Zubarev, Ivan; Zetter, Rasmus; Halme, Hanna Leena; Parkkonen, Lauri

Adaptive neural network classifier for decoding MEG signals

Published in:
NeuroImage

DOI:
10.1016/j.neuroimage.2019.04.068

Published: 15/08/2019

Document Version
Publisher's PDF, also known as Version of record

NeuroImage 197 (2019) 425–434

Contents lists available at ScienceDirect

NeuroImage

journal homepage: www.elsevier.com/locate/neuroimage

Adaptive neural network classifier for decoding MEG signals

Ivan Zubarev [a,*], Rasmus Zetter [a], Hanna-Leena Halme [a], Lauri Parkkonen [a,b]

[a] Department of Neuroscience and Biomedical Engineering, Aalto University School of Science, FI-00076, Aalto, Finland
[b] Aalto NeuroImaging, Aalto University, FI-00076, Aalto, Finland

ARTICLE INFO

Keywords:
Convolutional neural network
Magnetoencephalography
Brain–computer interface

ABSTRACT

We introduce two Convolutional Neural Network (CNN) classifiers optimized for inferring brain states from magnetoencephalographic (MEG) measurements. Network design follows a generative model of the electromagnetic (EEG and MEG) brain signals allowing explorative analysis of neural sources informing classification. The proposed networks outperform traditional classifiers as well as more complex neural networks when decoding evoked and induced responses to different stimuli across subjects. Importantly, these models can successfully generalize to new subjects in real-time classification enabling more efficient brain–computer interfaces (BCI).

1. Introduction

Deep Neural Networks have revolutionized many domains such as image recognition and natural language processing. To date, their application in the analysis of electro- and magnetoencephalographic (EEG and MEG) data has been limited by several domain-specific factors.

First of all, electromagnetic brain signals are characterized by very low signal-to-noise ratio (SNR). Here, the term "noise" is understood widely and includes external interference, physiological (e.g. cardiac or oculomotor) artifacts as well as background brain activity unrelated to the studied phenomena. SNR in single-trial EEG and MEG measurements is typically assumed to be < 1 for evoked responses and ≈ 1 for oscillatory activity, which puts these data to stark contrast with those in traditional applications of deep learning. Typical EEG/MEG analysis employs a wide range of techniques to increase the SNR, e.g. spatial and temporal filtering, averaging a large number of observations, source-separation algorithms, and other complex feature extraction methods (e.g. wavelet transform). Thus, efficient noise suppression is required for high-accuracy classification of EEG/MEG signals.

Second, these data have a complex, high-dimensional spatiotemporal structure. Modern EEG and MEG systems comprise several hundreds of sensors capable of sampling brain activity with sub-millisecond temporal resolution. In the case of MEG, these sensors may also measure different components of the neuromagnetic field. On one hand, this multitude of data points enables sophisticated analysis methods to extract finer details of brain function. On the other hand, manual analysis and interpretation of these data become increasingly complex and time-consuming. Machine-learning algorithms can be of great help in such tasks but the mere classification result is often not sufficient; ideally, the experimenter should understand why the algorithm is able to classify the data, i.e., the learned model should be interpretable in neurophysiological terms. A model able to reliably identify those neural sources that contribute to the discrimination between given experimental conditions could enable efficient exploitative analysis of these complex data sets and ultimately allow more complex experimental designs.

Finally, deep-learning models require large numbers of training samples to perform optimally. In a typical EEG/MEG or BCI experiment, however, time constraints and data acquisition costs limit the sample sizes severely. Open datasets usually comprise data from a large number of individuals, providing a promising strategy to overcome this limitation, but in this case, the classifier needs to be robust to high inter-individual variability stemming from differences in cortical anatomy, mappings of function to structure and physiological state.

Taken together, these factors may easily lead to over-fitting (especially in more complex models) and poor interpretability of the findings. To address these challenges, we propose a Convolutional Neural Network (CNN) whose architecture is based on a generative model of non-invasive electromagnetic measurements of the brain activity (Daunizeau and Friston, 2007).

This network utilizes spatiotemporal structure in the MEG data to extract informative components of the MEG signal from the noisy observations. Since the model structure reflects our understanding of the data generation process, the extracted components can be interpreted in terms of the underlying neural activity. Specifically, this model assumes

* Corresponding author.
E-mail address: ivan.zubarev@aalto.fi (I. Zubarev).

https://doi.org/10.1016/j.neuroimage.2019.04.068
Received 6 November 2018; Received in revised form 30 March 2019; Accepted 25 April 2019
Available online 4 May 2019

that the MEG measurements are generated by a linear (spatial) mixture of a limited number of latent sources, which evolve non-linearly over time. Each of these sources is characterized by a spatial distribution and a time course, which are relatively stable within each individual but may vary across individuals. For example, the spatial topography of a sensory evoked response may vary across subjects due to small differences in cortical anatomy but its latency can be relatively constant. Conversely, the phase of an event-related (induced) oscillatory response may vary considerably across trials while its spectral content and spatial distribution remain the same.

We demonstrate that utilizing a generative model makes the algorithm robust to inter-individual differences in spatial, temporal and spectral properties of the signal. Critically, such across-subject generalization makes it possible to train the model on pooled data from multiple subjects and successfully apply it to new subjects. Here, we applied this neural network to classify evoked responses to visual, auditory and somatosensory stimuli (5 classes; Experiment 1) and induced responses to hand motor imagery (3 classes; Experiment 2). We verified the feasibility of this approach by running the algorithm in real time brain–computer interface (BCI) experiment (2 classes; Experiment 3). Finally, we tested the algorithm on a large open MEG Cam-CAN dataset and provide the code to demonstrate the reproducibility of the results (2 classes; Experiment 4).

2. Methods

2.1. Generative latent state-space model for MEG signal

Magnetoencephalography (MEG) is a non-invasive, time-resolved technique for measuring electric brain activity through the magnetic field it generates (Hamalainen et al., 1993). The MEG signal is complementary to that of electroencephalography (EEG), in which the potential distribution caused by electric brain activity is measured using electrodes placed on the scalp. MEG is considered to have higher spatial resolution than EEG, as the EEG signal is distorted by the heterogeneous conductivity profile of head tissues to a much larger extent than the MEG signal (see e.g. Baillet, 2017).

MEG data typically include 1) evoked responses (event-related fields; ERF) that are phase-locked to specific sensory, cognitive or motor events, and 2) induced modulations of ongoing oscillatory brain activity that is not phase-locked to external events. MEG measurements are typically contaminated by noise and interference originating from external sources as well as by ongoing unrelated brain activity. Single-trial ERFs typically have a signal-to-noise ratio (SNR) ~ 1.

An MEG measurement can be represented by an $n\times t$ data matrix $\mathbf{X}$ containing measurements from n sensors (magnetometers or gradiometers; typ. 200–300) at t time points sampled at a high temporal frequency (typ. $\sim$ 1000 Hz). These data have a complex spatiotemporal structure because activation of a single neural source is picked up by several sensors at different spatial locations and these signals exhibit temporal correlations. Thus, simultaneously active neural sources result in a high degree of linear spatial mixing as well as non-linear temporal dependencies in the measured data. Fortunately, dense spatial and temporal sampling allow efficient source-separation by utilizing local spatiotemporal correlations (Cardoso, 1998). We argue that an effective approach towards decoding brain states should take into account these properties of the signal.

The proposed network architecture is broadly based on an extension of a model describing the generation of MEG signal (Daunizeau and Friston, 2007). The model is motivated by the assumption that a single event-related MEG observation $\mathbf{X} \in \mathbb{R}^{n\times t}$ is generated by a mixture of k latent sources s such that at each time point t

$$\mathbf{x}_t = \mathbf{C}\mathbf{s}_t + \boldsymbol{\varepsilon} \tag{1}$$

where $\mathbf{C}$ is an $n \times k$ matrix describing the spatial mixing of the k underlying latent sources whose time courses $\mathbf{s}$ are the rows of a matrix $\mathbf{S}$ and ε is additive Gaussian white observation noise. The number of such latent sources is small relative to the number of MEG channels (i.e., $k \ll n$). These k sources evolve in time in a structured way and may or may not be statistically dependent. Importantly, we do not restrict the definition of a latent source to neural activity as some of these sources may correspond to e.g. ocular or muscular artifacts. In the simplest case, the mapping from $\mathbf{S}$ to $\mathbf{X}$ is linear, and non-linearities in temporal dynamics of $\mathbf{S}$ can be locally approximated by a linear autoregressive (AR) model $\mathbf{A}$ of order L. Given the fast temporal sampling of MEG, such local linearity in the temporal domain is a reasonable assumption. Thus,

$$\mathbf{s}_t = \sum_{l=1}^{L} \mathbf{A}_l \mathbf{s}_{t-l} \tag{2}$$

Assuming no interaction between the latent sources, $\mathbf{A}_l$ becomes a diagonal submatrix, where the k-th diagonal element of each of the L submatrices form a $L-$ th order univariate AR model of the temporal dynamics of the k-th source. The coefficients of these AR models contain information about spectral properties of the sources. Furthermore, if $\mathbf{A_l}$ is a full submatrix, its off-diagonal elements model the interactions between the latent sources, leading to $\mathbf{A}$ being a full L-th order Vector-Autoregressive (VAR) model of k interacting sources.

2.2. Network architecture

The proposed classifier incorporates the assumptions of the generative model described above into the discriminative neural network model (see Fig 1). The first and the second layers of the network learn spatial and temporal filters, which extract a compact representation of MEG signal features contributing to the discrimination between the classes. These features make use of spatial and temporal correlations in the data to suppress noise and to obtain sufficient separation between the simultaneously active neural sources. The l_1 -regularized output layer assigns non-zero weights only to features that are informative for each class of the stimuli. Finally, the spatial and temporal filters are further optimized by back-propagating errors from the output-layer nodes with non-zero weights.

Input layer: Spatial de-mixing. The linear input layer trains a set of spatial filters $\mathbf{W}$ with each column $\mathbf{w}_k$ extracting a timecourse of k-th latent source. These filters are related to the spatial activation patterns $\mathbf{C}$ of the latent sources in the generative model (Eq. (1)) via

$$\mathbf{W}^T\mathbf{x}_t = \widehat{\mathbf{s}}_\mathbf{t} \tag{3}$$

$$\mathbf{C} = \boldsymbol{\Sigma}_x \mathbf{W} \boldsymbol{\Sigma}_{\widehat{s}}^{-1} \tag{4}$$

where $\boldsymbol{\Sigma}_x$ is the spatial data covariance and $\boldsymbol{\Sigma}_{\widehat{s}}^{-1}$ is the precision matrix of

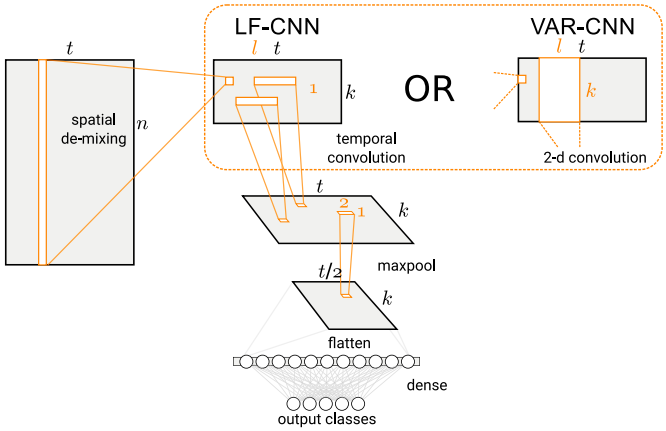

Fig. 1. Architecture of the two variants of the CNN.

the latent time courses(Haufe et al., 2014).

The input layer can be viewed as a linear projection layer performing dimensionality reduction in the spatial domain. The weights of the input layer are applied to each time point in the MEG epoch by computing dot product between **W** and the whole MEG epoch **X**. This layer implements a set of spatial filters that project the channel data onto a k-dimensional subspace. This layer has several functions: (1) it obtains a lower-dimensional and spatially-decorrelated representation of the signal time courses, (2) it learns and projects out irrelevant activity such as physiological artifacts, and (3) it provides an interpretable linear mapping from model weights to the channels in the original signal space. These weights can be used to extract spatial activation patterns (Haufe et al., 2014), or topographic maps, corresponding to neural sources informing the classifiers. Alternatively, this layer can be viewed as a spatial convolution layer with 'valid' padding applied to all channels at each time-point. Here, we prefer to refer to it as a linear projection layer for ease of interpretation. Linear projections similar to those for the function (2) above are typically used to suppress ocular and cardiac artifacts in MEG data. However, in contrast to e.g. independent component analysis (ICA), the projection basis is defined by back-propagation without introducing an explicit assumption of statistical independence.

Temporal convolution layer: Extraction of activation dynamics. This layer operates on the time courses of the latent sources (corresponding to the rows of **S**) and implements a filter, which extracts a temporal activation pattern of the informative neural event (e.g. a peak of an evoked response).

We used two variants of this layer. The simpler one (LF-CNN) applies k separate 1-dimensional convolution filters of the l-th order to the time courses of the k spatial components produced by the input layer. The model assumes that these time courses do not interact and that they have unique spectral fingerprints. This layer variant can be viewed as applying linear finite-impulse-response filters (hence LF) that specifically capture the informative features in time courses of each spatial component. Thus, each spatial pattern from the input layer has a corresponding impulse-response function learned by the temporal convolution layer. These functions extract the frequency bands specific to each component.

The more complex variant allows estimating the interactions between the spatial components and can be viewed as a vector autoregressive model (VAR-CNN) of the component time courses. This structure is implemented by applying k spatiotemporal convolution kernels of shape $l \times k$ and it results in a larger set of trainable parameters (lk^2) in this layer.

For both variants of this layer, the convolution is followed by a non-linearity using rectified linear units (*ReLU*) and a max-pooling layer with a pooling factor of 2 and a stride of 2 applied to the time dimension. Temporal max-pooling provides robustness against variation in the latency of the informative responses across subjects.

Output layer: Imposing sparsity. The mapping from the temporal convolution layer to the output is provided by a single, fully-connected layer followed by a soft-max normalization. The total number of (flattened) inputs to this layer $n_{inputs} = k*t/p$, where k is the number of latent components, and t/p is a number of time points after pooling with a factor of p. This final layer outputs a vector of logits with length equal to the number of classes (m). Thus, the weight matrix of the final layer has dimensions $n_{inputs} \times m$, with each column corresponding to the contribution of all the features to a given class. Sparsity is imposed on the weights of the output layer using l_1 -norm regularization suppressing most of the activity that is unrelated to the classification. Thus, as we show in the next section, exploring the non-zero weights in the output layer thus allows to identify the temporal and spatial patterns which contribute to the discrimination.

2.3. Model inspection and parameter interpretation

Interpretation of the model parameters in terms of the underlying neural activity is a desirable property. Our network design is based on a generative model of MEG signal to allow such interpretation.

To identify spatial and temporal features that contribute to assignment of a given sample to a particular class, we identified the nodes of the final classification layer containing the maximum positive weights (contributions) to each particular class. Because the weights of the final layer contain information about the contribution of all the features extracted by the previous layers to each class, we can map the index of the feature that has a maximum contribution to a given class in the output layer onto the shape of the temporal convolution layer. Thus, we can identify the latent component(s) as well as the (approximate) timepoints corresponding to this most informative feature. This information is then used to extract the corresponding spatial filter from the input layer and temporal convolutional filter from the hidden layer. Thus, a single feature in the output layer maps to the information about the spatial pattern, approximate latency and spectral properties of the latent component (see Fig. 2). This argument, however, holds only for LF-CNN, because the more complex VAR-CNN variant allows to capture non-linear interactions between the latent sources and thus cannot be interpreted in a straightforward way.

Spatial activation patterns and source estimates. Assuming statistical independence between the latent components one can extract the corresponding activation pattern by multiplying the spatial filter by the spatial covariance matrix of the data (Haufe et al., 2014). To relax the independence assumption further, one can also consider multiplying the

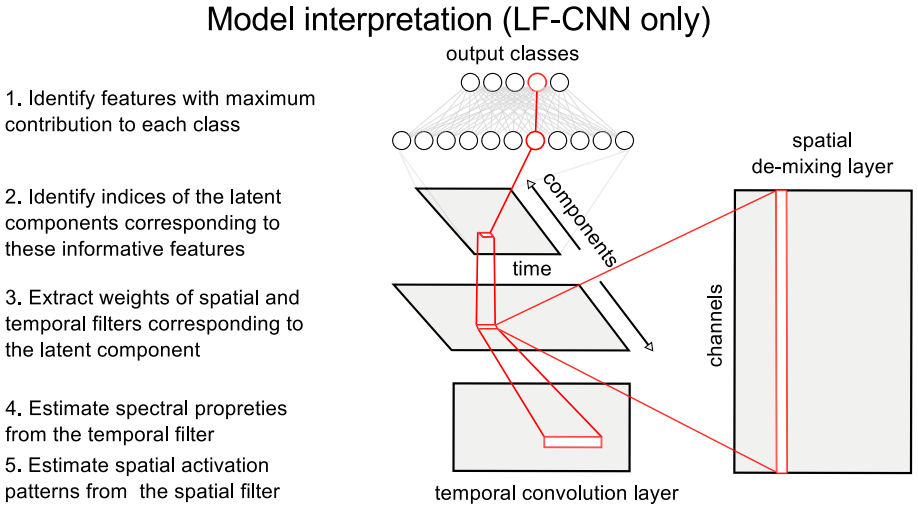

Fig. 2. Interpretation of model parameters (LF-CNN only).

resulting pattern by the precision matrix of the latent time courses that can be obtained by inverting the covariance matrix of the data, projected onto the latent component space with the spatial filters extracted during the previous step. In this study, this latter step was omitted, because it is unclear whether the spatial filters optimized to extract only the informative parts of the latent time course could be used to estimate the latent precision matrix without distortion (Supplementary Fig. 1).

The obtained sensor-level spatial pattern can further be mapped onto the individual source space using standard source estimation algorithms (Fig. 3). To demonstrate this mapping we trained an LF-CNN model on pooled data from Experiment 1 and updated it with a single training epoch performed on the data of a single held-out subject. For each of the 5 classes of stimuli, we extracted the model parameters corresponding to spatial filters of the single most informative component using the procedure described above. These patterns were compared to components that had minimal or no contribution to any of the classes as defined by the absolute sum of the corresponding weights in the output layer (Fig. 4). We then estimated the neural sources of each informative component by computing dynamic statistical parametric maps (dSPM) (Dale et al., 2000) for sources constrained onto the individual cortical surface and with orientations favoring the direction perpendicular to the local cortical surface (loose orientation constraint 0.2) as implemented in the MNE-Python software package (Gramfort et al., 2013a).

We then conducted a standard evoked-response analysis by averaging the MEG responses within each class of stimuli. We compared the features extracted from our model to the most prominent components of the corresponding averaged evoked responses in terms of their latency and spatial distribution and the corresponding source estimates. Source estimation of the averaged evoked responses utilized noise covariance estimated from a 300-ms pre-stimulus baseline whereas identity covariance was used for activation patterns.

Investigating spectral properties of extracted latent components. Since each temporal convolution filter in the hidden layer of the LF-CNN represents a univariate autoregressive model of the temporal dynamics of *k*-th latent source, we can use the coefficients of this filter to obtain an estimate of the power spectral density properties of the time course of this latent source. To test whether this approach can be used to identify oscillatory activity that is informative for classification we extracted the spatial and spectral filters from the model trained on the data from Experiment 2. Because we expected informative activity to be extended over the whole 1.5 s time window, since event-related desynchronization associated with motor imagery is not phase-locked to the stimulus onset, we used a different approach to identify the most informative latent component compared to the Experiment 1. Instead of identifying a single

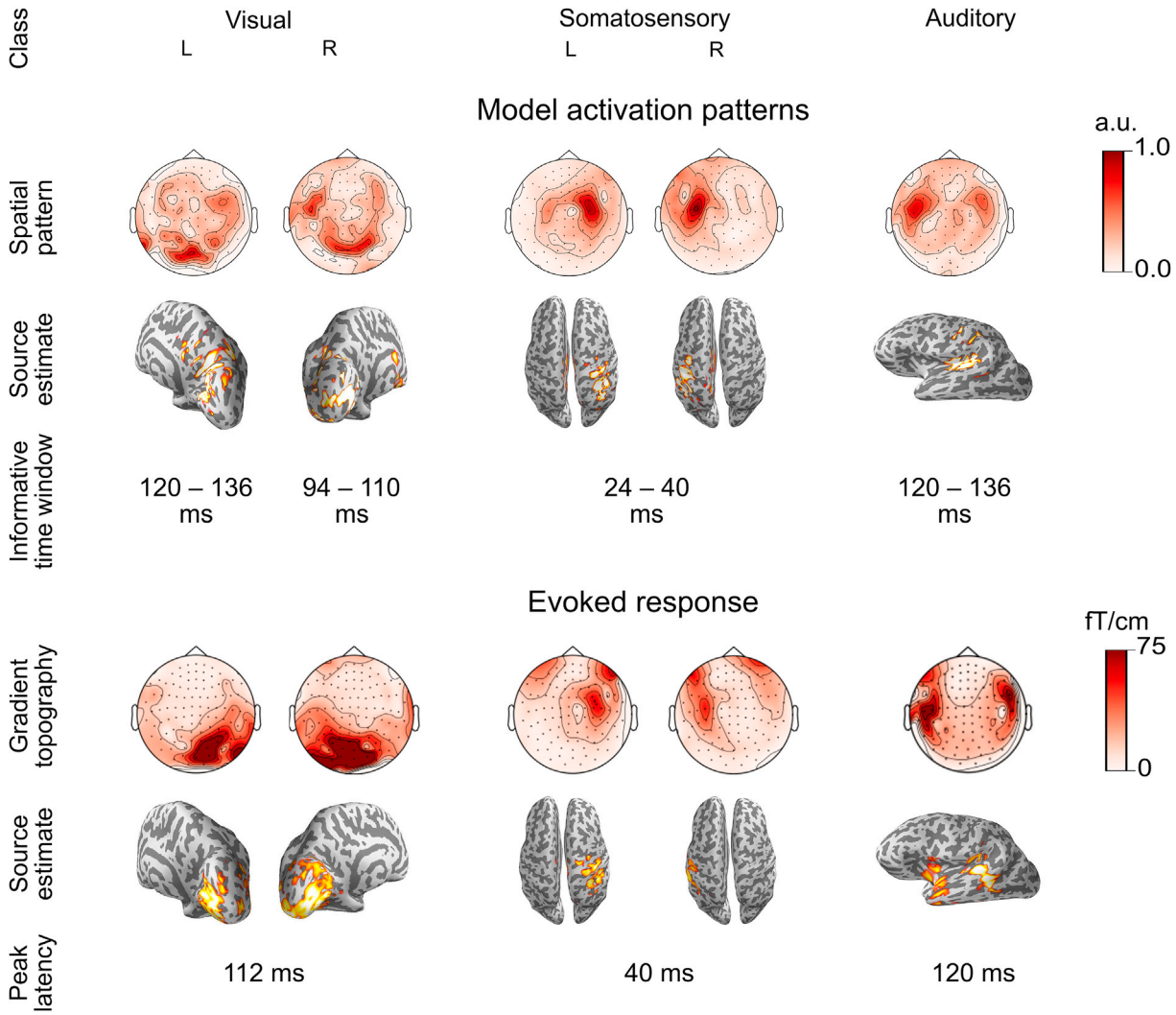

Fig. 3. Interpretation of informative LF-CNN model parameters in a representative subject from Experiment 1. **A.** Components having the maximum contribution to the decoding of each class were extracted from the model and interpreted in terms of their spatial topographies (top), source estimates (middle) and latency estimates (bottom). **B.** Spatial topographies(top), source estimates (middle) and peak latencies (bottom) of the corresponding evoked responses. Source estimates visualization is thresholded to 95% of the peak source activity.

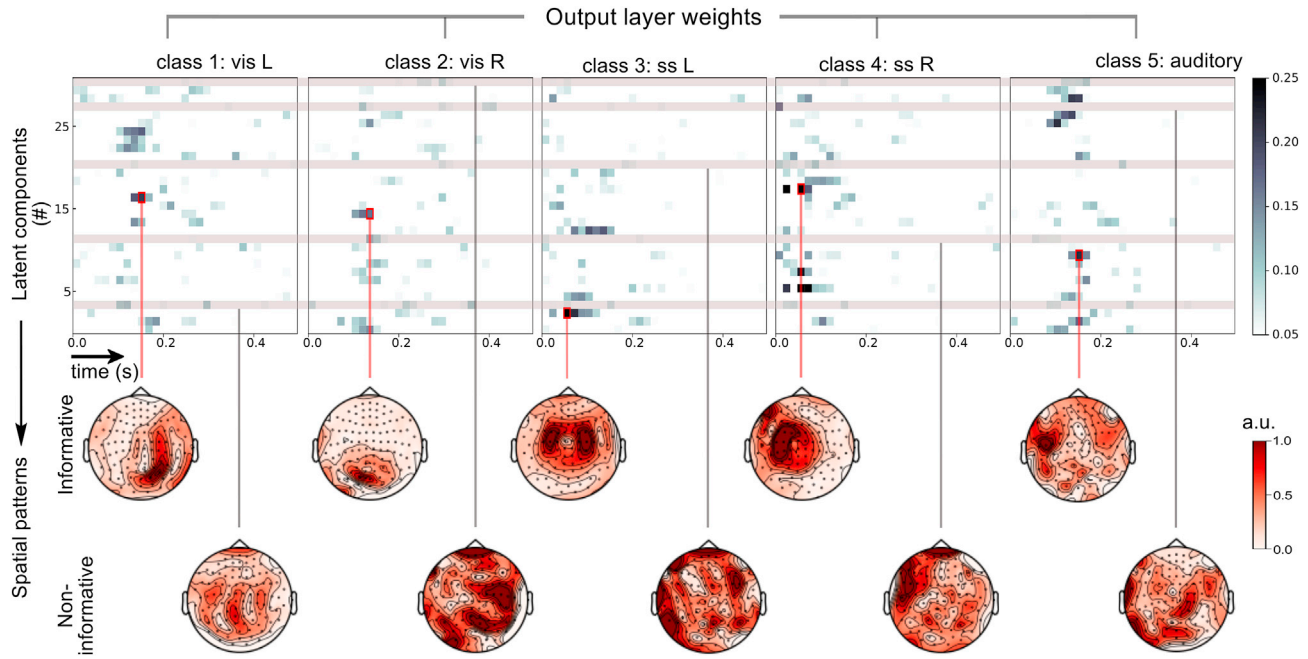

Fig. 4. Identification of informative latent components in a single representative subject from Experiment 1. The weights of the contributions of the final layer to each class are represented as a raster plot (top) with rows corresponding to the index of latent component and columns corresponding to (pooled) time points. Informative components (middle) are found by identifying single features (indicated by red boxes) with maximum positive weight for each class. Non-informative components are defined as having the minimal absolute sum of weights across all classes. Component topographies are scaled to interval [0,1] for comparability.

feature in the output layer, corresponding to the highest activation at a single time-point, we took the index of spatial component that had the largest sum of weights over all time-points. We extracted spatial patterns corresponding to two most informative latent components for each class. We also estimated the frequency content of the component's time course by computing power spectral density using the weights of the corresponding temporal convolutional filters.

2.4. Implementation and training

Design choices and hyperparameters. The neural network was implemented using the Tensorflow library (Abadi et al., 2016). The code is publicly available at https://version.aalto.fi/gitlab/zubarei1/aalto-megnet. Model development and hyperparameter tuning were performed on the data of a single randomly-picked subject from Dataset 1. The model was then applied to across-subject classification and other experiments as is. Table 1 summarizes the tested values of tunable hyperparameters.

Initialization and training. The initial values of the weight matrices were drawn from a uniform distribution following the procedure introduced by He et al. (2015).

Table 1
Tested and optimal hyperparameter values.

Parameter	Tested	Optimal
Number of latent sources	16,32,64	32
Temporal filter length	3,5,7,9,11	7
Learning rate	1·10−3, 3·10−4, 1·10−4	3·10−4
l_1 -penalty	1·10−3, 3·10−4, 1·10−4	3·10−4
Pooling	*max*	*max*
Pooling factor	2,3,5	2
Drop-out coefficient	0.25, 0.50, 0.75, 0.90	0.50
Input layer link function	*identity*, *ReLU*, *ELU*, *tanh*	*identity*
Hidden layer link function	*identity*, *ReLU*, *ELU*, *tanh*	*ReLU*
Output nonlinearity	*sigmoid*, *softmax*	*softmax*
Number of dense hidden layers	1,2	1

We initialized the bias variables to a constant value of 0.1. We used the Adam optimization algorithm with a batch size of 100 and the learning rate of $3.0 \cdot 10^{-4}$ to optimize multinomial cross-entropy between the model predictions and true labels. Higher learning rates were also used but they did not improve performance. We used an early-stopping strategy to prevent over-fitting; for every 1000 iterations, we computed the validation cost (multinomial cross-entropy) and stopped the iterations immediately if the cost function value was increasing or decreasing by less than $1.0 \cdot 10^{-5}$. The early-stopping criteria were typically met within 20 000 iterations, corresponding to a maximum training time of 32 min using a normal workstation CPU only.

Regularization. We examined several regularization approaches including drop-out, l_1 and l_2 penalties on the model weights as well as the pairwise combinations of drop-out and weight penalties. A combination of drop-out regularization applied to the output layer and l_1 penalty applied to all weight variables resulted in the highest model performance and was used with all datasets.

Performance evaluation. Since our main focus was on developing a model that generalizes across subjects, we used the leave-one-subject-out method to evaluate model performance. Thus, for a dataset of m subjects, the training (90% of all trials) and validation (10% of all trials) sets comprised pooled data from randomly-selected $m-1$ subjects. The model was then applied to the data of the held-out subject, and the following two scores were computed. As all our datasets comprised an equal number of trials for each category, we used classification accuracy as the performance metric. *Initial test accuracy* was defined as the proportion of correct predictions on the held-out subject. *Pseudo-real-time accuracy* was defined as the mean prediction accuracy in a simulated real-time design where the model predicted new observations in batches of 20 trials and was updated after each prediction. For the true real-time BCI experiment (Experiment 3), the actual BCI accuracy (with and without model updates) is reported.

2.5. Benchmark classifiers

RBF and linear SVMs. SVMs (Vapnik, 2000) are widely used in

classification of MEG data (e.g. Gramfort et al., 2013b; Westner et al., 2018). We used incremental versions of linear and radial basis function (RBF) -kernel Support Vector Machines (SVM) as benchmark classifiers. Data preprocessing, scaling and classifier training procedure for these methods were identical to the one reported for LF-CNN and VAR-CNN. Data points from all channels/timepoints were concatenated forming a single feature vector. No additional feature extraction methods were used. We refer the reader to our previous study for comparing various feature extraction methods in combination with linear classifiers for across-subject classification of MEG data (Halme and Parkkonen, 2018). Nyström RBF kernel approximation was used for incremental RBF-SVM as implemented in the Scikit-Learn package(Pedregosa et al., 2011). The SVM inverse regularization parameter C and the kernel lengthscale parameter for RBF kernel γ were set by performing a search over a 2-d grid of 5 logarithmically spaced values from 10^3 to 10^5 for C and from 10^{-2} to 10^{-7} for γ. The classifier that gave the highest validation set accuracy was evaluated on the test set.

CNNs developed for EEG classification. We used two CNN models developed for classification of EEG data. Shallow FBCSP-CNN (Schirrmeister et al., 2017) is a model inspired by Filter-Bank Common Spatial Pattern (FBCSP), a state-of-the-art method for extracting band-power features in EEG/MEG. Its architecture comprises a 1-d temporal-convolution input layer (40 filters) followed by a spatial-filter layer (40 filters) and mean pooling. The outputs of the pooling layer are then combined linearly to produce label predictions by applying the softmax function. We used the Shallow FBCSP-CNN implementation provided in the Braindecode library with default parameters, only modifying temporal filters and pooling factors to match the sampling rate of our data (125 Hz). We were not able to perform pseudo-real-time test for FBCSP-CNN due to the differences in implementation.

EEGnet (Lawhern et al., 2018) is a compact model designed specifically to optimize across-subject generalization. The model uses a combination of 1-d depth-wise and separable convolution layers (a total of 4 layers) and has been shown to generalize well across subjects in a large number of datasets. We implemented EEGNet-8 in Tensorflow following the description provided in Lawhern et al. (2018) and tested it in a simulated real-time set-up, similarly to VAR-CNN and LF-CNN.

Deep CNNs developed for image classification. As an example of general-purpose deep convolutional network, we used VGG19 –a 19-layer convolutional network and winner of ImageNet Challenge 2014 (Simonyan and Zisserman, 2015). The VGG19 architecture includes 5 blocks of convolutional layers followed by three fully-connected layers. Each of the 5 blocks includes a stack of several (2, 2, 4, 4, and 4) convolutional layers with (64, 128, 256, 512, and 512) 3×3 convolution kernels and a 2×2 max-pooling layer with stride 2. Due to the fact that the scaling of MEG data was different from that of the ImageNet dataset, we have introduced a batch-normalization layer after each block of convolutional layers to mitigate the risk of exploding or vanishing gradients. The final layer uses softmax non-linearity while all hidden layers are equipped with the ReLU non-linearity. For further details on VGG19 implementation, see Simonyan and Zisserman (2015).

3. Datasets

MEG recordings were acquired using an Elekta Neuromag Vectorview (MEGIN/Elekta Oy, Helsinki, Finland) MEG system, which includes 306 sensors at 102 positions around the head; two orthogonal planar gradiometers and a magnetometer at each position. Only data from the planar gradiometers (204 channels) were used in this work. All experiments had their respective approvals from Aalto University Committee on Research Ethics.

All MEG data were sampled at 1000 Hz, band-pass filtered to 1–45 Hz and thereafter downsampled to 125 Hz, as higher bandwidth and sampling rates did not significantly improve the performance while increasing computational time. We used the same basic preprocessing and scaling approaches for all four experiments. Each MEG epoch was scaled independently by subtracting the mean and dividing by the standard deviation of the MEG signal in all channels during the pre-stimulus interval (the −300 … 0-ms interval with the zero time corresponding to the stimulus onset). This approach provided reasonable scaling suitable for a real-time experiment while preserving the spatial structure of the signal.

The MEG dataset of Experiment 1 was preprocessed using a state-of-the-art off-line pipeline that is not applicable in a real-time BCI setting. External magnetic interference was suppressed and head movements compensated for using the temporally-extended signal-space separation (tSSS) method implemented in the MaxFilter software (version 2.2; MEGIN/Elekta Oy, Helsinki, Finland) (Taulu and Simola, 2006). Thereafter, cardiac and ocular artifacts were projected out using the FastICA algorithm as implemented in the MNE-Python software (Gramfort et al., 2013a). Preprocessing the datasets from Experiments 2 and 3 was performed exclusively using methods available for real-time processing. The rationale for different preprocessing approaches was to progress from an optimal MEG pipeline (Experiment 1) to a reduced one, comprising only those methods that are available in a real-time setting (Experiment 2), and finally to demonstrate a real BCI experiment (Experiment 3). Thus, no external magnetic interference or other artefact suppression was performed in Experiments 2 and 3, and the cardiac and oculomotor artifacts, typical for MEG measurements were present in these data. In Experiment 4 data from Cam-CAN dataset were pre-processed using tSSS, but no artifact correction was performed.

3.1. Experiment 1: classification of 5 types of sensory event-related fields

Dataset 1 comprised single-trial event-related-field (ERF) responses in 7 healthy human subjects (mean age = 30.1, 4 males, 3 females, 1 left-handed) to 5 types of sensory stimuli; checkerboard patterns presented in the right or left visual hemifield (Classes 1 and 2), 1-kHz 50-ms auditory tones presented to the left or right ear (pooled into Class 3), and transient transcutaneous electrical stimulation of the median nerve at the left or right wrist (Classes 4 and 5). This dataset comprised 500-ms segments of MEG measurements sampled at 1000 Hz (500 samples measured by 204 MEG channels) starting at the onset of each stimulus. Total trial counts per subject were 1622 ± 322 (mean ± SD).

3.2. Experiment 2: classification of event-related oscillatory activity in a 3-class motor imagery task

Dataset 2 comprised MEG measurements of 17 healthy human subjects who performed a motor imagery task (for details, see Halme and Parkkonen, 2018), in which they imagined moving either their left or right hand (without any actual movement) when a visual cue was presented. The data comprised 1500-ms segments of MEG measurements sampled at 1000 Hz (1500 samples measured by 204 MEG channels) starting at the onset of the visual cue. Based on the measured MEG signal, we decoded whether the subject imagined moving his/her left or right hand, or nothing at all (rest condition). Total trial counts per subject were 120 ± 5 (mean ± SD).

3.3. Experiment 3: real-time motor imagery BCI

Here we applied the same experimental paradigm as in Experiment 2, but with true real-time decoding and updating of the model. For this experiment, a network trained using a subset of trials from 17 subjects (Experiment 2) including only two classes (left vs. right motor imagery) was integrated into a real-time motor imagery BCI. Two subjects performed a task where they had to imagine moving the left or right hand following the presentation of a visual cue (an arrow pointing to the left or to the right). The VAR-CNN model which showed the highest performance in Experiment 2 performed 2-class classification (left vs. right hand motor imagery) in real time.

None of the subjects had used motor imagery-based BCIs before or

were involved in the collection of the MEG data used to train the model, i.e in Experiment 2. The experiment comprised three sessions. In the first session, the classifier was applied without the online updates to estimate the baseline performance. In sessions 2 and 3, the model parameters were updated following each trial where the classifier decoded the subject's intention correctly. The update was performed with a single back-propagation step using the MEG data from this trial and the associated label. Since our subjects had no prior experience with motor imagery tasks we chose to avoid updating the classifier after trials where the misclassification was due to the subject struggling to perform the mental task (as opposed to classifier errors). Updating the decoder with these erroneous trials would likely yield decreased decoding accuracy in subsequent trials. For the same reason we constrained this experiment to have only two classes. Trials in each session were presented in three different pre-defined sequences (50 trials per session) such that the true labels were available for real-time accuracy estimation and incremental model optimization.

3.4. Experiment 4: classification of 2 types of sensory stimuli on an Cam-CAN dataset

We used MEG data from 250 healthy subjects chosen (subjects: 110033–321594) from the open Cam-CAN data repository (Taylor et al., 2017). The subjects performed a (passive) audio–visual task comprising 120 trials of unimodal stimuli (60 visual stimuli: bilateral/full-field circular checkerboards; 60 auditory stimuli: binaural tones at one of three equiprobable frequencies), presented at a rate of approximately 1 per second (for details, see Taylor et al., 2017). Similarly to Experiment 1, planar gradiometer data were preprocessed with tSSS, scaled, and split into 500-ms segments, starting at the onset of the (visual or auditory) stimuli. No artefact suppression was performed. Due to a large number of subjects, the model evaluation procedure differed from the one used in Experiments 1 and 2. The training (90%) and validation (10%) sets comprised pooled MEG data from 200 subjects (2 classes of stimuli, 60 trials per class). Each model was trained only once and then applied to each of the 50 held-out subjects of the test set independently. *Initial test accuracy* and *pseudo-real-time accuracy* were estimated on these held-out subjects in the same way as in Experiments 1 and 2.

4. Results

4.1. Experiment 1

In this 5-class decoding of sensory ERFs, VAR-CNN outperformed other models in terms of accuracy on a pooled validation set (95.8%± 0.7%), when applied to held-out subjects (85.9% ± 7.4%), and in pseudo-real time tests (94.2% ± 2.8%). The results from LF-CNN were the second best, and RBF-SVM was the closest benchmark. Detailed statistical comparisons between the performance of LF-CNN, VAR-CNN and the benchmark classifiers in Experiment 1 are summarized in Supplementary Table 2.

In the simulated real-time experiments, the CNN-based models EEGNet-8, LF-CNN and VAR-CNN significantly improved their performance (+ 12.4, + 10.2, and +8.3 percent points, respectively compared to the initial test accuracies. In case of SVM classifiers, this improvement was considerably smaller (+6.8 for Linear SVM, and +1.2 percent points for RBF-SVM). These results are summarized in Table 2.

Table 2
Across-subject performance in a 5-class sensory stimulation task. Grand-average accuracy scores (mean ± SD) from leave-one-subject-out cross-validation. Highest-performing model in each test is indicated in bold.

Model	Validation (%)	Initial test (%)	Pseudo-real-time (%)
LF-CNN	95.0 ± 0.8	83.1 ± 8.4	93.3 ± 3.6
VAR-CNN	**95.8 ± 0.7**	**85.9 ± 7.4**	**94.2 ± 2.8**
Linear SVM	93.3 ± 1.2	80.2 ± 9.7	87.0 ± 5.4
RBF-SVM	93.6 ± 1.7	82.7 ± 8.3	83.9 ± 8.4
ShallowFBCSP-CNN	85.3 ± 2.4	60.1 ± 11.7	n.a.
EEGNet-8	88.7 ± 2.0	76.8 ± 11.7	89.2 ± 5.0
VGG19	80.5 ± 3.3	70.1 ± 12.8	73.9 ± 10.5

Table 3
Across-subject performance in a 3-class motor imagery task. Grand-average accuracy scores (mean ± SD) from leave-one-subject-out cross-validation. Highest-performing model in each test is indicated in bold.

Model	Validation (%)	Initial test (%)	Pseudo-real-time (%)
LF-CNN	84.3 ± 2.7	74.2 ± 6.5	78.0 ± 6.5
VAR-CNN	**86.7 ± 7.4**	**76.3 ± 6.8**	**82.3 ± 6.1**
Linear SVM	76.9 ± 3.0	68.2 ± 7.2	71.4 ± 7.3
RBF-SVM	80.3 ± 2.6	74.1 ± 8.4	73.6 ± 8,8
ShallowFBCSP-CNN	70.2 ± 4.1	60.2 ± 10.3	n.a.
EEGNet-8	80.8 ± 2.4	72.1 ± 5.8	80.9 ± 6.7
VGG19	71.4 ± 9.6	60.2 ± 6.8	57.4 ± 9.2

4.2. Experiment 2

With the 3-class motor imagery dataset, VAR-CNN again gave the highest classification accuracies when applied to pooled validation set (86.7% ± 7.4%), held-out subjects (76.3% ± 6.8%) and pseudo-real-time tests (82.3% ± 6.1%). The closest benchmarks classifiers were RBF-SVM for held-out subjects (74.1% ± 8.4%), and EEGnet-8 for the validation set (80.8% ± 2.4%) and pseudo-real-time test (80.9% ± 6.7%). Detailed statistical comparisons between the performance of LF-CNN, VAR-CNN and the benchmark classifiers in Experiment 2 are summarized in Supplementary Table 3. Similarly to Experiment 1, EEGNet-8, LF-CNN and VAR-CNN were able to significantly improve their performance in a simulated real-time test using incremental updates (+ 8.9, + 3.8, and + 5.7 percent points, respectively) compared to the initial test accuracies. The linear SVM classifier performance improved to a smaller extent through pseudo-real-time updates (+3.2 percent points), while the performance of RBF-SVM did not improve at all (−0.5 percent points). These results are summarized in Table 3.

4.3. Experiment 3

Results of the real-time application of VAR-CNN are summarized in Table 4. Comparing the accuracies achieved by VAR-CNN with and without stochastic updates clearly shows the significant increase in performance. We note that since the subjects had no prior experience with motor imagery BCIs, the improvement in performance after Session 1 may be partly attributed to their improved motor imagery skills.

4.4. Experiment 4

Table 5 summarizes the results obtained when classifying auditory *vs* visual stimuli in a passive audio–visual task from the Cam-CAN dataset (Taylor et al., 2017). Similarly to Experiments 1 and 2, LF-CNN and VAR-CNN achieved roughly equal accuracies, outperforming benchmark models on the validation set (LF-CNN: 94.9% and VAR-CNN: 95.6%) as well as on the held-out subjects with (96.5% and 96.0%, respectively) and without online updates (95.1% and 95.8%). Among the benchmark models, RBF-SVM and EEGNet-8 achieved the highest performance. Results for ShallowFBCSP-CNN are not available for this dataset, because

Table 4
VAR-CNN classification accuracy in the real-time motor imagery BCI experiment.

Subject	Run 1, no updates (%)	Run 2, online updates (%)	Run 3, online updates (%)
s01	80.0	88.0	92.0
s02	62.0	90.0	82.0

Table 5
Across-subject performance on 2-class Cam-CAN dataset. Grand-average accuracy scores (mean ± SD) estimated on a test set comprising 50 held-out subjects. Highest-performing model in each test is indicated in bold.

Model	Validation (%)	Initial test (%)	Pseudo-real-time (%)
LF-CNN	94.9	95.1 ± 4.2	**96.5 ± 2.8**
VAR-CNN	**95.6**	**95.8 ± 4.0**	96.0 ± 2.8
Linear SVM	91.6	92.1 ± 5.5	92.7 ± 7.6
RBF-SVM	93.1	93.9 ± 5.7	94.4 ± 7.3
ShallowFBCSP-CNN	n.a.	n.a.	n.a.
EEGNet-8	93.8	93.0 ± 4.3	94.3 ± 3.9
VGG19	94.7	92.3 ± 5.0	93.2 ± 5.2

the authors were not able to adapt its available implementation to the increased memory requirements.

4.5. Interpretation of learned model parameters

Fig. 3 shows the activation patterns and the corresponding informative time-windows of the components with the maximum contribution to the decoding of each class in the LF-CNN model, trained on the pooled data from Experiment 1 and updated using the pseudo-real-time update procedure described above on single held-out subject. For all of the five classes, our model extracted spatial patterns whose source estimates showed overall good correspondence to the locations and lateralizations of the peaks of the evoked responses (Fig. 3).

Similarly, activation patterns extracted from the data of Experiment 2 following a similar training and updating procedure resulted in spatial patterns focused over parietal as well as the more posterior occipital sensors contralateral to the imagined movement. Interestingly, spectral estimates of the most informative latent components resulted in density estimates peaking at around 10 Hz corresponding to well-known μ-rhythm desynchronization associated with motor imagery (Fig. 5). Although anatomical information was not available for this dataset these results suggest that apart from motor and pre-motor cortices more posterior sources might also contribute to motor imagery.

We also estimated spatial properties of the latent components that had the least overall contribution to any of the classes. Inspecting the weights of the output (classification) layer (Fig. 4) further allowed us to identify 5 components which provided a minimum contribution to either class. Although none of these components directly corresponded to known signatures of e.g. oculomotor artifacts (due to applied artefact suppression), their overall limited contribution to either class may suggest that these patterns were used for projecting the irrelevant activity out.

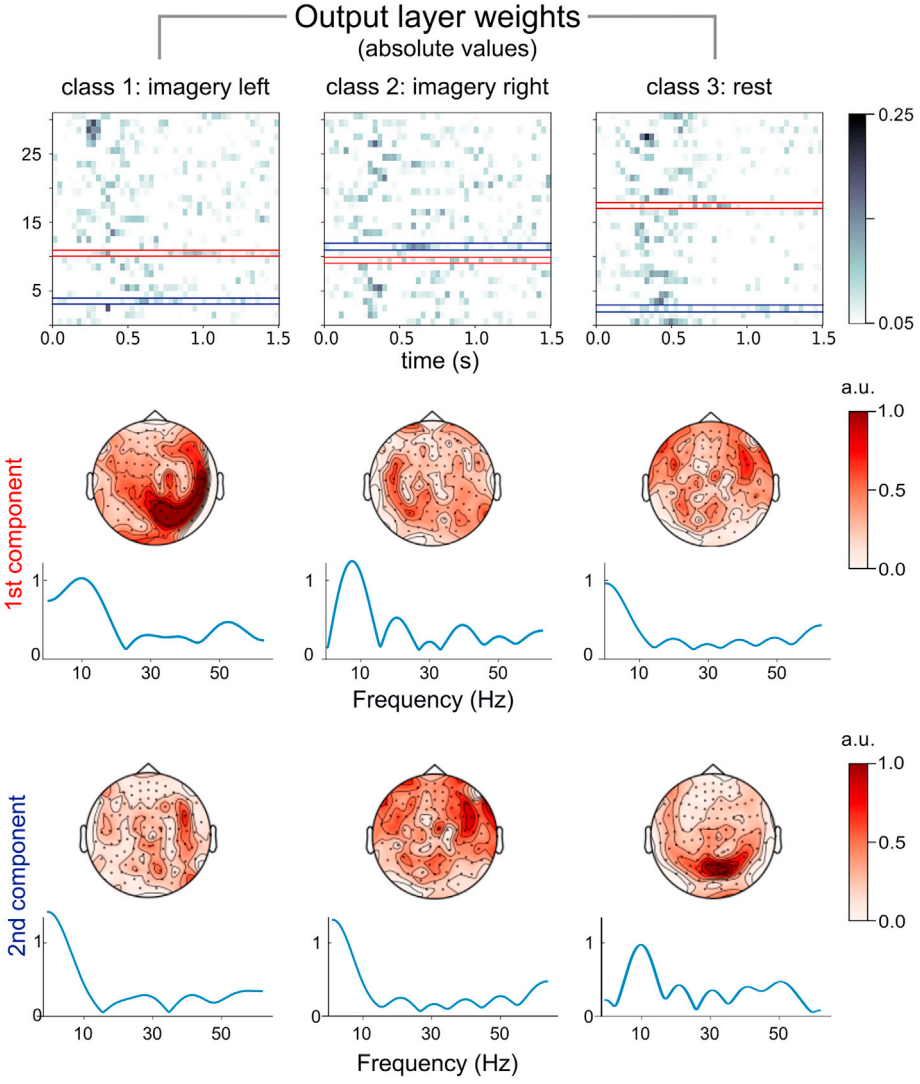

Fig. 5. Interpretation of informative LF-CNN model parameters in a representative subject from Experiment 2. Latent components having the maximum positive (red) and negative (blue) sum of weights over all time points for each class were extracted from the model and interpreted in terms of their spatial topographies, and spectral estimates.

5. Discussion

In this paper, we report two neural network models optimized for across-subject classification of electromagnetic brain signals. These models outperform traditional approaches as well as more complex deep neural networks in terms of accuracy, across-subject generalization, and simulated real-time performance. Arising from inter-individual variability in structural and functional cortical anatomy, across-subject generalization has proved to be a challenging problem for machine-learning methods applied to EEG and particularly to MEG data due to its higher spatial resolution. To illustrate the severity of this problem, we estimated the correlations between the spatial patterns extracted from our model fine-tuned to individual subjects. Mean spatial correlation coefficient between the most consistent components was $r = 0.67$ in Experiment 1 ($N = 7$) and $r = 0.52$ in Experiment 2 ($N = 17$). One advantage of our models is that they introduce reasonable assumptions based on the signal generation model. These assumptions include e.g. linear separability in the spatial domain, consistency of temporal or spectral properties of the signal across trials etc. Thus, we would expect it to be most sensitive to features common across subjects while allowing for reasonable variability.

We further show that the relatively low complexity of our LF-CNN architecture allows interpretation of the learned model parameters in terms of the underlying neural activity (Fig. 2). Such interpretation can prove to be a convenient tool for quickly exploring complex, high-dimensional MEG and EEG datasets and ultimately allow extracting more information from these rich data. Interpretation of discriminative models can provide valuable insights into these data by e.g. allowing to dissociate "most active" sources from those that contribute the most to discrimination. One potential application of these methods could be investigating neural sources contributing to BCI control. In the case of VAR-CNN, however, such interpretation is not possible, because this model allows complex interactions between the latent components, destroying one-to-one correspondence between spatial and temporal features.

Moreover, we propose a procedure where the classifier is initialized from the data of other subjects and then updated online during the real-time experiment. We demonstrate that using this approach our models perform accurately on new subjects in a real-time BCI experiment, allowing new subjects to efficiently use the system without separate calibration. Importantly, this procedure allows to omit a dedicated BCI calibration session, facilitating the use of BCIs in research and clinical settings.

The results reported here show considerable improvement in performance compared to our previous results in the across-subject decoding of MEG. Using state-of-the-art feature extraction methods in combination with linear classifiers Halme and Parkkonen (2018) achieved a classification accuracy of 70.6%. Our best-performing model was able to classify 3 classes of stimuli on the same dataset with 76.3% accuracy. Furthermore, adding incremental updates to the classifier allowed us to achieve even higher accuracy (82.3%) in a simulated real-time test. We confirmed that the feasibility of the latter approach in a real-time BCI experiment, achieving similarly high accuracy in Experiment 3.

Several studies report applying deep neural networks to single-trial classification of non-invasive neurophysiological measurements, typically multichannel EEG data. Most successful models make use of the spatiotemporal structure of EEG (Bashivan et al., 2015; Hajinoroozi et al., 2016; Lawhern et al., 2018; Schirrmeister et al., 2017) including the real-time applications (Burget et al., 2017; Fahimi et al., 2019). Bashivan et al. (2015) exploit the spatiotemporal and spectral structure of EEG by transforming the signals into a sequence of multidimensional spatio–spectral images using time–frequency and polar transforms to achieve significant performance improvement in classifying cognitive load. By contrast, we use a simpler linear projection in the spatial domain. Similarly to our approach, Hajinoroozi et al. (2016) tested spatial ICA as a preprocessing step followed by temporal 1-d convolution filters. However, such a combination did not result in a significant improvement in model performance in that study, as compared to using EEG channel data as the input. We argue that – when performed separately from the classification – ICA decomposition may not be optimal due to the independence assumption which may not hold for real EEG and MEG signals. In our design, the analogous linear projection in the spatial domain was obtained by back-propagation. When trained in conjunction with the frequency filters in the temporal convolution layer, such projection results in a separable decomposition related to a combination of Linear Discriminant Analysis (McLachlan, 1992), and Spatio–Spectral Decomposition (Nikulin et al., 2011) used in EEG–MEG analysis.

Lawhern et al. (2018) introduced the EEGNet model as a compact, interpretable CNN architecture that can be applied to different EEG-based BCI paradigms (including both evoked and induced responses). EEGNet has also been shown to generalize well to held-out subjects. Importantly, in our simulated real-time tests, EEGNet came close to the accuracies of our models, demonstrating the advantages of neural networks in combination with stochastic optimization.

Similarly to other models (e.g. Schirrmeister et al., 2017), Lawhern et al. (2018) applied 1-d temporal convolutions to the raw EEG channel data. Our results also indicate that models using 1-d convolutions are better suited for classification of MEG data than the deeper image classification networks relying on 2-d convolutions. In contrast to previous studies, we apply spatial decomposition first, followed by a temporal depth-wise (LF-CNN) or spatiotemporal (VAR-CNN) convolution. The motivation for this is three-fold: first, it allows effective spatial decorrelation and dimensionality reduction; second, when nested into the CNN architecture, it allows the network to learn and project out physiological artifacts similarly to other linear projection methods; finally, it can contribute to improved generalization across subjects by increasing model robustness to inter-individual differences in spatial distributions of the informative features. The latter argument is particularly important for MEG, since its spatial resolution is considerably higher than that of EEG, and even minor differences in source location or orientation may change the signal patterns at the sensors and lead to different channels being most sensitive to the brain activity of interest. This is the likely reason why the CNN models optimized for EEG decoding did not perform optimally in our study.

Our results demonstrate that with an adequate choice of hyperparameters, support vector machines may perform equally well as the best-performing CNN models in off-line classification. We thus recommend using SVMs as benchmark models in future studies. Our simulated real-time tests, however, demonstrate that incremental versions of SVMs do not gain as much in performance from the real-time updates as the CNN-based models do. This result probably reflects the fact that incremental SVMs rely on kernel approximations and thus their fine-tuning to individual subjects is limited.

Thus, the proposed models outperformed the benchmark methods and provided means to investigate spatial and temporal patterns that contribute to discrimination between the stimuli. Inspecting these patterns may prove a useful tool to obtain a fast preliminary estimate of the neural activity informing the classification. Yet, the degree of correspondence between the patterns informing the classification and the actual neural activity is not clear and requires further systematic testing.

In this study, we restrict the interpretation of the model parameters to the single most-informative latent component. However, it is expected that several distinct components can contribute to a single class and provide valuable information (as e.g. shown in Fig. 5). We suggest two heuristic approaches to identify such informative components. Similarly, we limit ourselves to the interpretation of the simpler variant (LF-CNN) of the proposed networks due to its low complexity and predominantly linear activation functions. A more comprehensive model investigation would require a more systematic approach, suggested e.g. in Kindermans et al. (2018) and Alber et al. (2018).

Finally, we show that the proposed models outperform other neural networks designed for EEG classification when applied to MEG data.

Given a large number of channels and a higher spatial resolution MEG datasets may benefit from these designs. It remains to be seen, however, how the proposed models can perform when applied to EEG data.

The proposed models can be efficiently used to optimize the existing and to develop novel BCI paradigms. Moreover, inspecting the parameters of the LF-CNN model variant can be used for quick explorative analysis of new experimental paradigms in cognitive neuroscience. Finally, the proposed architectures can be used as building blocks of the more complex models. For example, the proposed models can be used for extracting MEG signal features for information fusion across measurement modalities, or for more sophisticated experimental designs (e.g. single-trial regression problems and inter-subject similarity measures).

6. Conclusions

We have introduced a Convolutional Neural Network model optimized for off- and online (real-time) classification of MEG data. Incorporating prior knowledge about the processes generating MEG observations allowed us to substantially reduce model complexity while preserving high accuracy and interpretability. We show that this model successfully classifies evoked as well as oscillatory activity and generalizes efficiently across subjects. When combined with incremental real-time model updates, the time-consuming calibration sessions in MEG-based brain–computer interfaces could be omitted provided that a sufficient amount of training data from other subjects is available.

Acknowledgements

Research reported here was supported by the Academy of Finland, Grant/Award Number: “NeuroFeed”/295075 and the European Research Council under ERC Grant Agreement n. 678578. The content is solely the responsibility of the authors and does not necessarily represent the official views of the funding organizations. The calculations presented above were performed using computer resources within the Aalto University School of Science “Science-IT” project.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.neuroimage.2019.04.068.

References

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G.S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., Levenberg, J., Mane, D., Monga, R., Moore, S., Murray, D., Olah, C., Schuster, M., Shlens, J., Steiner, B., Sutskever, I., Talwar, K., Tucker, P., Vanhoucke, V., Vasudevan, V., Viegas, F., Vinyals, O., Warden, P., Wattenberg, M., Wicke, M., Yu, Y., Zheng, X., 2016. TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems, vol. 3. http://arxiv.org/abs/1603.04467.

Alber, M., Lapuschkin, S., Seegerer, P., Hägele, M., Schütt, K.T., Montavon, G., Samek, W., Müller, K.-R., Dähne, S., Kindermans, P.-J., 2018. iNNvestigate Neural Networks!. https://github.com/albermax/innvestigate.

Baillet, S., 2017. Magnetoencephalography for brain electrophysiology and imaging, 3 Nat. Neurosci. 20 (3), 327–339. https://doi.org/10.1038/nn.4504. ISSN 1097-6256. http://www.nature.com/articles/nn.4504.

Bashivan, P., Rish, I., Yeasin, M., Codella, N., 2015. Learning Representations from EEG with Deep Recurrent-Convolutional Neural Networks. https://doi.org/10.1080/03610928808829796. ISSN 03610926.

Burget, F., Fiederer, L.D.J., Kuhner, D., Völker, M., Aldinger, J., Schirrmeister, R.T., Do, C., Boedecker, J., Nebel, B., Ball, T., Burgard, W., 2017. Acting Thoughts: towards a Mobile Robotic Service Assistant for Users with Limited Communication Skills, vol. 7. https://doi.org/10.1109/ECMR.2017.8098658. http://arxiv.org/abs/1707.06633Ωhttps://doi.org/10.1109/ECMR.2017.8098658.

Cardoso, J.-F., 1998. Blind signal separation: statistical principles. Proc. IEEE 86 (10), 2009–2025. https://doi.org/10.1109/5.720250. ISSN 00189219. http://ieeexplore.ieee.org/document/720250/.

Dale, A.M., Liu, A.K., Fischl, B.R., Buckner, R.L., Belliveau, J.W., Lewine, J.D., Halgren, E., 2000. Dynamic statistical parametric mapping: combining fMRI and MEG for high-resolution imaging of cortical activity, 4 Neuron 26 (1), 55–67. https://doi.org/10.1016/S0896-6273(00)81138-1. ISSN 0896-6273. http://www.ncbi.nlm.nih.gov/pubmed/10798392.

Daunizeau, J., Friston, K.J., 2007. A mesostate-space model for EEG and MEG. Neuroimage 38 (1), 67–81. https://doi.org/10.1016/j.neuroimage.2007.06.034. ISSN 10538119.

Fahimi, F., Zhang, Z., Goh, W.B., Lee, T.-S., Ang, K.K., Guan, C., 2019. Inter-subject transfer learning with an end-to-end deep convolutional neural network for EEG-based BCI, 4 J. Neural Eng. 16 (2), 026007. https://doi.org/10.1088/1741-2552/aaf3f6. ISSN 1741-2560. http://stacks.iop.org/1741-2552/16/i=2/a=026007?key=crossref.7c244fde88bbf289575afd594f61eb05.

Gramfort, A., Luessi, M., Larson, E., Engemann, D.A., Strohmeier, D., Brodbeck, C., Goj, R., Jas, M., Brooks, T., Parkkonen, L., Hämäläinen, M., 2013a. MEG and EEG data analysis with MNE-Python, 12 Front. Neurosci. 7 (7 DEC), 267. https://doi.org/10.3389/fnins.2013.00267. ISSN 1662453X.

Gramfort, A., Luessi, M., Larson, E., Engemann, D.A., Strohmeier, D., Brodbeck, C., Goj, R., Jas, M., Brooks, T., Parkkonen, L., Hämäläinen, M., 2013b. MEG and EEG data analysis with MNE-Python, 12 Front. Neurosci. 7, 267. https://doi.org/10.3389/fnins.2013.00267. ISSN 1662-4548.

Hajinoroozi, M., Mao, Z., Jung, T.-P., Lin, C.-T., Huang, Y., 2016. EEG-based prediction of driver's cognitive performance by deep convolutional neural network, 9 Signal Process. Image Commun. 47, 549–555. https://doi.org/10.1016/J.IMAGE.2016.05.018. ISSN 0923-5965. https://www.sciencedirect.com/science/article/pii/S0923596516300832?via%3Dihub.

Halme, H.-L., Parkkonen, L., 2018. Across-subject offline decoding of motor imagery from MEG and EEG, 7 Sci. Rep. 8 (1), 10087. https://doi.org/10.1038/s41598-018-28295-z. https://www.nature.com/articles/s41598-018-28295-z.

Hamalainen, M., Hari, R., Ilmoniemi, R.J., Knuutila, J., Lounasmaa, O.V., 1993. Magnetoencephalography Theory, Instrumentation, and Applications to Noninvasive Studies of the Working Human Brain. https://doi.org/10.1103/revmodphys.65.413. http://journals.aps.org/rmp/abstract/10.1103/RevModPhys.65.413.

Haufe, S., Meinecke, F., Görgen, K., Dähne, S., Haynes, J.-D., Blankertz, B., Bießmann, F., 2014. On the interpretation of weight vectors of linear models in multivariate neuroimaging. Neuroimage 87, 96–110. https://doi.org/10.1016/j.neuroimage.2013.10.067. ISSN 10538119.

He, K., Zhang, X., Ren, S., Sun, J., 2015. Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification, vol. 2. http://arxiv.org/abs/1502.01852.

Kindermans, P.-J., Brain, G., SchüttSch, K.T., Alber Berlin, M.T., Berlin, T., Dähne, S., 2018. Learning How to Explain Neural Networks: PaternNet and PaternAttribution. https://arxiv.org/pdf/1705.05598.pdf.

Lawhern, V.J., Solon, A.J., Waytowich, N.R., Gordon, S.M., Hung, C.P., Lance, B.J., 2018. EEGNet: A compact convolutional neural network for EEG-based brain–computer interfaces, 10 J. Neural Eng. 15 (5), 056013. https://doi.org/10.1088/1741-2552/aace8c. ISSN 1741-2560. http://stacks.iop.org/1741-2552/15/i=5/a=056013?key=crossref.926d8be8d11477a9c1ce0a164d5c869a.

McLachlan, G.J., 1992. Discriminant Analysis and Statistical Pattern Recognition Wiley Series in Probability and Statistics, vol. 3. John Wiley & Sons, Inc., Hoboken, NJ, USA https://doi.org/10.1002/0471725293. ISBN 9780471725299.

Nikulin, V.V., Nolte, G., Curio, G., 2011. A novel method for reliable and fast extraction of neuronal EEG/MEG oscillations on the basis of spatio-spectral decomposition. Neuroimage 55 (4), 1528–1535. https://doi.org/10.1016/j.neuroimage.2011.01.057. ISSN 1095-9572. https://doi.org/10.1016/j.neuroimage.2011.01.057Ωhttp://www.ncbi.nlm.nih.gov/pubmed/21276858.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E., 2011. Scikit-learn: machine learning in Python. J. Mach. Learn. Res. 12 (Oct), 2825–2830. http://jmlr.csail.mit.edu/papers/v12/pedregosa11a.html.

Schirrmeister, R.T., Springenberg, J.T., Fiederer, L.D.J., Glasstetter, M., Eggensperger, K., Tangermann, M., Hutter, F., Burgard, W., Ball, T., 2017. Deep learning with convolutional neural networks for EEG decoding and visualization, 11 Hum. Brain Mapp. 38 (11), 5391–5420. https://doi.org/10.1002/hbm.23730. ISSN 10659471. http://doi.wiley.com/10.1002/hbm.23730.

Simonyan, K., Zisserman, A., 2015. Very Deep Convolutional Networks for Large-Scale Image Recognition. Technical Report. http://www.robots.ox.ac.uk/.

Taulu, S., Simola, J., 2006. Spatiotemporal signal space separation method for rejecting nearby interference in MEG measurements. Phys Med Biol Phys. Med. Biol. 51 (51), 1759–1768. https://doi.org/10.1088/0031-9155/51/7/008. http://www.iop.org/journals/pmbΩhttp://stacks.iop.org/pmb/51/1759.

Taylor, J.R., Williams, N., Cusack, R., Auer, T., Shafto, M.A., Dixon, M., Tyler, L.K., Cam-CAN, Henson, R.N., 2017. The Cambridge Centre for Ageing and Neuroscience (Cam-CAN) data repository: structural and functional MRI, MEG, and cognitive data from a cross-sectional adult lifespan sample, 1 Neuroimage 144, 262–269. https://doi.org/10.1016/J.NEUROIMAGE.2015.09.018. ISSN 1053-8119. https://www.sciencedirect.com/science/article/pii/S1053811915008150?via%3Dihub.

Vapnik, V.N., 2000. The Nature of Statistical Learning Theory. Springer. ISBN 9781441931603.

Westner, B.U., Dalal, S.S., Hanslmayr, S., Staudigl, T., 2018. Across-subjects classification of stimulus modality from human MEG high frequency activity, 3 PLoS Comput. Biol. 14 (3), e1005938. https://doi.org/10.1371/journal.pcbi.1005938. ISSN 15537358.