DEEP LEARNING WITH CONVOLUTIONAL NEURAL NETWORKS FOR DECODING AND VISUALIZATION OF EEG PATHOLOGY

R. Schirrmeister[1], L. Gemein[2], K. Eggensperger[3], F. Hutter[3] and T. Ball[1]

[1] Translational Neurotechnology Lab, Medical Center — University of Freiburg, Germany
[2] Department of Computer Science, University of Freiburg, Germany
[3] Machine Learning Lab, University of Freiburg, Germany

{robin.schirrmeister, tonio.ball}@uniklinik-freiburg.de, {gemeinl, eggenspk, fh}@cs.uni-freiburg.de

Abstract—We apply convolutional neural networks (ConvNets) to the task of distinguishing pathological from normal EEG recordings in the Temple University Hospital EEG Abnormal Corpus. We use two basic, shallow and deep ConvNet architectures recently shown to decode task-related information from EEG at least as well as established algorithms designed for this purpose. In decoding EEG pathology, both ConvNets reached substantially better accuracies (about 6% better, $\approx$85% vs. $\approx$79%) than the only published result for this dataset, and were still better when using only 1 minute of each recording for training and only six seconds of each recording for testing. We used automated methods to optimize architectural hyperparameters and found intriguingly different ConvNet architectures, e.g., with max pooling as the only nonlinearity. Visualizations of the ConvNet decoding behavior showed that they used spectral power changes in the delta (0-4 Hz) and theta (4-8 Hz) frequency range, possibly alongside other features, consistent with expectations derived from spectral analysis of the EEG data and from the textual medical reports. Analysis of the textual medical reports also highlighted the potential for accuracy increases by integrating contextual information, such as the age of subjects. In summary, the ConvNets and visualization techniques used in this study constitute a next step towards clinically useful automated EEG diagnosis and establish a new baseline for future work on this topic.

I. INTRODUCTION

Electroencephalography (EEG) is widely used in clinical practice because of its low cost and its lack of side effects due to its noninvasive nature. It is important both as a screening method as well as for hypothesis-based diagnostics, e.g., in epilepsy or stroke. One of the main limitations of using EEG for diagnostics is the required time and specialized knowledge of experts that need to be well-trained on EEG diagnostics to reach reliable results. Therefore, a machine-learning approach that aids in the diagnostic process could make EEG diagnosis more widely accessible, reduce time and effort for clinicians and potentially make diagnoses more accurate.

In recent years researchers have increasingly addressed the field of computer-aided EEG diagnosis. So far, the applications were mostly limited to specific diagnoses such as Alzheimer's disease [1], depression [2, 3], traumatic brain injuries [4], or stroke [5]. They used a large variety of machine-learning techniques, including k-nearest neighbors, random forests, support vector machines, linear discriminant analysis, logistic regression, neural networks, and more. This large variety of used methods indicates that the search for the best decoding approach for diverse types of EEG diagnosis is still ongoing.

To overcome the lack of large datasets representative of the variety of EEG-diagnosable diseases and the heterogeneity of clinical populations, the Temple University Hospital (TUH) has published an unprecedented public dataset of clinical EEG recordings [6]. From this dataset with over 16000 clinical recordings, the TUH Abnormal EEG Corpus with about 3000 recordings has been created specifically to foster the development of methods for distinguishing pathological from normal EEG. Due to its size and rich annotation, this data set has a lot of potential to contribute to the progress of automated EEG diagnosis. Baseline results on this dataset have already been reported by TUH using a convolutional neural network (ConvNet) with multiple fully connected layers that uses precomputed EEG bandpower-based features as input and reached 78.8% accuracy [7].

Deep learning approaches recently receive increasing attention in many types of machine learning problems in healthcare [8]. Deep ConvNets trained end-to end from the raw signals are a promising deep learning technique. These ConvNets exploit the hierarchical structure present in many natural signals. Recently, deep ConvNets trained end-to-end were, for example, able to more accurately diagnose skin cancer types from images than human dermatologists [9] and could segment retinal vessels better than human annotators [10].

Deep ConvNets are nowadays also being applied to EEG analyses, such as decoding task-related information from EEG [11–16]. We have recently developed and validated the Braindecode toolbox[1] for this purpose, and showed that the performance of deep ConvNets trained end-to-end is comparable to that of algorithms using hand-engineered features to decode task-related information. We also introduced novel visualization methods to gain a better understanding of ConvNet decoding behavior.

In this study, we apply deep ConvNets to the problem of distinguishing normal from pathological EEG on the TUH EEG Abnormal Corpus and show that they can reach better accuracies than the only published baseline result we are aware

[1]https://github.com/robintibor/braindecode, code to reproduce the results of this study is available under https://github.com/robintibor/auto-eeg-diagnosis-example

of, establishing a new improved baseline for future work in this field.

II. Methods
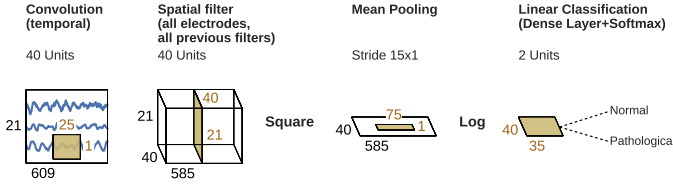
A. EEG ConvNet architectures and training

Fig. 2. **Shallow ConvNet architecture.** Conventions as in Fig. 1

We used two convolutional network architectures, for both of which we recently showed that they decode task-related information from raw time-domain EEG with at least as good accuracies as previous state-of-the-art algorithms relying on hand-engineered features [11]. Our deep ConvNet is a fairly generic architecture (Fig. 1), while our shallow ConvNet is specifically tailored to decode band-power features (Fig. 2). For more details on these models, see [11]. To accommodate the longer duration of the EEG inputs as compared to our previous study, we adapted the architectures by changing the final layer filter length so the ConvNets have an input length of about 600 input samples, which correspond to 6 seconds for the 100 Hz EEG input. Additionally, we moved the pooling strides of the deep ConvNet to the convolutional layers directly before each pooling. This modification, which we initially considered a mistake, allowed us to grow the ConvNet input length without strongly increased computation times and provided good accuracies in preliminary experiments on the training data; therefore we decided to keep it. We optimized the ConvNet parameters using stochastic gradient descent with the optimizer Adam [17]. To make best use of the available data, we trained the ConvNets on maximally overlapping time crops using cropped training as described by [11]. Code to reproduce the results of this study is available under https://github.com/robintibor/auto-eeg-diagnosis-example.

B. Decoding from reduced EEG time segments

We also evaluated the ConvNets on reduced versions of the datasets, using only the first 1, 2, 4, 8, or 16 minutes after the first minute of the recording (the first minute of the recordings was always excluded because it appeared to be more prone to artifact contamination than the later time windows). We reduced either only the training data, only the test data, or both. These analyses were carried out to study how long EEG recordings need to be for training and for predicting EEG pathologies with good accuracies.

C. Automatic architecture optimization

We also carried out a preliminary study of automatic architecture optimization to further improve our ConvNet architectures. To that end, we used the automatic hyperparameter optimization algorithm SMAC [18] to optimize architecture hyperparameters of the deep and shallow ConvNets, such as filter lengths, strides and types of nonlinearities. As the objective function to optimize via SMAC, we used 10-fold cross-validation performance obtained on the first 1500 recordings of the training data (using each fold as an instance for SMAC to speed up the optimization). We set a time limit of 3.5 hours for each configuration run on a single fold. Runs that timed out or crashed (e.g., networks configurations that did not fit in GPU memory) were scored with an accuracy of 0%.

D. Visualizations of the spectral differences between normal and pathological recordings

To understand class-specific spectral characteristics in the EEG recordings, we analyzed band powers in five frequency ranges: delta (0–4 Hz), theta (4–8 Hz), alpha (8–14 Hz), low beta (14–20 Hz), high beta (20–30 Hz) and low gamma (30–50 Hz).

For this, we performed the following steps:
1) Compute a short-term Fourier transformation with window size 12 seconds and overlap 6 seconds using a Blackman-Harris window.
2) Compute the median over all band powers of all windows and recordings in each frequency bin; independently for pathological and normal recordings.
3) Compute the log ratio of these median band powers of the pathological and normal recordings.
4) Compute the mean log ratio over all frequency bins in each desired frequency range for each electrode.
5) Visualize the resulting log ratios as a topographical map.

E. Visualizations based on the effects of amplitude perturbations on decoding decisions

Understanding the ConvNet behavior and decoding predictions is important for automatic EEG diagnosis to become practically useful as an assistive diagnosis technology. To better understand the ConvNets used in this study, we used the input-perturbation network-prediction correlation maps that we recently developed specifically for ConvNets for EEG decoding. This method shows the effect of perturbing the input amplitudes in different frequencies on the ConvNet decoding predictions. This visualization can provide spatial maps that show where on the scalp an amplitude change in a given frequency range correlates negatively or positively with the ConvNet classification decision. For more details, see [11].

F. Analysis of word frequencies in the medical reports

Furthermore, to better understand what kind of recordings are easier or harder for the ConvNets to correctly decode, we analyzed the textual clinical reports of each recording as included in the TUH Abnormal EEG Corpus. Specifically, we investigated which words were relatively more or less frequent in the incorrectly compared with the correctly predicted recordings. We performed this analysis independently for both the normal and the pathological class of recordings. Concretely, for each class, we first computed the relative frequencies f_{i-} for each word w_{i-} in the incorrectly predicted
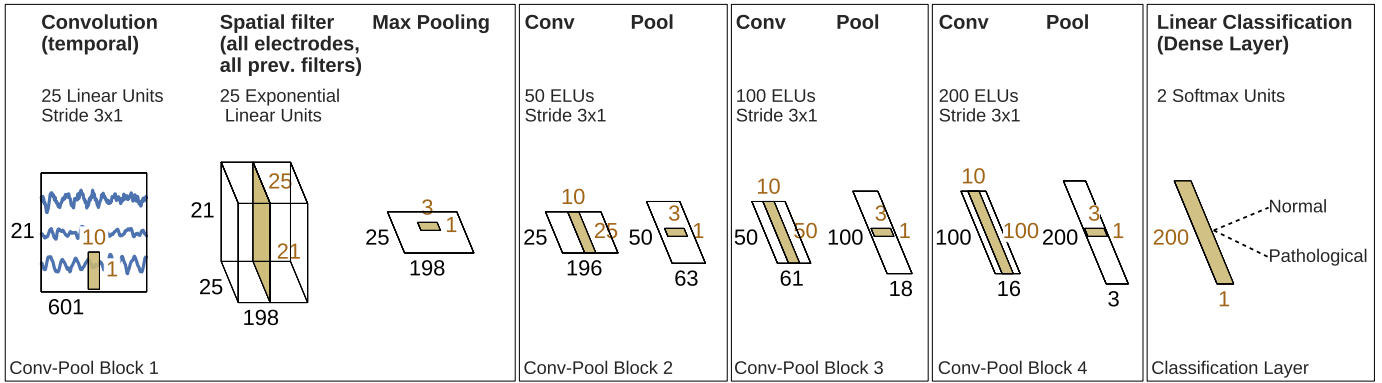
Fig. 1. **Deep ConvNet architecture.** Black cuboids: inputs/feature maps; brown cuboids: convolution/pooling kernels. The corresponding sizes are indicated in black and brown, respectively. Each spatial filter has weights for all possible pairs of electrodes with filters of the preceding temporal convolution. Note that in these schematics, proportions of maps and kernels are only approximate.

		Files	Patients
Train	Normal	1379 (50%)	1238 (58%)
	Pathological	1361(50%)	894 (42%)
	Rater Agreement[2]	2704 (99%)	2107 (97%)
	Rater Disagreement [2]	36 (1%)	25 (0%)
Evaluation	Normal	150 (54%)	148 (58%)
	Pathological	127 (46%)	105 (42%)
	Rater Agreement [2]	277 (100%)	253 (100%)
	Rater Disagreement [2]	0 (0%)	0 (0%)

[1] Obtained from https://www.isip.piconepress.com/projects/tuh_eeg/.
[2] These fields refer to the agreement between the annotator of the file and the medical report written by a certified neurologist.

recordings, i.e.: $f_{i-} = \frac{|w_{i-}|}{\sum_i |w_{i-}|}$, where $|w_{i-}|$ denotes the number of occurrences for word w_i in the incorrectly predicted recordings. We then computed the frequencies f_{i+} in the same way and computed the ratios $r_i = f_{i-}/f_{i+}$. Finally, we analyzed words with very large ratios ($\gg 1$) and very small ratios ($\ll 1$) by inspecting the contexts of their occurrences in the clinical reports. This allowed us to gain insights into which clinical/contextual aspects of the recordings correlated with ConvNets failures.

G. Dataset

The Temple University Hospital (TUH) EEG Abnormal Corpus 1.1.2 is a dataset of manually labeled normal and pathological clinical EEG recordings. It is taken from the TUH EEG Data Corpus which contains over 16000 clinical recordings of more than 10000 subjects from over 12 years [6]. The Abnormal Corpus contains 3017 recordings, 1529 of which were labeled normal and 1488 of which were labeled pathological. The Corpus was split into a training and evaluation set, see Table I.

Recordings were acquired from at least 21 standard electrode positions and with a sampling rate of in most cases 250 Hz. Per recording, there are around 20 minutes of EEG data. The inter-rater agreement on between the medical report of a certified neurologist and another annotator was 99% for the training recordings and 100% for the evaluation recordings.

H. Preprocessing

We minimally preprocessed the data with these steps:

1) Select a subset of 21 electrodes present in all recordings.
2) Remove the first minute of each recording as it contained stronger artifacts.
3) Use only up to 20 minutes of the remaining recording to speed up the computations.
4) Clip the amplitude values to the range of ± 800 mV to reduce the effects of strong artifacts.
5) Resample the data to 100 Hz to further speed up the computation.

III. Results

A. Deep and shallow ConvNets reached state-of-the-art results

	Accuracy	Sensitivity	Specificity	Crop-accuracy
Baseline	78.8	75.4	81.9	n.a.
Deep	85.4	75.1	94.1	82.5
Shallow	84.5	77.3	90.5	81.7
Linear	51.4	20.9	77.3	50.2

Results on the evaluation set of the TUH EEG Abnormal Corpus. For deep and shallow ConvNets, mean over five independent runs with different random seeds. Sensitivity and specificity are, as commonly defined, the ratio of the number of true positives to the number of all positives and the ration of the number of true negatives to the number of all negatives, respectively. Deep and shallow ConvNet outperformed the feature-based deep learning baseline [7]. n.a.: not applicable.

Both the deep and the shallow ConvNet outperformed the only results published on the TUH Abnormal EEG Corpus so far (see Table II). Both ConvNets were more than 5% better than the baseline method of a convolutional network that included multiple fully connected layers at the end and took precomputed EEG features of an entire recording as one

input [7] [2]. The ConvNets as applied here reduced the error rate from about 21% to about 15%. We also tested a linear classifier on the same 6-second inputs as our ConvNets. The linear classifier did not reach accuracies substantially different from chance (51.4%).
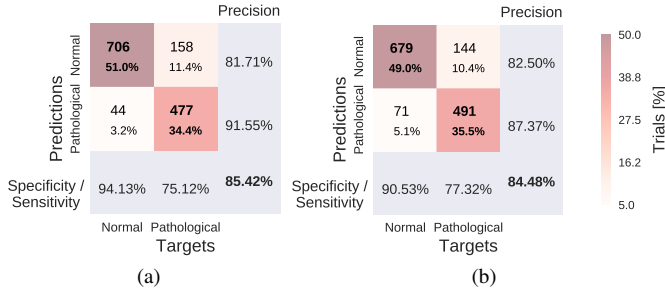

Fig. 3. **Confusion Matrices for deep and shallow ConvNets**, summed over five independent runs. Each entry of row r and column c for upper-left 2x2-square: Number of trials of target r predicted as class c (also written in percent of all trials). Bold diagonal corresponds to correctly predicted trials for both classes. Percentages and colors indicate fraction of trials in each cell relative to all trials. The lower-right value: overall accuracy. The first two values in the bottom row correspond to sensitivity and specificity. Rightmost column corresponds to precision defined as the number of trials correctly predicted for class r/number of trials predicted as class r. .

Both of our ConvNets made more errors on the pathological recordings, as can be seen from Fig. 3. Both ConvNets reached a specificity of above 90% and a sensitivity of about 75-78%. Confusion matrices between both approaches were very similar. Relative to the baseline, they reached a similar sensitivity (0.3% smaller for the deep ConvNet, 1.9% higher for the shallow ConvNet), and a higher specificity (12.2% higher for the deep ConvNet and 8.6% higher for the shallow ConvNet).

Interestingly, both of our ConvNet architectures already reached higher accuracies than the baseline when evaluating single predictions from 6-second crops. The average per-crop accuracy of individual predictions was only about 3% lower than average per-recording accuracy (averaged predictions of all crops in a recording). Furthermore, the individual prediction accuracies were already about 3% higher than the per-recording accuracies of the baseline. This implies that predictions with high accuracies can be made from just 6 seconds of EEG data.

B. Deep ConvNet reached best accuracies using only 1 minute per test-recording

Deep ConvNets already reached their best trialwise accuracies with only one minute of data used for the prediction. While the reduction of the amount of length of the training data led to crop- and trialwise accuracy decreases on the test data, reductions in the test data did not have such an effect (see Fig. 4). Remarkably, both crop- and trialwise accuracies slightly decreased when going from 1 minute to 2 or 4 minutes of
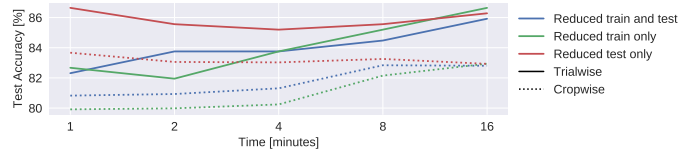

Fig. 4. **Results on reduced datasets for deep ConvNet.** Train and/or test (evaluation) dataset was reduced from 20 minutes per recording to 1,2,4,8, or 16 minutes per recording, results are shown on the test set. Notably, when only reducing the duration of the test set recordings, maximal accuracies were observed when using just 1 minute. We note that these results are each based on one run only; the slightly better performance than in Table II may thus be due to noise.
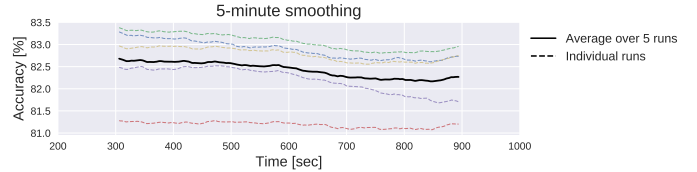

Fig. 5. **Moving average of cropwise accuracies for the deep ConvNet.** 5-minute moving averages of the cropwise accuracies of the deep ConvNet, averaged over all test set recordings. Dashed lines represent 5 individual training runs with different random seeds, solid black line represents mean over results for these runs. x-axis shows center of 5-minute averaging window.

test data. To investigate whether earlier parts of the recordings might be more informative, we also computed a 5-minute moving average of the cropwise accuracies on the test data for the Deep ConvNet trained on the full data. We show the average over all recordings for these moving averages in Fig. 5. Noticeably, as expected, accuracies slightly decreased with increasing recording time. However, the decrease is below 0.5% and thus should be interpreted cautiously.

C. Architecture optimization yielded unexpected new models
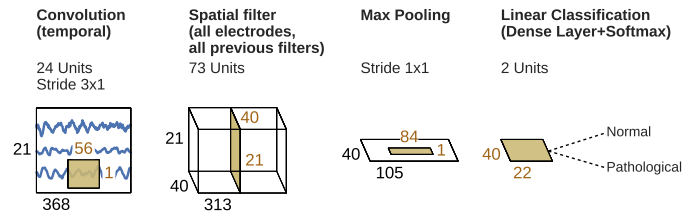

Fig. 6. **Final shallow ConvNet architecture selected by SMAC.** Conventions as in Fig. 2. Note that max pooling is the only nonlinearity SMAC decided to use.

TABLE III
DECODING ACCURACIES ON TRAINING AND TEST SET

		Train[1]		Test	
	Config	**Trial**	**Crop**	**Trial**	**Crop**
Deep	Default	84.2	81.6	85.4	82.5
	Optimized	86.3	80.9	84.5	81.3
Shallow	Default	84.5	82.1	84.5	81.7
	Optimized	85.9	80.3	83.0	79.8

[1] 10-fold cross-validation on the 1500 chronologically earliest recordings of the training data

---

[2]Note that the baseline was evaluated on an older version of the Corpus that has since been corrected to not contain the same patient in training and test recordings among other things.

The models discovered by automated architecture optimization were markedly different from our original deep and shallow ConvNets, which were designed based on the experience in a previous study on decoding of task-related information from EEG [11]. For example, the optimized architectures used only 1.8 and 3.7 seconds of EEG data for the optimized deep and shallow ConvNet, respectively, in contrast to about 6 seconds in the original versions. While the improved performance of these modified architectures for the 10-fold cross-validation on the training dataset (2.1% and 1.4% improvement for deep and shallow ConvNets, respectively) did not generalize to the evaluation set (0.9% and 1.5% deterioration for deep and shallow ConvNets, respectively, see Table III), the modifications to the original network architectures already provided interesting insights for further exploration: For example, in the case of the shallow ConvNet, the modified architecture did not use any of the original nonlinearities, but used max pooling as the only nonlinearity (see Fig. 6), a configuration we had not considered in our manual search so far.

D. Power spectra and ConvNet visualizations

Before moving to ConvNet visualization, we examined the spectral power changes of pathological compared to normal recordings. Power was broadly increased for the the pathological class in the low frequency bands (delta and theta range) and decreased in the beta and low gamma ranges (Fig. 7a). Alpha power was decreased for the occipital electrodes and increased for more frontal electrodes.

Scalp maps of the input-perturbation effects on predictions for the pathological class for the different frequency bands showed effects consistent with the power spectra in Fig. 7a. Both networks strongly relied on the lower frequencies in the delta and theta frequency range for their decoding decisions.

E. Insights from the textual reports of the clinicians

Most notably, "small" and "amount" had a much larger word frequency (15.5 times larger) in the incorrectly predicted pathological recordings compared with the correctly predicted pathological recordings. Closer inspection showed this is very sensible, as "small amount" was often used to describe more subtle EEG abnormalities ("small amount of temporal slowing", "Small amount of excess theta", "Small amount of background disorganization", "A small amount of rhythmic, frontal slowing"), as this subtlety of changes was likely the cause of the classification errors.

Secondly, other words with a notably different frequency were "age" (9.7 times larger) and "sleep" (3 occurrences in 630 words of texts of incorrectly predicted recordings, not present in texts of correctly predicted recordings). Both typically indicate the clinician used the age of the subject or the fact that they were (partially) asleep during the recording to interpret the EEG ("Somewhat disorganized pattern for age", "Greater than anticipated disorganization for age.", "A single generalized discharge noted in stage II sleep."). Obviously, our ConvNets trained only on EEG do not have access to this context information, leaving them at a disadvantage compared to the clinicians and highlighting the potential of including contextual cues such as age or vigilance in the training/decoding approach.

Inspection of the textual records of misclassified normal recordings did not provide much insight, as they are typically very short (e.g., "Normal EEG.", "Normal EEG in wakefulness.").

Finally, consistent with the strong usage of the delta and theta frequency range by the ConvNets as seen in the input-perturbation network-prediction correlation maps (Fig. 7), "slowing" and "temporal" are the 6th and 10th most frequently occurring words in the textual reports of the pathological recordings, while never occurring in the textual reports of the normal recordings (irrespective of correct or incorrect predictions).

IV. Discussion

To the best of our knowledge, the ConvNet architectures used in this study achieved the best accuracies published so far on the TUH EEG Abnormal Corpus. The architectures used were only very slightly modified versions of ConvNet architectures that we previously introduced to decode task-related information. This suggests that these architectures might be broadly applicable both for physiological and clinical EEG. The identification of all-round architectures would greatly simplify the application of deep learning to EEG decoding problems and expand their potential use cases.

Remarkably, the ConvNets already reached good accuracies based on very limited time segments of the EEG recordings. Further accuracy improvements could thus be possible with improved decoding models that can extract and integrate additional information from longer timescales. The exact nature of such models, as well as the amount of EEG they would require, remains to be determined. More accurate decoding models could either be ConvNets that are designed to intelligently use a larger input length or recurrent neural networks, since these are known to inherently work well for data with information both on shorter and longer term scales. Furthermore, combinations between both approaches, for example using a recurrent neural network on top of a ConvNet, as they have been used in other domains like speech recognition [19–21], are promising.

Our automated architecture optimization provided interesting insights by yielding configurations that were markedly different from our hand-engineered architectures, yet reached similar accuracies. Since the marked improvements in training performance did not improve the evaluation accuracies in this study, in future work, we plan to use more training recordings in the optimization and study different cross-validation methods to also improve evaluation accuracies. A full-blown architecture search [22–26] could also further improve accuracy. With such improved methods it would also be important not only to decode pathological vs. normal EEG in a binary fashion, but to also evaluate the possibility to derive more fine-grained clinical information, such as the type of pathological change (slowing, asymmetry, etc) or the likely underlying disorder (such as epilepsy).
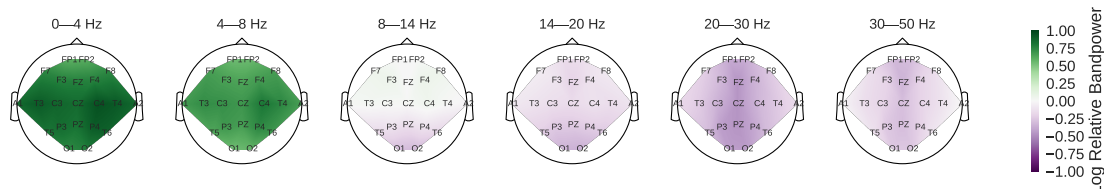
(a) Pathological vs. normal relative spectral bandpower differences for the training set. Shown is the logarithm of the ratio of the median bandpower of the pathological vs. normal (according to the experts' ratings) EEG recordings.
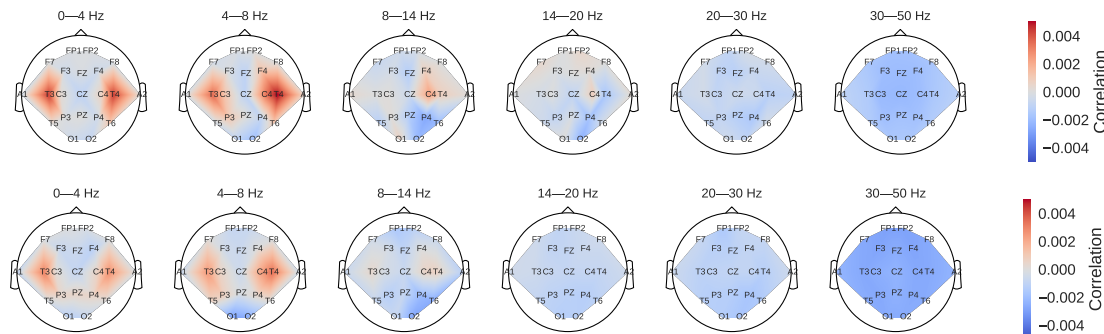
(b) Input-perturbation network-prediction correlation maps for the deep (top) and shallow (bottom) ConvNet. Correlation of predictions for the pathological class with amplitude perturbations. Scalp maps revealed for example a bilateral positive correlation for the delta and theta frequency ranges and a spatially more broadly distributed negative correlation for the beta and low gamma frequency ranges, indicating that the ConvNets used these frequency components in their decisions

Fig. 7. **Spectral power differences and input-perturbation network-prediction correlation maps.**

Any of these or other improvements might eventually bring the machine-learning decoding performance of pathological EEG closer to human-level performance. Since clinicians make their judgments from patterns they see in the EEG and other available context information, there is no clear reason why machine learning models with access to the same information could not reach human-level accuracy. This human-level performance is a benchmark for decoding accuracies that does not exist for other brain-signal decoding tasks, e.g. in decoding task-related information for brain-computer interfaces, where there is inherent uncertainty what information is even present in the EEG and no human-level benchmark exists.

Our perturbation visualizations of the ConvNets' decoding behavior showed that they used spectral power changes in the delta (0-4 Hz) and theta (4-8 Hz) frequency range, particularly from temporal EEG channels, possibly alongside other features (Fig. 7). This observation is consistent both with the expectations implied by the spectral analysis of the EEG data (Fig. 7a) and by the textual reports that frequently mentioned "temporal" and "slowing" with respect to the pathological samples, but never in the normal ones. Our perturbation visualization showed results that were consistent with expectations that the ConvNets would use the bandpower differences between the classes that were already visible in the spectra to perform their decoding. Similarly, the textual reports also yielded plausible insights, e.g., that "small amounts" of abnormalities as indicated in the written clinical reports were more difficult for the networks to decode correctly. Additionally, inspection of the textual reports also emphasized the importance of integrating contextual information such as the age of the subject.

Still, to yield more clinically useful insights and diagnosis explanations, further improvements in ConvNet visualizations are needed. Deep learning models that use an attention mechanism might be more interpretable, since these models can highlight which parts of the recording were most important for the decoding decision. Other deep learning visualization methods like recent saliency map methods [27, 28] to explain individual decisions or conditional generative adversarial networks [29, 30] to understand what makes a recording pathological or normal might further improve the clinical benefit of deep learning methods that decode pathological EEG.

Conclusion

In summary, the deep ConvNets as presented in this study yielded the best accuracies published so far on the largest available dataset for decoding EEG pathology and by that, made a next step towards clinically useful automated EEG diagnosis.

REFERENCES

[1] C. Lehmann, T. Koenig, V. Jelic, L. Prichep, R. E. John, L.-O. Wahlund, Y. Dodge, and T. Dierks, "Application and comparison of classification algorithms for recognition of Alzheimer's disease in electrical brain activity (EEG)," *Journal of Neuroscience Methods*, vol. 161, no. 2, pp. 342–350, Apr. 2007. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0165027006005425

[2] H. Cai, X. Sha, X. Han, S. Wei, and B. Hu, "Pervasive EEG diagnosis of depression using Deep Belief Network with three-electrodes EEG collector," in *2016 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, Dec. 2016, pp. 1239–1246.

[3] B. Hosseinifard, M. H. Moradi, and R. Rostami, "Classifying depression patients and normal subjects using machine learning techniques and nonlinear features from EEG signal," *Computer Methods and Programs in Biomedicine*, vol. 109, no. 3, pp. 339–345, Mar. 2013.

[4] B. Albert, J. Zhang, A. Noyvirt, R. Setchi, H. Sjaaheim, S. Velikova, and F. Strisland, "Automatic EEG Processing for the Early Diagnosis of Traumatic Brain Injury," *Procedia Computer Science*, vol. 96, pp. 703–712, Jan. 2016. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1877050916320646

[5] E. P. Giri, M. I. Fanany, and A. M. Arymurthy, "Ischemic Stroke Identification Based on EEG and EOG using 1d Convolutional Neural Network and Batch Normalization," *arXiv:1610.01757 [cs]*, Oct. 2016, arXiv: 1610.01757. [Online]. Available: http://arxiv.org/abs/1610.01757

[6] I. Obeid and J. Picone, "The Temple University Hospital EEG Data Corpus," *Frontiers in Neuroscience*, vol. 10, May 2016. [Online]. Available: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4865520/

[7] S. Lopez, "Automated Identification of Abnormal EEGs," MS Thesis, Temple University, 2017. [Online]. Available: http://www.isip.piconepress.com/publications/ms_theses/2017/abnormal

[8] R. Miotto, F. Wang, S. Wang, X. Jiang, and J. T. Dudley, "Deep learning for healthcare: review, opportunities and challenges," *Briefings in Bioinformatics*, May 2017.

[9] A. Esteva, B. Kuprel, R. A. Novoa, J. Ko, S. M. Swetter, H. M. Blau, and S. Thrun, "Dermatologist-level classification of skin cancer with deep neural networks," *Nature*, vol. 542, no. 7639, pp. 115–118, Feb. 2017. [Online]. Available: http://www.nature.com/nature/journal/v542/n7639/full/nature21056.html?foxtrotcallback=true

[10] K. K. Maninis, J. Pont-Tuset, P. Arbeláez, and L. V. Gool, "Deep Retinal Image Understanding," in *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2016.

[11] 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*, Aug. 2017. [Online]. Available: http://dx.doi.org/10.1002/hbm.23730

[12] M. Hajinoroozi, Z. Mao, T.-P. Jung, C.-T. Lin, and Y. Huang, "EEG-based prediction of driver's cognitive performance by deep convolutional neural network," *Signal Processing: Image Communication*, vol. 47, pp. 549–555, Sep. 2016. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0923596516300832

[13] V. J. Lawhern, A. J. Solon, N. R. Waytowich, S. M. Gordon, C. P. Hung, and B. J. Lance, "EEGNet: A Compact Convolutional Network for EEG-based Brain-Computer Interfaces," *arXiv:1611.08024 [cs, q-bio, stat]*, Nov. 2016, arXiv: 1611.08024. [Online]. Available: http://arxiv.org/abs/1611.08024

[14] R. Manor and A. B. Geva, "Convolutional Neural Network for Multi-Category Rapid Serial Visual Presentation BCI," *Frontiers in Computational Neuroscience*, vol. 9, p. 146, 2015.

[15] S. Stober, A. Sternin, A. M. Owen, and J. A. Grahn, "Deep Feature Learning for EEG Recordings," *arXiv:1511.04306 [cs]*, Nov. 2015, arXiv: 1511.04306. [Online]. Available: http://arxiv.org/abs/1511.04306

[16] Y. R. Tabar and U. Halici, "A novel deep learning approach for classification of EEG motor imagery signals," *Journal of Neural Engineering*, vol. 14, no. 1, p. 016003, 2017. [Online]. Available: http://stacks.iop.org/1741-2552/14/i=1/a=016003

[17] D. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," in *arXiv:1412.6980 [cs]*, 2015, arXiv: 1412.6980. [Online]. Available: http://arxiv.org/abs/1412.6980

[18] F. Hutter, H. H. Hoos, and K. Leyton-Brown, "Sequential Model-Based Optimization for General Algorithm Configuration," in *Proceedings of the conference on Learning and Intelligent OptimizatioN (LION 5)*, Jan. 2011, pp. 507–523.

[19] X. Li and X. Wu, "Constructing long short-term memory based deep recurrent neural networks for large vocabulary speech recognition," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Apr. 2015, pp. 4520–4524.

[20] T. N. Sainath, O. Vinyals, A. Senior, and H. Sak, "Convolutional, Long Short-Term Memory, fully connected Deep Neural Networks," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Apr. 2015, pp. 4580–4584.

[21] H. Sak, A. Senior, K. Rao, and F. Beaufays, "Fast and accurate recurrent neural network acoustic models for speech recognition," *arXiv preprint arXiv:1507.06947*, 2015.

[22] H. Mendoza, A. Klein, M. Feurer, J. Springenberg, and F. Hutter, "Towards Automatically-Tuned Neural Networks," in *ICML 2016 AutoML Workshop*, Jun. 2016.

[23] R. Miikkulainen, J. Liang, E. Meyerson, A. Rawal, D. Fink, O. Francon, B. Raju, H. Shahrzad, A. Navruzyan, N. Duffy, and B. Hodjat, "Evolving Deep Neural Networks," *arXiv:1703.00548 [cs]*, Mar. 2017, arXiv: 1703.00548. [Online]. Available: http://arxiv.org/abs/1703.00548

[24] E. Real, S. Moore, A. Selle, S. Saxena, Y. L. Suematsu, J. Tan, Q. Le, and A. Kurakin, "Large-Scale Evolution of Image Classifiers," *arXiv:1703.01041 [cs]*, Mar. 2017, arXiv: 1703.01041. [Online]. Available: http://arxiv.org/abs/1703.01041

[25] B. Zoph and Q. V. Le, "Neural Architecture Search with Reinforcement Learning," *arXiv:1611.01578 [cs]*, Nov. 2016, arXiv: 1611.01578. [Online]. Available: http://arxiv.org/abs/1611.01578

[26] B. Zoph, V. Vasudevan, J. Shlens, and Q. V. Le, "Learning Transferable Architectures for Scalable Image Recognition," *arXiv:1707.07012 [cs]*, Jul. 2017, arXiv: 1707.07012. [Online]. Available: http://arxiv.org/abs/1707.07012

[27] P.-J. Kindermans, K. T. Schütt, M. Alber, K.-R. Müller, and S. Dähne, "PatternNet and PatternLRP - Improving the interpretability of neural networks," *CoRR*, vol. abs/1705.05598, 2017. [Online]. Available: http://arxiv.org/abs/1705.05598

[28] G. Montavon, W. Samek, and K.-R. Müller, "Methods for Interpreting and Understanding Deep Neural Networks," *CoRR*, vol. abs/1706.07979, 2017. [Online]. Available: http://arxiv.org/abs/1706.07979

[29] M. Mirza and S. Osindero, "Conditional generative adversarial nets," *arXiv preprint arXiv:1411.1784*, 2014.

[30] J. T. Springenberg, "Unsupervised and semi-supervised learning with categorical generative adversarial networks," *arXiv preprint arXiv:1511.06390*, 2015.