Adequately Wide 1D CNN facilitates improved EEG based Visual Object Recognition

Subhranil Bagchi
Department of Computer Science and Engineering
Indian Institute of Technology Ropar
Ropar, India
2018csy0002@iitrpr.ac.in

Deepti R. Bathula
Department of Computer Science and Engineering
Indian Institute of Technology Ropar
Ropar, India
bathula@iitrpr.ac.in

***Abstract*—Accurate classification of visual objects from Single-Trial EEG signals is a challenging task due to the low signal-to-noise ratio (SNR) associated with the brain signals. Recently, machine learning frameworks based on deep neural networks have shown great potential. Network architectures have grown increasingly complex with sophisticated modules to achieve the state-of-the-art performance. Unfortunately, finding the optimal network configuration is a tedious trial-and-error process. In this work, we propose to use a wider version of the simple 1D – CNN architecture with residual connections for EEG based visual object recognition. Experimental results establish that this fairly simple architecture outperforms existing techniques across five different classification tasks. Comprehensive ablation studies analyze the sensitivity of the model to varying parameters, especially the width of the network. We further showcase the features extracted for different classes using t-SNE plots, and demonstrate the superior discriminating quality of suitable network configuration through representational dissimilarity analysis.**

***Index Terms*—CNN, Wide, Residual, EEG, Representation**

I. Introduction

One of the recent interests in the scientific community has been to comprehend how different categories of objects are conceptualized in the brain by finding their representational similarities [1], [2]. Although earlier works have shown that distinct responses are processed in the ventral temporal cortex for different categories [3], it is, however, arguably difficult to deploy such Functional Magnetic Resonance Imaging (fMRI) based systems for day to day experimentation and applications due to its data scarcity, stringent and prolonged acquisition protocols and sheer per-sample data volume. In contrast, Electroencephalography (EEG), being a cheap and non-invasive method to record the brain's electrical activities, is one of the simplest and earliest proposed techniques for Brain-Computer Interface (BCI), which, over the past few decades, has long been researched for a wide range of feasible applications such as motor imagery, emotion recognition and medical ones.

Previous work [2] brings forward the use of Principal Component Analysis for feature extraction and Linear Discriminant based Classification amongst the categories and examples of different visual stimuli. One limitation of these single-trial EEG systems is the poor *signal-to-noise* (SNR) ratio. To overcome that, signal averaging and ERP is used [4]. Similar work [5] investigates the spatio-temporal dynamics of representational similarity using three different modalities: EEG, Magnetoencephalography (MEG) and fMRI, and employ averaged pattern vectors for EEG and MEG, followed by linear Support Vector Machine (SVM) to determine the decoding accuracy between pairs of images or categories across time. These hand-crafted features may lead to sub-par performance, as evident from their classification accuracies.

The advent of AlexNet [6] in Computer Vision eliminated the need for manual feature extractors, rather introduced the end-to-end training of deep Convolutional Neural Network (CNN), which learns local patterns from raw data at the lower layers, later decreasing the extent of this localization at higher layers by increasing its receptive fields, and produce superior results. Due to the simplistic recording procedure of EEG's, data of a relatively larger sample size could be procured, increasing the efficacy for these deep models if applied.

One work [7] shows the effectiveness of CNN with respect to Filter Bank Common Spatial Pattern (FBCSP), a commonly used feature extractor for motor imagery. The proposed models split the convolution operation at the first layer as a combination of two linear operations before passing through an activation: the first operation performs temporal convolution on individual input channels only, whereas, in the second operation, the output from the first stage is convolved over channels and features to learn spatial filter. It proposes four architecture types: Deep, Shallow, Hybrid and very-deep Residual (residual connections allow the training of extremely deep models [8]), and through a comprehensive analysis shows that Deep and Shallow networks perform on par with FBCSP, but the performance of deep residual network suffers.

Other deep learning approaches to classify raw EEG data use Long Short-Term Memory (LSTM), which learn intrinsic temporal dependencies [9]. However, CNNs have been evidenced to outperform these models for identification of the visual objects [10], [11]. One way to apply CNN is to convert the multi-channel EEG signals into an image like structure and apply deep models from computer vision, like AlexNet [11]. These models, however, result in huge parameter space. Current literature suggests that CNNs used for EEG based systems, in general, are often shallower, unlike computer vision, as they perform better [7], [12]. Particularly, EEG driven object recognition systems [10] may employ a shallow model with few filters, increasing in number at subsequent layers,

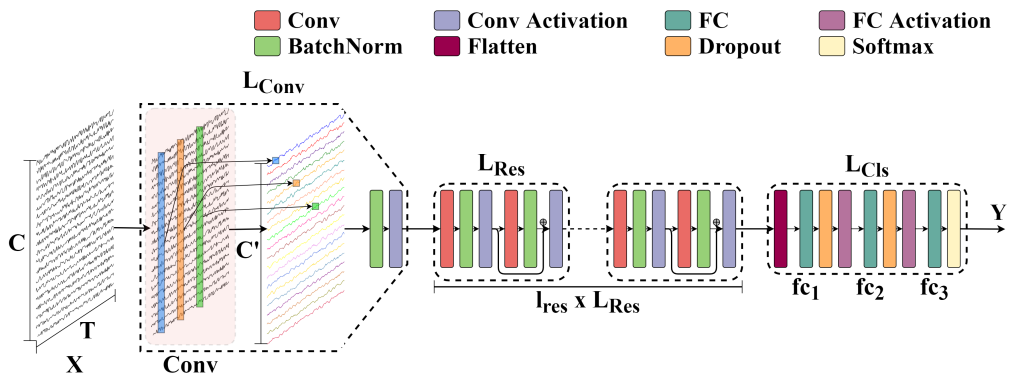


Fig. 1: Proposed framework of Wide, 1-D CNN with Residual Connections

along with an attentional mask for the occipital electrodes. We, however, argue that the number of filters plays a huge role in the model's performance [13] and the presence of a few filters may lead to information loss [14], and definitely limits the performance. We also argue that introducing residual connections for shallow networks may be beneficial, unlike deep ones, as residuals may push forward a representation from the lower layer and act like an ensemble [15].

To this regard, the main contributions of our work are as follows: (1) We propose a wider version of the simple 1-D CNN with residual connections for single-trial EEG based visual object classification, (2) Demonstrate the superiority of the proposed model on different classification tasks, (3) Analyze the sensitivity of the model with ablation studies, and lastly (4) Illustrate the importance of network width in learning class discriminating features.

II. Methodology

A. Dataset

We used the EEG dataset [16] from [2], where 10 subjects of different ages with normal color vision were shown photographs of 72 different stimuli from 6 different categories, with each category encompassing 12 distinct stimuli. The data was acquired over 2 experimental sessions for individual subjects, with each session containing 3 blocks of 864 trials and each stimuli being shown 12 times in random order in a block. This paradigm resulted in $5,184$ trials per subject.

The data collected from the 128 channel EGI HydroCel Geodesic Sensor Net [17], sampled at a frequency of 1 kHz, was preprocessed by a high-pass fourth-order Butterworth filter and a low-pass eighth-order Chebyshev Type I filter to preserve frequency components of and between 1 Hz and 25 Hz only, before it was downsampled to 62.5 Hz and channels $125 - 128$ were discarded. Finally, extended Infomax ICA removed ocular artifacts before it was average referenced and periodized into multiple trials, each of 32 time points [2].

B. Architecture

The proposed architecture is depicted in Fig. 1. The network takes as input (**X**) the 124 channel (**C**) EEG signal with 32 time samples (**T**) and passes it through the 1-D Convolutional Block ($\mathbf{L_{Conv}}$) to *temporally* convolve, integrating all the input channels, using filters of kernel size (**K**), and form newer channels ($\mathbf{C}'$), which then pass through a series of Residual Blocks ($\mathbf{L_{Res}}$) [8] and a Classifier Block ($\mathbf{L_{Cls}}$) of 3 fully-connected (FC) layers to output the final prediction (**Y**). Following each 1-D Convolution (Conv) layer, Batch Normalization (BatchNorm) layer [18] is added to help with training. For intermediate FC layers, $\mathbf{fc_1}$ and $\mathbf{fc_2}$ with output sizes of **500** and **100**, respectively, Dropout [19] with drop probability of 0.5 has been incorporated for regularization.

For the standard baseline, we set $\mathbf{l_{res}} = 2$ and $\mathbf{C}' = 512$. Exponential Linear Unit (ELU), being the suitable choice [7], was set as activation for $\mathbf{L_{Conv}}$, $\mathbf{L^1_{Res}}$ and $\mathbf{L^2_{Res}}$, whereas, for $\mathbf{fc_1}$ and $\mathbf{fc_2}$, Rectified Linear Unit (ReLU) was used. The preprocessed data was already downsampled by a factor of 16, so we excluded the use of pooling from the baseline.

III. Experimental Protocols and Analysis

A. Training and Evaluation Procedure

We trained the models by minimizing the Cross-Entropy Loss using Adam optimizer [21] with weight decay regularization. We conducted evaluations for within-subject object classification using *repeated 10-fold stratified cross validation*. This strategy splits the data for each subject into 10 folds, while ensuring uniform distribution of per-class samples across folds. For precise evaluation, the cross-validation procedure is repeated thrice with varying splits. We report the mean accuracy and the corresponding standard deviation obtained from the accuracies of all the subjects. For different classification tasks, we determined the different weight decays and number of training epochs based on the repeated cross validation accuracy. As for the individual variants of the architecture, we fine-tuned the weight decays pertaining to each task.

B. Classification Results

We evaluated the performance of our proposed architecture for 5 different classification tasks, as in [2]: 6 class category, 72 class exemplar, Human Face (HF) vs Inanimate Object (IO) category, HF exemplar and IO exemplar. Table I compares the classification performance of our 1-D Wide-Res CNN to existing methods. Though a fairly simple architecture, the

TABLE I: Comparison of classification performance using different approaches across different tasks

Method	Classification Accuracy (%)				
	HF vs IO	**HF Exemplar**	**IO Exemplar**	**6 Category**	**72 Exemplar**
LDA [2]	81.06 ± 3.66	18.30 ± 5.63	28.87 ± 10.57	40.68 ± 5.54	14.46 ± 6.43
ICA-ERP [20]	—	—	—	43.50	—
Shallow [10]	—	—	—	49.04 ± 6.99	23.72 ± 10.95
LSTM [10], [11]	80.67	—	—	44.77 ± 6.30	15.39 ± 6.01
LSTM + CNN [10]	—	—	—	46.18 ± 6.79	23.23 ± 10.48
CNN [10], [11]	83.10	—	—	50.00 ± 6.61	25.93 ± 10.67
Attention CNN [10]	—	—	—	50.37 ± 6.56	26.75 ± 10.38
CNN-ResNet101 [11]	85.50	—	—	—	—
1-D Wide-Res CNN (our)	$\mathbf{88.83 \pm 3.49}$	$\mathbf{24.64 \pm 7.90}$	$\mathbf{47.12 \pm 16.26}$	$\mathbf{51.29 \pm 7.57}$	$\mathbf{28.68 \pm 12.58}$

— *represents values not reported in their respective paper.*

proposed method outperforms all other reported methods, attributing the performance boost to the network's width and residuals. To get a better understanding of the latent-space representation of different categories learned by the model, we extracted the $\mathbf{fc_2}$ outputs for individual validation folds and visualized them using t-SNE plots [22]. As depicted in Fig. 2, the *t-SNE* showcases similarities of samples while preserving both the local and global structures from the extracted features.

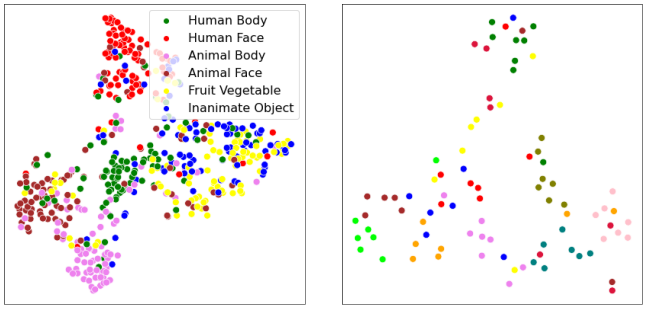
C. Ablation Studies

We performed comprehensive ablation studies to investigate the sensitivity of our proposed model to the network configurations as shown in Table II.

In general, the classification performance has a positive correlation with the increasing width till a certain threshold, after which it either saturates or deteriorates. The results suggest that, ideally, the number of filters used in the 1-D convolutional layers should be more than the number of input channels for learning reasonably good discriminating representations. It could further be inferred that the lower number of filters may have barred previous networks from realising their full potential [7], [10]. Interestingly, the influence of the width is striking for tasks with a significantly higher sample size - 6 class category and 72 class exemplar classifications. This is indicative of the optimal width being driven by the sample size, as established by [23].

In comparison to the width, the influence of depth appears to be more complex. While experimental results concur with previous studies that shallow networks outperform their deeper counterparts, determining the optimal depth still remains intricate. Results suggest that networks of one or two Residual Blocks perform better for a majority of tasks, while deeper ones may critically divest the performance. Furthermore, the effect of introducing Residual Blocks is strikingly visible for deeper networks of 5 Conv Layers, compared to 3 Conv Layers networks.

Ablation results indicate that kernel size is another important factor. Introduction of kernels of size 5 led to improved performance in three of the tasks. Additionally, although most deep architectures employ pooling, which not only reduces the computational complexity but also increases the receptive fields at higher layers, our results, in contrast, exhibit a statistically significant drop in accuracy for a majority of the tasks.



(a) 6 Category (Perplexity: 30) (b) IO Exemplars (Perplexity: 6)

Fig. 2: Sample t-SNE plots generated for different classification tasks for Subject 6. While the inherent structure of the data is depicted in both, the intra-class similarity is distinguishable for 6 category classification but not so evident for IO exemplar classification.

This drop may be attributed to the inherently low sampling resolution of the preprocessed dataset. It can be inferred that pooling should be not universally exercised but introduced judiciously based on the characteristics of the dataset.

D. Representational Analysis

For a deeper understanding of the influence of network's width, we compared the individual Conv layers' extent in discriminating the classes for networks with varying width. We incorporated the correlation distance based Representational Dissimilarity Matrices (RDM) [1] for all sample pairs per individual folds. The correlation distance is measured as $1-\rho$, where ρ is the Pearson's Correlation Coefficient (PCC). The RDM values range from 0 to 2, with 2 being the most dissimilar. The self-similarity must be 0, and is thus discarded. But the input signals for any distinct visual stimulus varies to some extent for different trials. Considering this variability for every within class sample pairs, it is the network's task to learn the underlying dissimilarities amongst the samples of different classes. Thus, the features learned by the Conv layers' filters, in learning these dissimilarities, should directly impact the performance.

To evaluate the discriminative capacity of each layer, we first calculated the pair-wise correlation distances for both

TABLE II: Ablation study analyzing the effect of varying network parameters

Variant	Architecture Type	Parameters (Approx.)	Classification Accuracy (%)				
			HF vs IO	***HF Exemplar***	***IO Exemplar***	***6 Category***	***72 Exemplar***
Standard	Baseline	$11.59M$	88.83 ± 3.49	24.64 ± 7.90	47.12 ± 16.26	51.29 ± 7.57	28.68 ± 12.58
Variation in Channel	64 Channels	$1.15M$	88.71 ± 3.52	$20.32 \pm 5.69^{\mathbf{a}}$	$38.45 \pm 13.00^{\mathbf{a}}$	$46.44 \pm 8.43^{\mathbf{a}}$	$21.75 \pm 10.50^{\mathbf{a}}$
	128 Channels	$2.35M$	88.84 ± 3.64	$23.10 \pm 6.93^{\mathbf{b}}$	$44.46 \pm 14.99^{\mathbf{a}}$	$49.53 \pm 7.52^{\mathbf{a}}$	$22.06 \pm 11.01^{\mathbf{a}}$
	256 Channels	$5.03M$	88.95 ± 3.60	24.01 ± 7.91	$45.65 \pm 15.62^{\mathbf{a}}$	$49.84 \pm 7.32^{\mathbf{a}}$	$25.21 \pm 12.59^{\mathbf{a}}$
	768 Channels	$19.71M$	88.80 ± 3.57	24.48 ± 7.15	47.12 ± 16.33	$51.91 \pm 7.63^{\mathbf{a}}$	$\mathbf{29.93 \pm 12.47^{\mathbf{a}}}$
	1028Channels	$29.42M$	88.72 ± 3.55	24.67 ± 7.64	46.91 ± 16.20	51.80 ± 7.51	$29.76 \pm 12.56^{\mathbf{b}}$
Variation in Depth	0 Residual Layer	$8.44M$	$86.73 \pm 4.46^{\mathbf{a}}$	22.96 ± 6.25	$41.82 \pm 13.48^{\mathbf{a}}$	$\mathbf{52.08 \pm 7.56}$	$19.85 \pm 9.72^{\mathbf{a}}$
	1 Residual Layer	$10.01M$	$89.21 \pm 3.64^{\mathbf{a}}$	24.47 ± 7.46	46.72 ± 16.21	$\mathbf{52.08 \pm 7.67^{\mathbf{a}}}$	$26.23 \pm 13.08^{\mathbf{a}}$
	3 Residual Layers	$13.16M$	88.73 ± 3.75	24.65 ± 7.75	46.62 ± 15.90	51.04 ± 7.21	$27.99 \pm 12.09^{\mathbf{b}}$
	4 Residual Layers	$14.74M$	88.50 ± 3.85	24.07 ± 7.30	46.15 ± 16.18	$50.72 \pm 7.07^{\mathbf{b}}$	$26.80 \pm 11.60^{\mathbf{a}}$
Absence of Residuals	3 Conv Layers	$10.01M$	88.82 ± 3.62	24.64 ± 7.22	47.05 ± 16.32	50.74 ± 7.45	$22.18 \pm 11.49^{\mathbf{a}}$
	5 Conv Layers	$11.59M$	$87.58 \pm 3.82^{\mathbf{a}}$	23.27 ± 6.70	$42.69 \pm 14.76^{\mathbf{a}}$	$48.31 \pm 7.18^{\mathbf{a}}$	$26.55 \pm 11.65^{\mathbf{a}}$
Pooling Techniques	Avg Pool Twice	$5.44M$	88.92 ± 3.59	$22.22 \pm 7.08^{\mathbf{a}}$	$42.90 \pm 15.69^{\mathbf{a}}$	$50.06 \pm 7.37^{\mathbf{a}}$	$25.00 \pm 11.78^{\mathbf{a}}$
	Max Pool Twice	$5.44M$	89.00 ± 3.62	$22.20 \pm 6.58^{\mathbf{b}}$	$41.76 \pm 14.39^{\mathbf{a}}$	$47.43 \pm 7.07^{\mathbf{a}}$	$19.90 \pm 8.63^{\mathbf{a}}$
Variation in Kernel Size	Kernel Size 5	$13.81M$	$\mathbf{89.54 \pm 3.39^{\mathbf{a}}}$	25.25 ± 8.04	$\mathbf{47.42 \pm 16.51}$	50.99 ± 7.55	28.44 ± 12.18
	Kernel Size 7	$16.04M$	$89.46 \pm 3.49^{\mathbf{a}}$	24.55 ± 7.76	46.66 ± 16.17	$50.65 \pm 7.53^{\mathbf{b}}$	$28.00 \pm 12.02^{\mathbf{b}}$
	Kernel Size 3 and 5	$12.70M$	$89.51 \pm 3.33^{\mathbf{a}}$	$\mathbf{25.36 \pm 7.88^{\mathbf{b}}}$	47.31 ± 16.25	51.26 ± 7.69	28.80 ± 12.37
	Kernel Size 5 and 7	$14.92M$	$89.44 \pm 3.53^{\mathbf{b}}$	24.75 ± 7.77	47.17 ± 15.91	50.90 ± 7.60	28.40 ± 12.14

'Parameters': respect to *12 output classes*. 'Statistical Significance' compared to the Baseline: $^{\mathbf{a}}p\text{-}Value \leq 0.01$, $^{\mathbf{b}}p\text{-}Value \leq 0.05$.

the *intra* and *inter* class sample pairs for all the validation folds. The hidden representation at each layer for each sample is obtained by reshaping all per-channel output vectors to a single output vector. This vector is standardized and converted to unit-variable [24]. PCC of unit-variable pair of samples allows for the correlation distance to be expressed as Euclidean distance, as:

$$d_{ij} = \sqrt{2}\sqrt{1 - \rho_{ij}}, \tag{1}$$

where, d is the Euclidean Distance, ρ is the PCC, and i and j are sample pair from same or different classes.

We hypothesize that the representational dissimilarity among the intra and inter class sample pairs' should provide us with some insights into the degree of separability at individual layers for networks of varying width. To this end, we incorporated the transformation from (1) to the dissimilarities and calculated this separability in terms of the Fisher Score [25]. This representational separability for the intra and inter pair groups is, thus, written as follows:

$$r_d = \frac{p_{inter}(\mu_{inter} - \mu_b)^2 + p_{intra}(\mu_{intra} - \mu_b)^2}{p_{inter}\sigma^2_{inter} + p_{intra}\sigma^2_{intra}}, \tag{2}$$

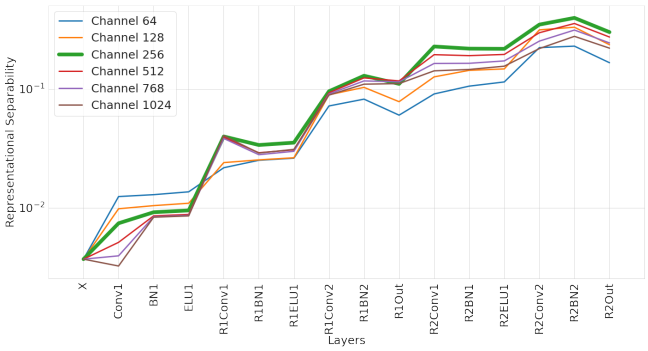
where, r_d is the measure of representational separability, and μ_{inter} , μ_{intra}, σ^2_{inter}, σ^2_{intra} are the means and variances of the dissimilarity measures (in terms of Euclidean distance) for all possible inter and intra class pairs, respectively, and the p_{inter} and p_{intra} are the respective fractions of inter and intra class dissimilarities from the total number of possible dissimilarities, and μ_b is the overall mean, calculated as $p_{inter}\mu_{inter} + p_{intra}\mu_{intra}$.

We measured r_d for individual layers of $\mathbf{L_{Conv}}$, $\mathbf{L^1_{Res}}$ and $\mathbf{L^2_{Res}}$, for networks of different widths. Two of these are depicted in Fig. 3. The separability is increasing at the higher layers, and for four of the cases, it was observed that the high separabilities at higher layers correlate well with the best performances (induced by the network's width). These results emphasize the implicit effect of the network's width on class discrimination capability.

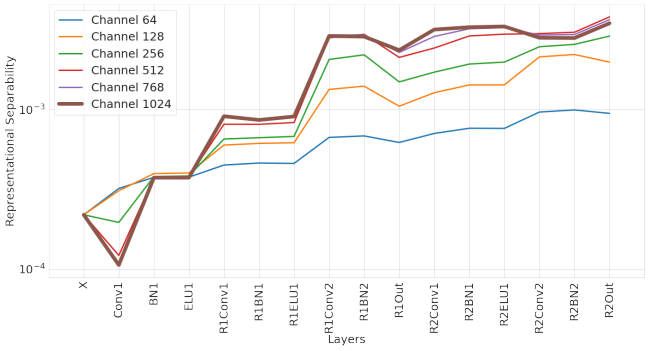
IV. Discussion

Single-trial EEG based visual object detection is a difficult task. Even with the state-of-the-art methods, the performance is still far from an acceptable threshold. The high noise associated with individual EEG signals affects the network's performance significantly. At times, methods such as averaging multiple trials of signals are used [4], which may improve the classification by a large extent. But the comparison of such a method's performance to single-trial models would not be fair. For single-trial approaches, deep neural networks can more effectively discriminate the noisy data than traditional machine learning techniques. However, due to many parameters, the performance of deep learning architectures varies significantly for each configuration. This makes determining the optimal set of parameters a laborious and time-consuming task. Additionally, variability in the performance observed for individual tasks, specifically for the width and depth parameters, makes us concur that the data sample size and the number of classes play a crucial role in determining proper parameter choices.

A limitation of this work is that it focuses on a single dataset. While our ablation studies have been comprehensive, the observed trends might not necessarily generalize to other datasets. One obvious instance would be the behaviour of pooling operation. To better understand the intricacies of such operations, future studies could be conducted on EEG data acquired at different resolutions. To observe meaningful and generalized trends, experiments need to be conducted on several datasets that differ in sample size, number of classes, temporal and spatial resolution under a range of experimental setting. While such *into the wild* datasets may be available for computer vision, but, to the best of our knowledge, not for EEG currently.



(a) HF vs IO Classification



(b) HF Exemplar Classification

Fig. 3: Plots for layer-wise representational separabilities (r_d), for $\mathbf{L_{Conv}}$, $\mathbf{L^1_{Res}}$ and $\mathbf{L^2_{Res}}$. The thick line corresponds to the network with the maximum accuracy. The plots are in log-scale to ensure better visibility. The low separability in the HF Exemplar Classification is apparent by its decoding accuracy.

V. Conclusion

In this work, we demonstrate the superior performance of a simple Wide-Residual 1-D CNN for single-trial EEG signal based visual object classification task in comparison to existing methods. Through in-depth ablation studies, we analyzed the effect of varying network parameters such as width, depth, residual connections, kernel sizes and pooling techniques for our proposed architecture. These studies revealed a distinct relation between the width and the performance of the network. These experiments concluded that an adequate width is necessary for a network to realize its full potential, and its parameters depend on the size of the training data and the number of classes. We further corroborate these observations by the latent-space representational separability achieved by the different network layers with varying width.

References

[1] N. Kriegeskorte, M. Mur, and P. A. Bandettini, "Representational similarity analysis-connecting the branches of systems neuroscience," *Frontiers in systems neuroscience*, vol. 2, p. 4, 2008.

[2] B. Kaneshiro, M. P. Guimaraes, H.-S. Kim, A. M. Norcia, and P. Suppes, "A representational similarity analysis of the dynamics of object processing using single-trial eeg classification," *Plos one*, vol. 10, no. 8, p. e0135697, 2015.

[3] J. V. Haxby, M. I. Gobbini, M. L. Furey, A. Ishai, J. L. Schouten, and P. Pietrini, "Distributed and overlapping representations of faces and objects in ventral temporal cortex," *Science*, vol. 293, no. 5539, pp. 2425–2430, 2001.

[4] X. Zheng, Z. Cao, and Q. Bai, "An evoked potential-guided deep learning brain representation for visual classification," in *International Conference on Neural Information Processing*. Springer, 2020, pp. 54–61.

[5] R. M. Cichy and D. Pantazis, "Multivariate pattern analysis of meg and eeg: A comparison of representational structure in time and space," *NeuroImage*, vol. 158, pp. 441–454, 2017.

[6] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in neural information processing systems*, vol. 25, pp. 1097–1105, 2012.

[7] R. T. Schirrmeister, J. T. Springenberg, L. D. J. Fiederer, M. Glasstetter, K. Eggensperger, M. Tangermann, F. Hutter, W. Burgard, and T. Ball, "Deep learning with convolutional neural networks for eeg decoding and visualization," *Human brain mapping*, vol. 38, no. 11, pp. 5391–5420, 2017.

[8] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[9] C. Spampinato, S. Palazzo, I. Kavasidis, D. Giordano, N. Souly, and M. Shah, "Deep learning human mind for automated visual classification," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 6809–6817.

[10] J. Kalafatovich, M. Lee, and S.-W. Lee, "Decoding visual recognition of objects from eeg signals based on attention-driven convolutional neural network," in *2020 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*. IEEE, 2020, pp. 2985–2990.

[11] Z. Jiao, H. You, F. Yang, X. Li, H. Zhang, and D. Shen, "Decoding eeg by visual-guided deep neural networks." in *IJCAI*, 2019, pp. 1387–1393.

[12] Y. Roy, H. Banville, I. Albuquerque, A. Gramfort, T. H. Falk, and J. Faubert, "Deep learning-based electroencephalography analysis: a systematic review," *Journal of neural engineering*, vol. 16, no. 5, p. 051001, 2019.

[13] S. Zagoruyko and N. Komodakis, "Wide residual networks," *arXiv preprint arXiv:1605.07146*, 2016.

[14] Z. Lu, H. Pu, F. Wang, Z. Hu, and L. Wang, "The expressive power of neural networks: A view from the width," *arXiv preprint arXiv:1709.02540*, 2017.

[15] A. Veit, M. Wilber, and S. Belongie, "Residual networks behave like ensembles of relatively shallow networks," *arXiv preprint arXiv:1605.06431*, 2016.

[16] B. Kaneshiro, M. P. Guimaraes, H.-S. Kim, A. M. Norcia, and P. Suppes, "EEG data analyzed in "A representational similarity analysis of the dynamics of object processing using single-trial EEG classification"," Stanford Digital Repository, 2015, Available at http://purl.stanford.edu/bq914sc3730.

[17] D. M. Tucker, "Spatial sampling of head electrical fields: the geodesic sensor net," *Electroencephalography and clinical neurophysiology*, vol. 87, no. 3, pp. 154–163, 1993.

[18] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *International conference on machine learning*. PMLR, 2015, pp. 448–456.

[19] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," *The journal of machine learning research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[20] A. S. Bobe, A. S. Alekseev, M. V. Komarova, and D. Fastovets, "Single-trial erp feature extraction and classification for visual object recognition task," in *2018 Engineering and Telecommunication (EnT-MIPT)*. IEEE, 2018, pp. 188–192.

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

[22] L. Van der Maaten and G. Hinton, "Visualizing data using t-sne." *Journal of machine learning research*, vol. 9, no. 11, 2008.

[23] T. Nguyen, M. Raghu, and S. Kornblith, "Do wide and deep networks learn the same things? uncovering how neural network representations vary with width and depth," *arXiv preprint arXiv:2010.15327*, 2020.

[24] M. Greenacre and R. Primicerio, *Multivariate analysis of ecological data*. Fundacion BBVA, 2014.

[25] Q. Gu, Z. Li, and J. Han, "Generalized fisher score for feature selection," *arXiv preprint arXiv:1202.3725*, 2012.