Hierarchical Deep Feature Learning for Decoding Imagined Speech from EEG

Pramit Saha, Sidney Fels

Department of Electrical and Computer Engineering, University of British Columbia, Vancouver, Canada
pramit@ece.ubc.ca, ssfels@ece.ubc.ca

Abstract

We propose a mixed deep neural network strategy, incorporating parallel combination of Convolutional (CNN) and Recurrent Neural Networks (RNN), cascaded with deep autoencoders and fully connected layers towards automatic identification of imagined speech from EEG. Instead of utilizing raw EEG channel data, we compute the joint variability of the channels in the form of a covariance matrix that provide spatio-temporal representations of EEG. The networks are trained hierarchically and the extracted features are passed onto the next network hierarchy until the final classification. Using a publicly available EEG based speech imagery database we demonstrate around 23.45% improvement of accuracy over the baseline method. Our approach demonstrates the promise of a mixed DNN approach for complex spatial-temporal classification problems.

Introduction

In the past decade, numerous methods have been proposed to decode speech and motor-related information from electroencephalography (EEG) signals for Brain-Computer Interface (BCI) applications. However, EEG signals are highly session-specific; infested with noises and artifacts. Interpreting active thoughts underlying vocal communication involving labial, lingual, naso-pharyngeal and jaw motion, is even more challenging than inferring motor imagery, since utterances involve higher degrees of freedom and additional functions in comparison to hand movement and gestures. As a result, it is extremely challenging to recognize phonemes, vowels and words from single-trial EEG data. The reported classification accuracy of existing methods (Nguyen, Karavas, and Artemiadis 2017; Min et al. 2016; DaSalla et al. 2009) are not satisfactory, showing that manual handcrafting of features or traditional signal processing algorithms lack sufficient discriminative power to extract relevant features for classification.

This work addresses these issues by implementing a deep learning based feature extraction scheme, targeting classification of speech imagery EEG data corresponding to vowels, short and long words. It is mentionworthy that most of the previous works applied to vowels and phonemes classification show a degradation of performance when applied to

words and vice versa. Therefore, this is the first work that aims to automatically learn a discriminative EEG manifold applicable to both word and vowel/phoneme based classification at the same time, using deep learning techniques.

Proposed framework

The problem of categorizing EEG data based on speech imagery can be formulated as a non-linear mapping $\hat{f}$ of a multivariate time-series input sequence $\mathbf{X}_t^c$ to fixed output $\boldsymbol{y}$, *i.e*, mathematically $\hat{\boldsymbol{f}} : \mathbf{X}_t^c \longrightarrow \boldsymbol{y}$.

We found that well-known deep learning techniques, like fully connected networks, CNN, RNN, autoencoders *etc.* fail to individually learn such complex feature representations from single-trial EEG data. Also, our investigation demonstrated that it is crucial to capture the information transfer between the electrodes rather than using the multi-channel high-dimensional EEG data that otherwise needs large training times and resource requirements. Therefore, instead of utilizing raw EEG, we compute channel cross-covariance, resulting in positive, semi-definite matrices encoding the joint variability of electrodes. We define channel cross-covariance (CCV) between any two electrodes c_1 and c_2 as: $Cov(X_t^{c_1}, X_{t+\tau}^{c_2}) = \mathbb{E}[X^{c_1}(t) - \mu_{X^{c_1}}(t)][X^{c_2}(t + \tau) - \mu_{X^{c_2}}(t + \tau)]$.

This is particularly important because higher cognitive processes underlying speech synthesis and utterances involve frequent information exchange between different parts of the brain. Hence, such matrices often contain more discriminative features and hidden information than mere raw signals. Besides, cognitive learning process underlying articulatory speech production involves incorporation of intermediate feedback loops and utilization of past information stored in the form of memory as well as hierarchical combination of several feature extractors. To this end, we develop our mixed neural network architecture composed of three supervised and single unsupervised learning step (Fig 1).

In order to decode spatial connections between the electrodes from the channel covariance matrix, we use six-layered 1D convolutional networks stacking two convolutional and two fully connected hidden layers with ReLU as activations. The kth feature map at a given CNN layer with input x, weight matrix W^k and bias b_k is obtained as: $h^k = ReLU(W^k * x + b_k)$. The network is trained with
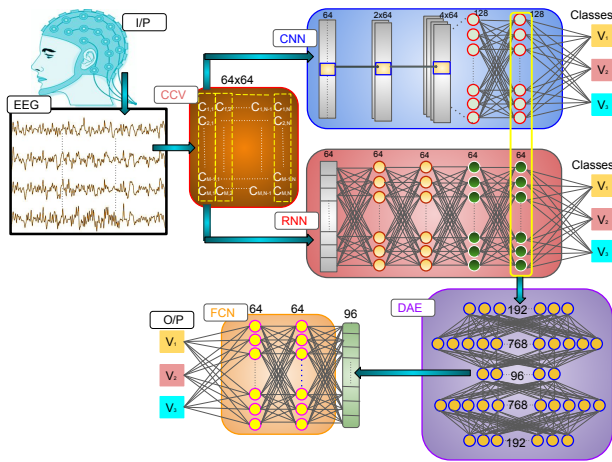
Figure 1: Overview of the proposed approach

Figure 2: Performance comparison for all subjects on vowels (left) and short words (right)

Table 1: Classification accuracy on long words

Method	S2	S3	S6	S7	S9	S11
Nguyen et al.	70.0	64.3	72.0	64.5	67.8	58.5
Proposed	77.5	90.7	73.7	86.8	80.1	71.1

the corresponding labels as target outputs, optimizing cross-entropy cost function via AdamOptimizer.

In parallel, we apply a six-layered recurrent neural network on the channel covariance matrices to explore the hidden temporal features of the electrodes. It consists of two fully connected hidden layers, stacked with two LSTM layers and is trained in a similar manner as CNN.

Since these parallel networks are trained individually and 5th layer of both the networks has a direct relationship with respective output layers, we claim that these layers are powerful discriminative spatial and temporal representations of the data. Therefore, we concatenate these feature vectors to form joint spatio-temporal encodings of covariance matrix.

The second level of hierarchy encompasses unsupervised training of deep autoencoders (DAE) having two encoder-decoder layers, with mean squared error (MSE) as the cost function. This leads to further dimensionality reduction of the spatio-temporal encodings (Zhang et al. 2018).

At the third level of hierarchy, the discrete latent vector representation of the deep autoencoder is fed into a two-layered fully connected network (FCN) followed by softmax classification layer. This is again trained in a supervised manner similarly as the CNN network, to output the final predicted classes corresponding to the speech imagery.

Experiments and Results

We evaluate our model on the publicly available imagined speech EEG dataset (Nguyen, Karavas, and Artemiadis 2017). It consists of imagined speech data corresponding to vowels, short words and long words, for 15 healthy subjects. The short words included in the study were 'in', 'out' and 'up'; the long words were 'cooperate' and 'independent'; the vowels were /a/, /i/ and /u/. The participants were instructed to pronounce these sounds internally in their minds, avoiding muscle movements and overt vocalization.

We trained the networks with 80% of the data in the training set and the remaining 20% in the validation set. Table 1 shows the average test accuracy for all subjects corresponding to long word classification. Fig 2 provides a compari-
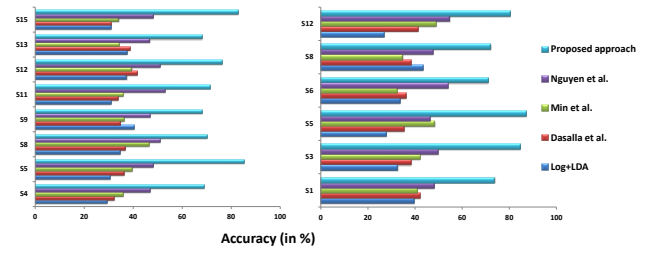
son of test accuracy of our approach with others on vowels and short word classification. Our results illustrate that our model achieves significant improvement over the other methods, thereby validating that deep learning based hierarchical feature extraction can learn a better discriminative EEG manifold for decoding speech imagery.

Conclusion

Towards recognizing active thoughts from EEG corresponding to vowels, short words and long words, this work presents a novel mixed neural network strategy as a combination of convolutional, recurrent and fully connected neural networks stacked with deep autoencoders. The network is trained hierarchically on a channel covariance matrix for categorizing respective EEG signals to the imagined speech classes. Our model achieves satisfactory performance with different types of target classification, on different subjects and hence can be considered as a reliable and consistent approach for classifying EEG based speech imagery.

References

DaSalla, C. S.; Kambara, H.; Sato, M.; and Koike, Y. 2009. Single-trial classification of vowel speech imagery using common spatial patterns. *Neural networks* 22(9):1334–1339.

Min, B.; Kim, J.; Park, H.-j.; and Lee, B. 2016. Vowel imagery decoding toward silent speech bci using extreme learning machine with electroencephalogram. *BioMed research international* 2016.

Nguyen, C. H.; Karavas, G. K.; and Artemiadis, P. 2017. Inferring imagined speech using eeg signals: a new approach using riemannian manifold features. *Journal of neural engineering* 15(1):016002.

Zhang, X.; Yao, L.; Sheng, Q. Z.; Kanhere, S. S.; Gu, T.; and Zhang, D. 2018. Converting your thoughts to texts: Enabling brain typing via deep feature learning of eeg signals. In *2018 IEEE PerCom*, 1–10. IEEE.