r}$ number of neurons and the ELU activation, where r is the reduction ratio with default value as in [8]. The output features from the shared network are combined by using element-wise summation to form $M_C \in R^{T\times C\times 1}$. For implementing self-attention, we used the $softmax$ function with inverse temperature β, defined as:

$$Softmax = \frac{\exp(\beta x_i)}{\sum_{i=1}^{N}\exp(\beta x_i)} \quad (1)$$

where x_i is a vector of length N. Inverse temperature β in the softmax function, helps the network avoid background noise in EEG signals and sharping the weights of important channel and spectral features [18]. The self-attention was applied as:

$$\alpha_c = Softmax(\tanh(M_C)) \quad (2)$$

a channel feature vector g_c is obtained by using α_c and the original input feature map I_m. The channel feature vector $g_c \in R^{T\times F}$ is obtained as:

$$g_c = \sum_{c=1}^{C}\alpha_c I_m \quad (3)$$

Spectral attention is computed by aggregating channel information by max-pooling and average-pooling operation, followed by concatenating the two feature descriptors to form $M_F \in R^{T\times F\times 1}$. Attention mechanism is applied to M_F using (2) and (3) to calculate a spectral feature vector $g_f \in R^{T\times C}$. The channel and spectral vector for each time point are multiplied to form attention feature map $G_m \in R^{T\times F\times C}$. Further, a refined feature map $M_m \in R^{T\times F\times C}$ is estimated by combining G_m and I_m using element-wise summation operation. The attention across temporal dimension is applied with $\beta = 1$.

The network was implemented on a NVIDIA P100 GPU using the *Keras* library [19] with *Tensorflow* [20] backend. DAN used the Adam algorithm [21] for weight optimization. The network was trained for 200 epochs with a learning rate of 0.001 and a mini-batch size of 5. Regularization in the network was implemented using the batch-normalization and residual connections.

4. RESULTS

To discriminate silently spoken words, we used EEG signals from 12 participants for two words: "Apple" and "Write".

Table 1: Classification accuracy of silently spoken words in leave-subject-out (LSO) evaluation method.

β	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	S11	S12	Average
1.5	75.5	77.0	55.0	74.4	98.8	88.9	92.7	53.9	55.6	71.5	86.6	72.0	75.1
1.8	74.5	77.5	54.4	75.5	99.4	87.7	95.0	56.5	53.7	73.5	88.8	69.5	75.3

Although each participant completed 10 trials of a silently spoken word, some participants ended up with 9 trials because the trials contaminated by noise were removed during pre-processing. Two different experimental approaches were used to access the performance of the proposed network. Specifically, two sets of results were obtained at a group-level, i.e., the data used for evaluation contained EEG signals from all the participants. The two experimental approaches are summarized in Table 2. To account for the stochastic nature of deep learning algorithm, the DAN was trained and evaluated ten times for each test sample.

Table 2: Two experimental approaches: leave-subject-out (LSO) and leave-trial-out (LTO).

Exp	Training		Testing	
	Subjects	Trials	Subjects	Trials
LSO	All but one	All	one	All
LTO	All	90%	All	10%

4.1. Leave Subject Out (LSO)

The LSO cross validation method assessed our network's performance by training it on EEG trials from 11 participants and testing it on trials from a separate participant. Table 1 shows the classification accuracy for the LSO evaluation approach. The network was evaluated for distinct temperature β values. The results indicate that our method is capable of accurately classifying EEG signals from participants who were not used in the network's training.

4.2. Leave Trial Out (LTO)

In this LTO cross-validation method, 90% of the EEG trials from 12 participants were used for training and 10% of the EEG trials from 12 participants were used for testing the network. The results were evaluated by varying the training and test trials, and the accuracy was estimated by averaging the results from all the test trials. As can be seen from Table 3, summarizing the results of the LTO method, the network achieved a recognition rate of 80.7%.

4.3. Comparison

To validate the effectiveness of the DAN, we compared its performance with the other baseline methods by implementing them on our EEG dataset. The previous methods tested in this work are as follow: (1) Bashivan [1] extracted frequency information from the EEG signals and converted them to pictures which were evaluated using a CNN-LSTM network; (2) Panachakel [4] used the time and wavelet domain features of EEG signals and classification was performed on each channel separately using a multi-layer perception followed by hard voting to reach the final decision; (3) Sereshkeh [22] used discrete wavelet transform (DWT) features such as the standard deviation (SD) and root mean square (RMS) and performed classification using a regularized neural network; and (4) Kumar [3] used a CNN-LSTM network to extract spatio-temporal features from windows of EEG signals and used majority voting to reach the final decision.

Table 3: Evaluation of baseline methods on our EEG dataset using: leave-subject-out (LSO) and leave-trial-out (LTO) methods.

Method	LSO	LTO
Bashivan [1]	63.2	68.8
Panachakel [4]	50.6	52.2
Sereshkeh [22]	63.7	65.0
Kumar [3]	55.8	59.3
Proposed (β = 1.5)	75.1	80.7
Proposed (β = 1.8)	75.3	80.6

The performance of the existing methods on our EEG dataset is shown in Table 3. As can be seen our proposed network outperformed existing methods in recognition of silently spoken words from EEG signals. Furthermore, the proposed network takes less training time, compared to the baseline networks in [1], [3].

5. CONCLUSION

We proposed the Dense Attention Network (DAN) to recognize silently spoken words from EEG signals, which uses the self-attention mechanism and dense layers to learn representations from spectro-spatial information and model temporal dependencies of EEG spectrograms. The proposed network is capable of learning essential patterns in EEG signals produced during silent speech, making it superior to other tested methods in recognizing silently spoken words. Our Future work will involve analysis using a larger EEG dataset and interpretability of the DAN.

6. REFERENCES

[1] Pouya Bashivan, Irina Rish, Mohammed Yeasin, and Noel Codella, "Learning representations from EEG with

deep recurrent-convolutional neural networks," *arXiv preprint arXiv:1511.06448*, 2015.

[2] Sahil Datta and Nikolaos V Boulgouris, "Recognition of Grammatical Class of Imagined Words from EEG Signals using Convolutional Neural Network," *Neurocomputing*, 2021.

[3] Pradeep Kumar and Erik Scheme, "A Deep Spatio-Temporal Model for EEG-Based Imagined Speech Recognition," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 995–999.

[4] Jerrin Thomas Panachakel, AG Ramakrishnan, and TV Ananthapadmanabha, "Decoding imagined speech using wavelet features and deep neural networks," in *2019 IEEE 16th India Council International Conference (INDICON)*. IEEE, 2019, pp. 1–4.

[5] Sahil Datta, Jorunn Jo Holmberg, and Elena Antonova, "Electrode selection and convolutional attention network for recognition of silently spoken words from eeg signals," in *2021 IEEE Symposium Series on Computational Intelligence (SSCI)*. IEEE, 2021, pp. 01–08.

[6] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio, "Neural machine translation by jointly learning to align and translate," *arXiv preprint arXiv:1409.0473*, 2014.

[7] Irwan Bello, Barret Zoph, Ashish Vaswani, Jonathon Shlens, and Quoc V Le, "Attention augmented convolutional networks," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 3286–3295.

[8] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon, "CBAM: Convolutional block attention module," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 3–19.

[9] Jie Hu, Li Shen, and Gang Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7132–7141.

[10] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin, "Attention is all you need," *arXiv preprint arXiv:1706.03762*, 2017.

[11] Fei Wang, Mengqing Jiang, Chen Qian, Shuo Yang, Cheng Li, Honggang Zhang, Xiaogang Wang, and Xiaoou Tang, "Residual attention network for image classification," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 3156–3164.

[12] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al., "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[13] Anne Porbadnigk, Marek Wester, and Tanja Schultz Jan-p Calliess, "EEG-based speech recognition impact of temporal effects," 2009.

[14] Steven J Luck, *An introduction to the Event-Related Potential Technique*, MIT Press Cambridge, MA, 2005.

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[16] Sergey Ioffe and Christian Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," *arXiv preprint arXiv:1502.03167*, 2015.

[17] Djork Arné Clevert, Thomas Unterthiner, and Sepp Hochreiter, "Fast and accurate deep network learning by exponential linear units (ELUs)," *arXiv preprint arXiv:1511.07289*, 2015.

[18] Jan Chorowski, Dzmitry Bahdanau, Dmitriy Serdyuk, Kyunghyun Cho, and Yoshua Bengio, "Attention-based models for speech recognition," *arXiv preprint arXiv:1506.07503*, 2015.

[19] François Chollet et al., "Keras," `https://keras.io`, 2015.

[20] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, et al., "Tensorflow: Large-scale machine learning on heterogeneous distributed systems," *arXiv preprint arXiv:1603.04467*, 2016.

[21] Diederik P Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[22] Alborz Rezazadeh Sereshkeh, Robert Trott, Aurélien Bricout, and Tom Chau, "EEG classification of covert speech using regularized neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 12, pp. 2292–2300, 2017.