Lights in the Brain

Pre-training neural signal prediction for downstream single-subject brain-computer interface thought classification with transfer learning

Esben Kran*[1]
[1] Stud. BSc. Cognitive Science, Aarhus University, 201909190
* Correspondence: Esben Kran, supervisor: Arthur Hjorth

Brain-computer interfaces (BCI) become useful for a wide array of everyday tasks but to accomplish this, we need an easy way to read the brain activity with high accuracy. Existing research in BCIs focus on single task classification accuracy and miss the broader view of multi-task accuracy with single-subject calibration to increase accuracy and usability. In this paper, one subject was recorded in an unstructured paradigm for 3.5 hours using the functional near-infrared spectroscopy (fNIRS) neuroimaging method that measures oxygen flow in the brain. A self-supervised bidirectional LSTM, a unidirectional LSTM and a fully-connected model were trained to predict the single-channel brain signal 4 seconds into the future given 10 seconds of multi-channel data and achieved absolute error far better than a mean prediction. Afterwards, the model weights from this brain signal prediction task were transferred and used for brain signal classification tasks by augmenting with a nonlinear layer and a sigmoid output. The signals to classify spanned three levels of intrinsicity, i.e. how much the activity stems from within the brain, from waving the arms and talking to subtracting numbers and imagining rotating a box. Each of these were low-resource recordings with 79-83 trials in each that should benefit the most from transferring weights from pre-trained models. All models overfit severely to the training dataset despite many measures to reduce this, e.g. data augmentation, normalisation, and dropout. However, when transferring layer weights from pre-trained models, a reduction in overfitting was seen. Investigations into how this reduction was affected by how many layers were transferred showed that early layers had learned different patterns than the non-transferred models, in accordance with earlier work. Over- and underfitting were plausibly caused by data limitations but the framework presented in this paper presents a bold vision for what the future might hold for BCIs and takes the first steps that future studies can build upon. See reproducible code, figures and data on https://github.com/esbenkc/fnirs-bci.

Keywords: fNIRS, BCI, neuroscience, NL

Introduction

Brain-computer interfaces (BCI) provide a new framework for studying and augmenting the human mind and have been successful in offering new perspectives for what human cognition is capable of. Implanted BCI enable patients to control robot arms and write using just their thoughts, however, BCIs on the head (non-invasive BCIs) are limited in how easily and reliably they can be implemented in everyday use because of algorithm limitations. With recent revolutions in artificial language processing, we can imagine novel solutions to these algorithm problems. I attempt to combine these fields and present the practical implementation of solutions to make the use of neuroimaging devices for everyday use better.

Specifically, I perform self-supervised training (LeCun & Misra, 2021) to pre-train a machine learning model using the LSTM architecture (Hochreiter & Schmidhuber, 1997) on functional near-infrared spectroscopic (fNIRS) neuroimaging data (Naseer & Hong, 2015) from the NIRx NIRSport2 system (NIRx, 2021) and transfer and fine-tune it for a BCI thought classification task (Yoo et al., 2018) as is done with language models (C. Sun et al., 2019). As far as I am aware, this is the first example of such work.

Brain-computer interfacing (BCI)

BCIs expand the ways the human brain can interact with the world (Eagleman, 2020), e.g. let tetraplegics (unable to use limbs) write digitally (Willett et al., 2021) and have ALS patients control a robotic arm (Aliakbaryhosseinabadi et al., 2021). BCIs include an algorithm that takes neural signals as input, pre-processes that input and generates an output from its internal algorithm, be that machine learning or a handmade algorithm (Hill & Wolpaw, 2016; Wolpaw et al., 2002). The popularity of the field is rising (Hill & Wolpaw, 2016) and with new technologies (Audette, 2013; Hochberg et al., 2012; Johnson & Kernel, 2020; Musk & Neuralink, 2019), we move further towards clinical and public usage of devices that can and will enable everyday use of these types of algorithms.

However, an important step on this journey is missing. Invasive technologies are as quick as our own limbs to adjust to neural activity (Buzsaki, 2019; Eagleman, 2020) but non-invasive neuroimaging technologies for the public do not have the same ease of use (Gürkök et al., 2011) with problems regarding brain data including, but not limited to, signal instability and low signal-to-noise ratio (Gagnon et al., 2012; Klein & Kranczioch, 2019) and a long adaptation time (Willett et al., 2021). These are especially true for the so-called *active* BCIs that read thoughts directly compared to *reactive* BCIs (Hill & Wolpaw, 2016) that classify brain states based on e.g. specific sounds in the environment. The present paper will contribute to solving these problems with solutions that take inspiration from the field of natural-language processing (NLP).

Pre-trained language models

In NLP, Pre-trained language models are large neural networks of different architectures that are trained on massive amounts of data (Devlin et al., 2019). These models gain a basic understanding of language and can be augmented to solve more specific tasks. The most-used of these are sometimes called foundation models (Bommasani et al., 2021).

An example of a pre-trained language model is BERT (Devlin et al., 2019) where a specially-designed transformer architecture is trained on 3.3 million words. This model is trained through self-supervision by generating training data from the sentences in the text and masking out words. This removes the need for human labelling. The model is then asked to predict the words in these masked locations. Other examples include the GPT series of models that attempt to predict the next word in a sequence given the previous words (Radford et al., 2018). They call this process an unsupervised pre-training, a semi-supervised training process or self-supervised training (LeCun & Misra, 2021).

The big language models of GPT-3, BERT, T5 and ELMo (Simon, 2021) learn semantically relevant encodings of text and can use this knowledge in similar ways as humans by generalising the high-level semantic understanding they get to specific tasks such as reading comprehension (Lai et al., 2017), answering questions (Z. Chen et al., 2018) and emotion analysis (Socher et al., 2013). By disabling training in the pre-trained model layers and adding extra trainable layers that fit the task, we avoid spending valuable and climate-damaging (Bommasani et al., 2021) compute resources every time we need to model a new task. Often, the added layers are a simple dense neural network and an output layer, e.g. a Softmax layer for text classification (Radford et al., 2018).

By transferring these practises to neuroimaging, the hypothesis is that we are able to perform semi-supervised pre-training of a model that learns parts of brain activity dynamics to encode different brain states into embeddings useful for fine-tuning and transfer learning. Existing work has looked into

similar transfer learning possibilities with electrocorticographic (ECoG) subcranial recordings (Elango et al., 2018; Makin et al., 2020) and implanted invasive microelectrode arrays (Schwemmer et al., 2018) with relatively good success. No similar work was found in the field of fNIRS, the neuroimaging technique with most potential in everyday BCI systems.

Functional near-infrared spectroscopy (fNIRS) for brain-computer interfacing

fNIRS records blood oxygen levels which is a functional measure of neuron activity, delayed by a few seconds. It fits on the head and only needs a clip-on device for wireless, live data transmission. fNIRS can only record neuron activity close to the scalp but it is the centre of high-level tasks such as speaking, understanding, thinking, hearing, and seeing. With fNIRS, the user is able to talk and move naturalistically, it is cheaper and less noisy compared to other non-implanted (non-invasive) neuroimaging technologies, and it is more spatially precise in its measurements (Solovey et al., 2009).

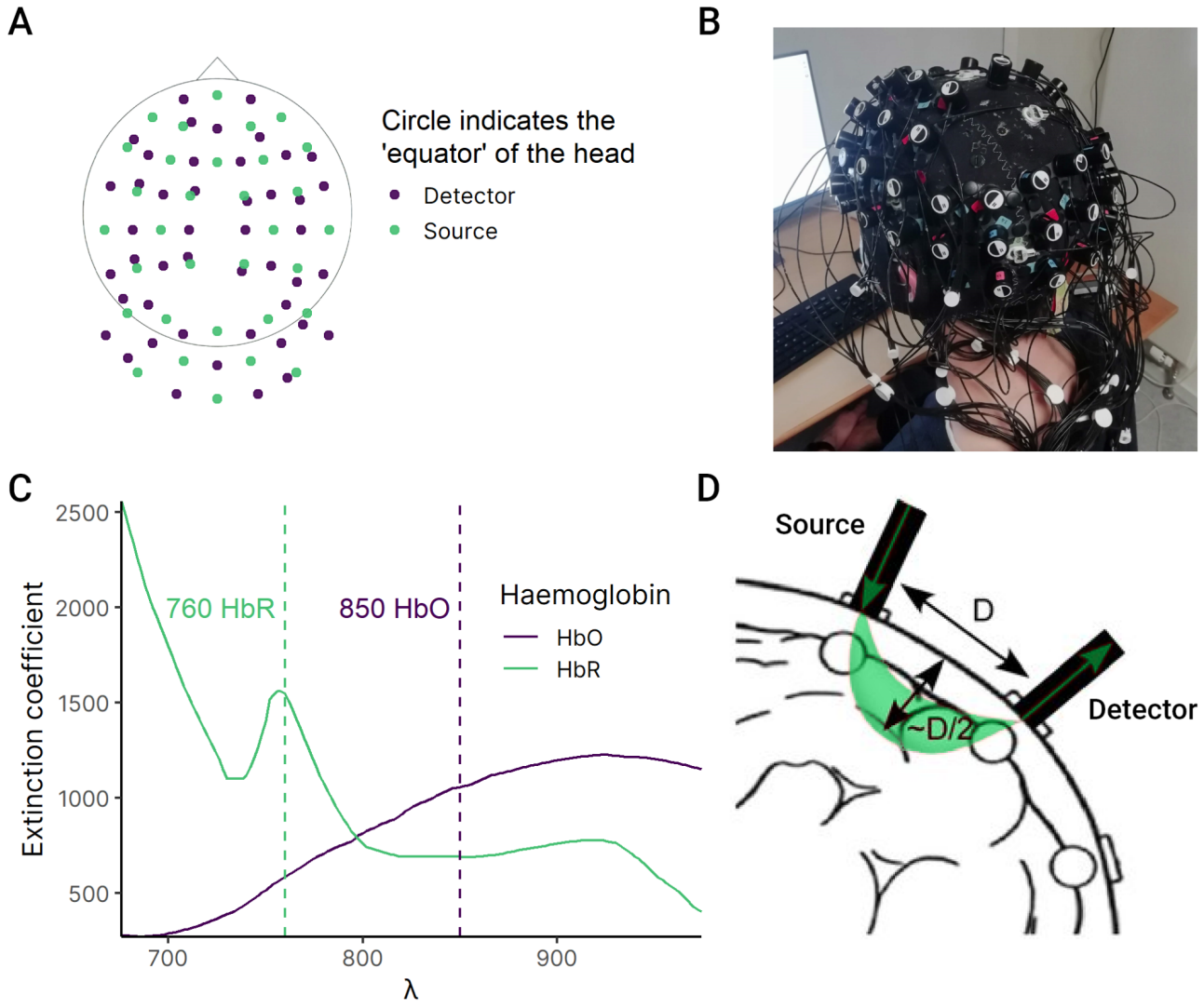

*Figure 1: A and B) The setup of sources and detectors used in this work's data, B) visualisation of source-detector pair photon migration (adapted from **Chen, 2016**), C) extinction coefficients for HbO and HbR with lines for wavelengths 760 nm and 850 nm that are used to infer HbR and HbO concentration*

E.g., electroencephalography (EEG) records highly diffuse neuron electricity levels from the scalp which will fluctuate wildly if speaking and moving (Repovs, 2010). Magnetoencephalography (MEG) records magnetic field fluctuations from neuron activity but needs a magnetically sealed room that can cost thousands per participant (Singh, 2014). Functional magnetic resonance imaging (fMRI) measures the same as fNIRS at a higher precision but the user has to lie still and it is a very noisy and expensive machine (Block Imaging, 2021).

The limitations of fNIRS come from the fact that neuron activity occurs 2~5 seconds before the blood-oxygen-level-dependent (BOLD) signal gives an indication of this (Arthurs & Boniface, 2002). For BCI, fNIRS is often combined with EEG (Naseer & Hong, 2015) which, despite its high noise, has millisecond timing of when neuron clusters activate (Subha et al., 2010).

fNIRS utilises the way oxy- and deoxygenated blood (oxy- and deoxyhaemoglobin, HbO and HbR) absorbs infrared light, and as oxygen is used by ATP to generate energy for neuronal activity, the BOLD signal measures neuron activity. An optode occupies positions in a 10-20 cap (Homan et al., 1987, p. 198), and is either a source or a detector (fig. 1A, 1B). The source is composed of two lights that each send a continuous stream of photons into the skull at wavelengths of around 760 nm and 850 nm respectively. These infrared wavelengths have the advantage of passing through most tissue, including the skull, but diffracting and absorbing against HbO and HbR. Since HbO and HbR have different absorption spectra that intersect at ~800 nm, the two lights will be a measure of HbR and HbO, respectively (fig. 1C). The measure of HbO and HbR each come from the amount of photons hitting the detectors compared to the source output of photons. This way, how much light has been absorbed of each wavelength is easily distinguishable. fNIRS records the BOLD signal of the curved path the photons on average pass through (fig. 1D). For this curve to pass through the neocortex, optodes are placed 3-5 cm apart. This signal is pre-processed in different ways, e.g. as described in this study's methods section.

Existing work in improvements of machine learning (ML) for time series neural signals

Previous work has found that deep learning is beneficial in neural decoding because of its ability to learn complex and nonlinear transformations from the data (Hasson et al., 2020; Livezey & Glaser, 2020) even rivalling primate visual inferior temporal cortex visual performance (Cadieu et al., 2014). For these reasons, a large body of literature has amassed in the intersection between machine learning and neuroimaging (Craik et al., 2019; Davatzikos, 2019; Hennrich et al., 2015; Pereira et al., 2009; Varoquaux & Thirion, 2014; Zhuang et al., 2020) and is directly related to the rise of BCI systems (Wolpaw et al., 2000). Here, I summarise two articles that represent fNIRS machine learning and time series neuroimaging transfer learning.

Yoo et al. (2018) compare a classic statistical learning-based fNIRS ML framework to using an LSTM and get a three-class performance of 87% compared to 37%. They use the same features of signal slope and signal mean of the 2~7 seconds window as Hong et al. (2015) does in their representation of the classic paradigm. Each task was performed for 10 seconds. Their results show a high performance gap with LSTM and points out issues with the classic analysis process that makes it hard to process in real-time, a severe limitation for use in BCIs.

Makin et al. (2020) uses the similar time series signals of electrocorticography (ECoG) recorded from electrodes directly on the brain during brain surgery. They classify the words participants thought of using an LSTM sequence-to-sequence framework (S2S; Sutskever et al., 2014) with a temporal convolutional neural network (CNN; Krizhevsky et al., 2012) and get the same accuracy as professional speech transcription. They transfer the weights of successful models to improve performance drastically for a subject with very little data. The LSTMs are 3 400-unit bidirectional cells for encoding and 1 similar for decoding into words.

The main takeaways from these articles is that machine learning is a strong contender to the classic fNIRS analysis pipeline, that ML is enabling novel potentials of time series neural signal analysis and that transfer learning works well in low-resource scenarios where the target data is limited which supports our mixing of NLP and BCI. ML introduces several improvements over traditional fNIRS pipelines such as 1) analysis during live use (a core part of BCIs), 2) higher accuracy and better ability to see differences between classes and 3) less manual feature engineering.

Machine learning (ML) models for time series

Since fNIRS data output is a $channels \times time$ matrix, an important part of using machine- and transfer learning on fNIRS is the ability to work with 2D time series. Notable deep learning solutions include LSTM, CNN, Transformers and the algorithmic variations of these. CNNs have shown good performance with fNIRS (Ma et al., 2021) and Transformers are at the forefront of the natural-language processing revolution (Devlin et al., 2019). This paper uses LSTMs because of their use in the fNIRS literature (Rojas et al., 2021; Yoo et al., 2018) along with its characteristics designed for sequential time series data (Hochreiter & Schmidhuber, 1997).

From a high-level perspective, the LSTM network is able to remember for longer than its predecessors (Hochreiter & Schmidhuber, 1997) by selectively forgetting. The LSTM runs for each step in a time series and receives the time step data and the output of the previous time step's LSTM run. Refer to (Hochreiter & Schmidhuber, 1997) for further information about LSTMs**.**

Fine-tuning and transfer learning

Transfer learning "is the process of first training a base network on a source dataset and task, and then transfer the learned features (the network's weights) to a second network to be trained on a target dataset and task" (Ismail Fawaz et al., 2018, p. 1). Like language comprehension, a model can get a general understanding of the patterns of language and then use that for more specific tasks. This can reduce overfitting, decrease training time (Devlin et al., 2019; Radford et al., 2018) and increase performance (Yosinski et al., 2014). Pre-trained models might perform worse in cases where it overfits to the pre-training data (Ismail Fawaz et al., 2018; Yosinski et al., 2014).

Yosinski et al. (2014) analysed the generalisability of features using the AlexNet 8-layer architecture (Krizhevsky et al., 2012) on ImageNet (Deng et al., 2009). They found that transferring features from the early layers resulted in no performance change on a similar dataset while transferring the features up till the later layers resulted in drastic performance reductions. This indicates that early-stage modelling learns a lower-level representation that is directly transferable while later-stage modelling is too representation specific.

One way to solve these issues is to use fine-tuning. In transfer learning, the features of the NNs are blocked from learning while layers added for the target task are trained because of learning speed advantages. This methodology is practical in cases where the pre-trained networks contain billions of parameters but fine-tuning is the practice of letting the pre-trained network train on the target task as well. Yosinski et al. (2014) found that this mitigated accuracy reduction and even improved performance comparatively with pre-trained models no matter how many layers were included in the transfer. This is what we hope happens here.

Why pre-train a brain-computer interface (BCI) system

The solutions presented in this paper are meant to alleviate problems of signal instability, low signal-to-noise ratio and long initial adaptation time in BCI systems. These are basic problems in BCI and improvements in any will result in massive improvements downstream for clinical and civil use cases, instigated by the fact that properly classifying brain states is better and easier.

Hypotheses

From the theory introduced above, I will attempt to implement a pre-trained LSTM architecture for signal prediction and transfer the weights to classify different mental tasks, irrespective of which.

1. *An LSTM model can be trained to predict future brain signals compared to an optimal baseline from a sample of earlier brain signals.*
2. *A pre-trained model will enable higher performance in both **a)** speed and **b)** accuracy compared to a randomly initialised model.*

Additionally, the BCI, transfer learning and machine learning theory give us two more hypotheses. A multi-layer LSTM will learn different features by layer which will affect the transfer learning performance based on which layers are transferred to the target task and active BCI are inherently harder to distinguish because of the signal/noise ratio.

3. ***a)** Transferring early layers of the pre-trained model will give higher performance compared to other layers and **b)** fine-tuning will significantly increase performance.*
4. *Mental-only (active) BCI will have a higher benefit from pre-training compared to input-based (reactive) BCI because of data limitations.*

Collectively, these hypotheses will be telling in how well pre-training and transfer learning works for BCI applications. **An LSTM can be trained unsupervised (1) to enhance performance on a small-data BCI task (2) through designed fine-tuning transfer learning (3) with better improvement for active compared to reactive BCI (4).**

Methods and neuroimaging tools

Experimental paradigm and data

The participant in this study is the author of this paper, a 22 year old WEIRD (Henrich et al., 2010) male, referred to as S1. For the pre-training task, S1 had the fNIRS cap on in an unstructured experimental paradigm for 10-60 minutes at a time for 6 sessions (fig. 2D). All data was collected on a NIRSport2 NIRx headset (NIRx, 2021) through the NIRx' Aurora interface (NIRx, 2021) at 3.8Hz. The data contains 232 channels of fNIRS blood oxygen level measurements.

The transfer learning tasks consist of three pairs of differing levels of mental tasks: 1) Waving arms and talking, 2) doing mental arithmetic and listening to audio, and 3) doing mental arithmetic and a fully imagined box rotation task. These tasks were performed for 10 seconds since the BOLD signal is clear in that duration with a variable break in-between of 3.1 to 5.1 seconds to avoid temporally dependent noise (e.g. heart beats and breathing) that would be systematically present in the tasks without variability in the break (fig. 2C).

Pre-processing of the raw light intensity

For all tasks (unstructured and structured), the data went through a standard pre-processing sequence implemented from the standards in the MNE (Gramfort et al., 2014) Python library. The 232 channels are represented as a matrix and the raw intensity of light is converted to the differences in optical density (OD, absorbance) of the material the light has passed through (Dell, 2012). $I(t)$ is the intensity at time t and I^0 is the baseline where the difference is calculated from:

$$\Delta OD = -log\left(\frac{I(t)}{I^0}\right) = -log\left(\frac{I(t)}{mean(I)}\right)$$

We then use Temporal Derivative Distribution Repair (TDDR) to remove motion artefacts, a technique better than prominent alternatives (Fishburn et al., 2019). It assumes that fluctuations are normally distributed, that most fluctuations do not relate to motion artefacts and that motion artefacts have much greater fluctuations. It calculates the change in OD (fluctuations), uses an optimization algorithm to remove noise in the changes and integrates the signal to get the motion-corrected signal. See (Fishburn et al., 2019) for more details.

Then we apply the modified Beer-Lambert law (Kocsis et al., 2006) to calculate oxy- and deoxyhaemoglobin (HbO/HbR) from the OD. It is a generalisation of the Beer-Lambert law, allowing it to derive the concentration of molecules from the OD from an fNIRS source-detector pair. $c_{HbX}(t)$ is the change in concentration of HbX, $\Delta A(t, \lambda)$ is the change in absorbance (OD) of light of wavelength λ, DPF is the unit-less differential pathlength factor that is an estimate of how much longer the curved path is compared to the direct path, given the distance between a source-detector pair ρ and $\alpha_{HbX}(\lambda)$ is the extinction coefficient ($\mu M^{-1} mm^{-1}$) of HbX at wavelength $lambda$:

$$\Delta c_{HbX}(t) = \frac{\Delta A(t, \lambda)}{\langle L \rangle} = \frac{\Delta A(t, \lambda)}{DPF \cdot \rho \cdot \alpha_{HbX}(\lambda)}$$

For a two-λ signal such as the two wavelengths in an fNIRS system, this is reformulated as a matrix calculation for computational efficiency:

$$\begin{bmatrix} \Delta c_{HbO}(t) \\ \Delta c_{HbR}(t) \end{bmatrix} = \begin{bmatrix} \alpha_{HbO}(\lambda_1) & \alpha_{HbR}(\lambda_1) \\ \alpha_{HbO}(\lambda_2) & \alpha_{HbR}(\lambda_2) \end{bmatrix}^{-1} \begin{bmatrix} \Delta A(t, \lambda_1) \\ \Delta A(t, \lambda_2) \end{bmatrix} \frac{1}{\rho \cdot DPF}$$

The haemoglobin signal is then bandpass filtered to remove noise from heart rate (0.8Hz), breathing (0.2Hz), Mayer waves (0.1Hz), drift and machine noise (Izzetoglu, 2012; Naseer & Hong, 2015; Weyand et al., 2015). A FIR, one-pass, zero-phase, non-causal bandpass filter with a windowed time-domain design (firwin) was implemented with a band of 0.01-0.70Hz and a transition bandwidth of 0.005Hz and 0.30Hz respectively for the high-pass and low-pass. Because it is a zero-phase FIR filter, the entire data has to be available before filtering since it compensates for the filter delay compared to the signal. It is non-causal which means that future values in the data can affect past data points. In a real-life BCI example, a causal (e.g. linear) filter would be used instead.

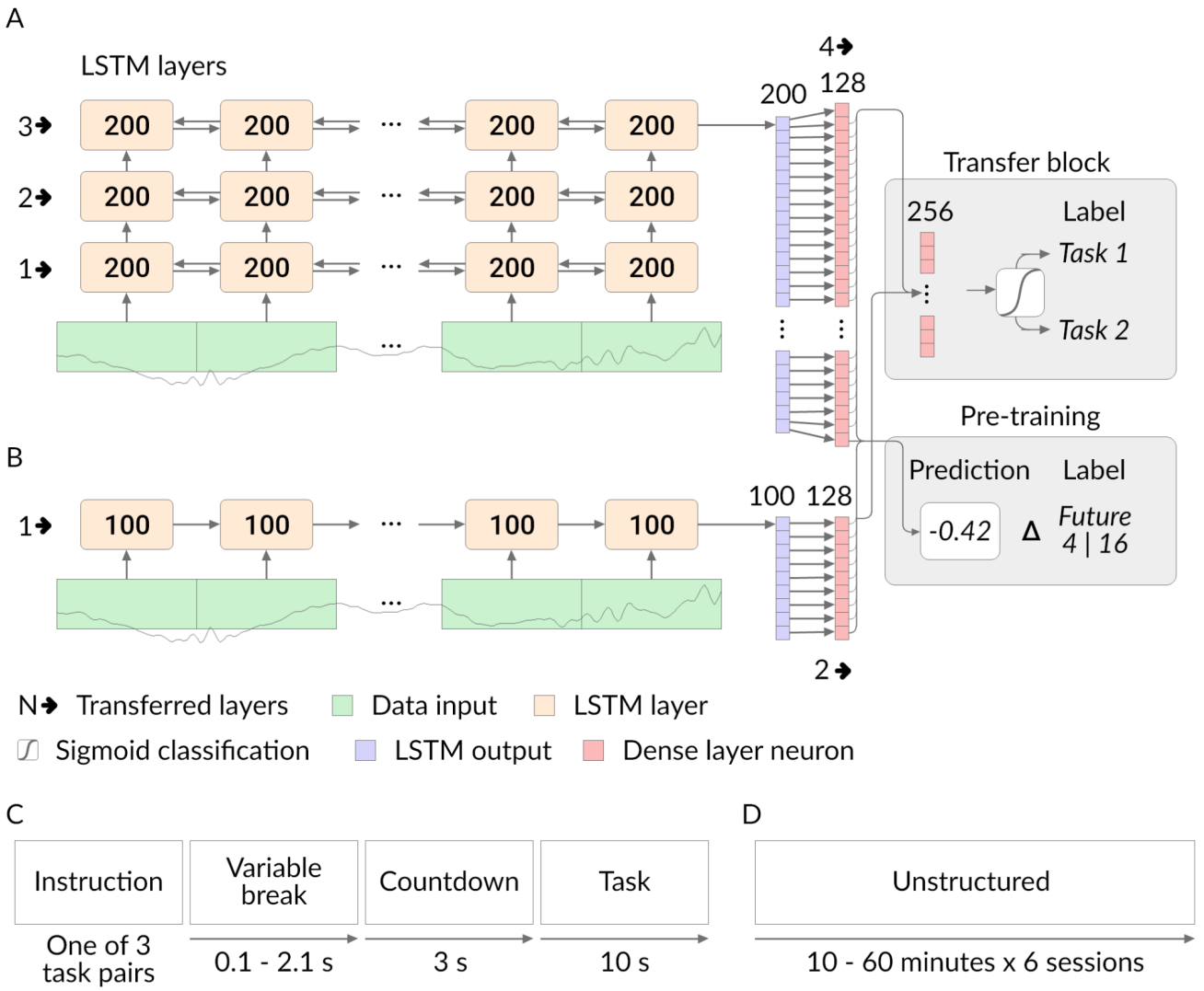

Figure 2: A) Model architecture for the 3-level stacked LSTM in both neural prediction and classification tasks along with how many layers are transferred for classification task, B) 2-layer pre-training architecture, C) BCI transfer tasks paradigm, D) pre-training experimental paradigm.

After the bandpass filter, the data is normalised around the Z-distribution by subtracting the mean and dividing by the standard deviation. C_{HbX} is the matrix for one type of haemoglobin concentration for all channels.

$$C_{HbX}^{norm} = \frac{C_{HbX} - \mu(C_{HbX})}{\sigma(C_{HbX})}$$

Constructing the pre-training and BCI task (transfer learning) datasets

All fNIRS output files were pre-processed with the above pipeline. The pre-training dataset was constructed in a manner where 10 seconds of data (39 samples) was set as the input and the value of the oxyhemoglobin concentration of the first channel 4 or 16 samples (~1 and ~4 seconds) ahead was extracted as the respective target value. This was repeated with a rolling window with a stride of 1, meaning that the 38 last samples of one input (39 samples) would overlap with the first 38 samples of the next input. One file was chosen as the test dataset and the other five was the training dataset. This created an effective train/test split of 71%.

Each of the BCI tasks datasets were pre-processed. The trials were marked in the fNIRS data with a trigger marker during the experiment. The 10-second samples of trial data were extracted and matched with their respective task, i.e. task 1 or task 2. Each task set of 10-second trials was split into train/test

and recombined to create the final train/test datasets, each consisting of a balanced number of trials for each task.

The three task pairs were waving arms and talking (42/39 trials), subtracting two-digit numbers from a three-digit number and listening to speech (43/41), and the same subtraction task and imagining rotating a box a specific amount of times up/down/left/right (39/40). These represent the reactive BCI paradigm where actions or stimuli create a neural reaction that the BCI can recognise and the active BCI where the internal workings of the brain are not visible from the outside but can be classified by the BCI.

Since the BCI tasks have between 79 and 84 trials, the data was augmented 10 times using the following three methods in sequence for HbX and HbO on each channel: 1) $scale(X)$ scales the whole vector by a uniformly random variable between 0.9 and 1.1, 2) $jitter(x_t)$ adds a Gaussian random number to x_t and 3) $M_\Delta(x_t)$ adds a vector that is a Gaussian random walk within a Gaussian probability space. Effectively, $M_\Delta(X)$ is X summed with a moving line constrained in the positive and negative direction:

$$
\begin{aligned}
scale(X) &= X \cdot Uniform(0.9, 1.1) \\
jitter(x_t) &= x_t + Gaussian(0, 0.05) \\
M_\Delta(x_t) &= x_t + N_{\Delta t}(0.1, 0.2) \\
&\text{where } N_{\Delta t}(x, y) = N_{\Delta(t-1)}(x, y) + \phi_{\text{sample}}(N_{\Delta(t-1)}(x, y) + N(0, x), 0, y) \cdot N(0, x) \\
&\text{where } (0, x) = Gaussian(0, x) \text{ and } \phi_{\text{sample}}(x, \mu, \sigma) \\
&\text{samples a normal distribution with mean } \mu \text{ and sd } \sigma
\end{aligned}
$$

This results in trial numbers between 869 and 924. The final datasets are the pre-training dataset and three BCI task datasets.

Designing the models

For the pre-training task, two baseline models were designed which consisted of the mean value of all data along with the last value of the test channel's input sequence. The main neural model architectures are seen on fig. 2A and fig. 2B. They are a three-layer LSTM and a one-layer LSTM, respectively. Both have a 128-neuron fully connected hidden layer before their output is evaluated with a linear single-neuron layer. An additional model consisting of one fully connected (dense) 128-neuron hidden layer between the input and output was designed as a baseline model. The inputs into the LSTMs are all the channels for 39 samples (~10 s), giving a token over the whole brain for each sample that the LSTMs train on. The dense model receives all 39 samples of only the test channel as a vector input.

Each LSTM layer has a dropout of 50% which means 50% of the neurons are deactivated for each session to avoid overfitting. The same is the case for the dense layer in all three models. The LSTM activation function is the tanh (Karlik & Olgac, 2010) which avoids activation run-off present in e.g. ReLU activation (Agarap, 2019) because of continuous multiplication through the input sequence, a traditional problem in sequence modelling (Hochreiter & Schmidhuber, 1997). Additionally, normalisation, early stopping and data augmentation were used to reduce overfitting to the training data.

Transfer learning means to transfer the weights of previous models to other models while fine-tuning means to adjust these weights on the new task as well. To perform binary classification on the BCI task pairs, a 256-neuron dense layer and a sigmoid output layer replaces the linear output of the pre-training phase (fig. 2A and 2B). The transferred layers on fig. 2A can either be trainable or not, leading to fine-tuning if they are.

For the pre-training task, mean absolute error is used as the loss function since it directly measures how imprecise the model predicts future activation. For the binary classification, binary cross-entropy is used since it measures how correctly confident the model is (Ramos et al., 2018). This is better than using accuracy since it also incorporates how confident the model is in its predictions. Cross-entropy is similar in nature to Kullback-Leibler divergence between the true distribution (test dataset) and the model's distribution (predictions).

Training the models

With the data already distributed into training and test datasets, the NAdam (Dozat, 2016) optimization algorithm was used with a learning rate of 0.0002. The pre-training models were each trained for 89-131 epochs (with a patience of 30) which showed strong convergence. All transfer learning models were

trained for 250 epochs. For model performance monitoring and testing management, the Weights & Biases software was used (Biewald, 2020). Other data used and visualised for this project are available at the associated Github link[1].

Results

Table 1 shows the prediction ability for the different model architectures trained to predict 4 samples or 16 samples ahead. The measure is the mean absolute error; lower is better. LSTM and LSTM-3 show similar performance while vastly outperforming both the fully-connected model and the last-value baseline. For comparison, the mean absolute error of predicting the global mean gives an error of 0.71.

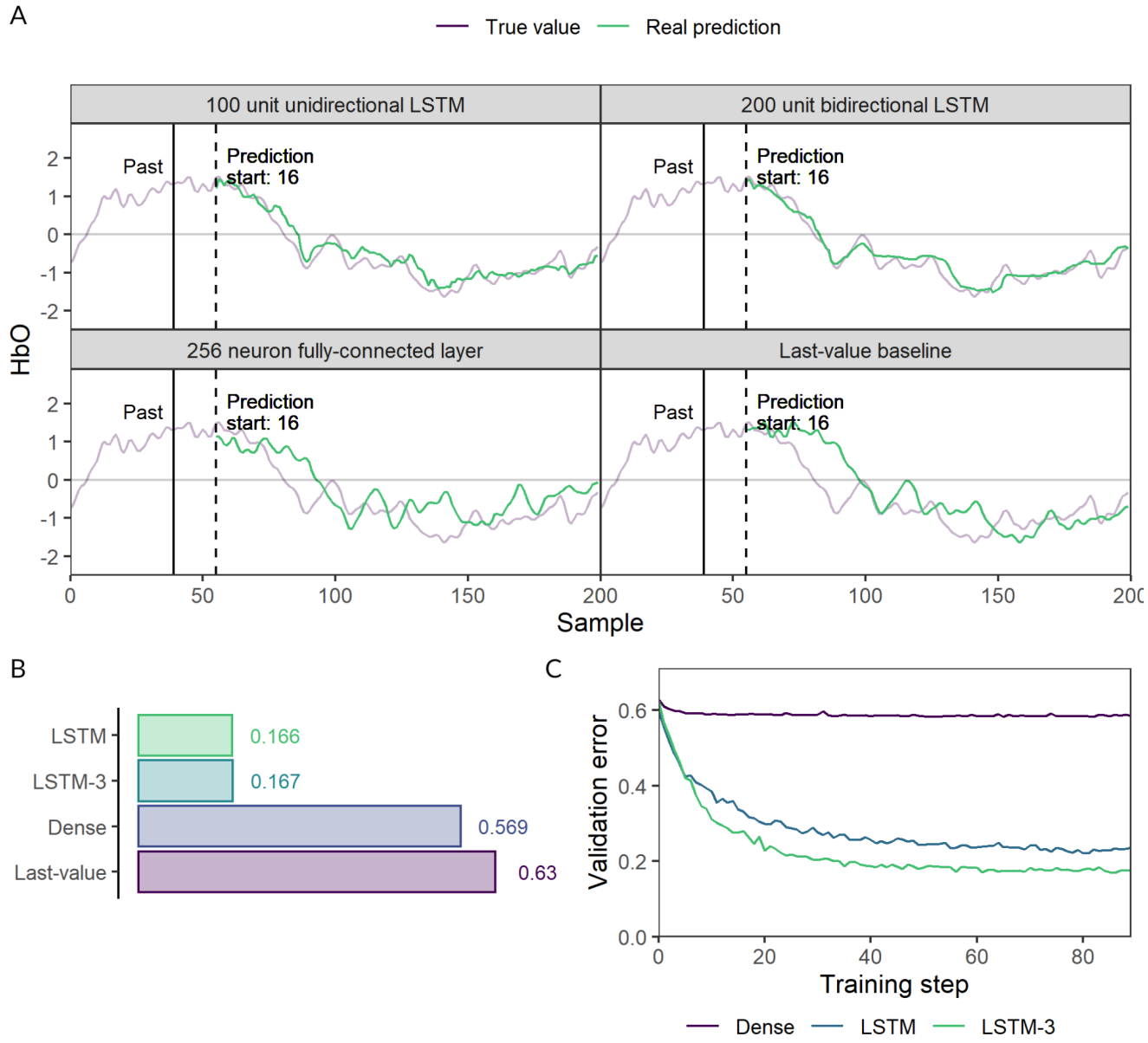

Figure 3: A) Predictions based on continuous real data for the three pre-trained models, B) Performance of the different models and the last-value baseline on all validation data, C) the three models' mean absolute error on the test dataset (notice that the LSTMs learn more than the Dense model)

[1] https://github.com/esbenkc/fnirs-bci

Table 1: Best (worst, n models). *The prediction's mean absolute error from the true value. Fully connected is a 128-neuron Dense ReLU layer with the input of the test channel's 10 second input. The best model is reported with the worst model in parenthesis. Smaller is better.*

	Last-value baseline	Fully connected	LSTM	3-stack LSTM
4-step	0.26	0.57 (0.58, n = 3)	**0.176** (0.21, n = 4)	0.183 (0.29, n = 4)
16-step	0.63	0.569 (0.57, n = 2)	**0.166** (0.26, n = 4)	0.167 (0.29, n = 5)

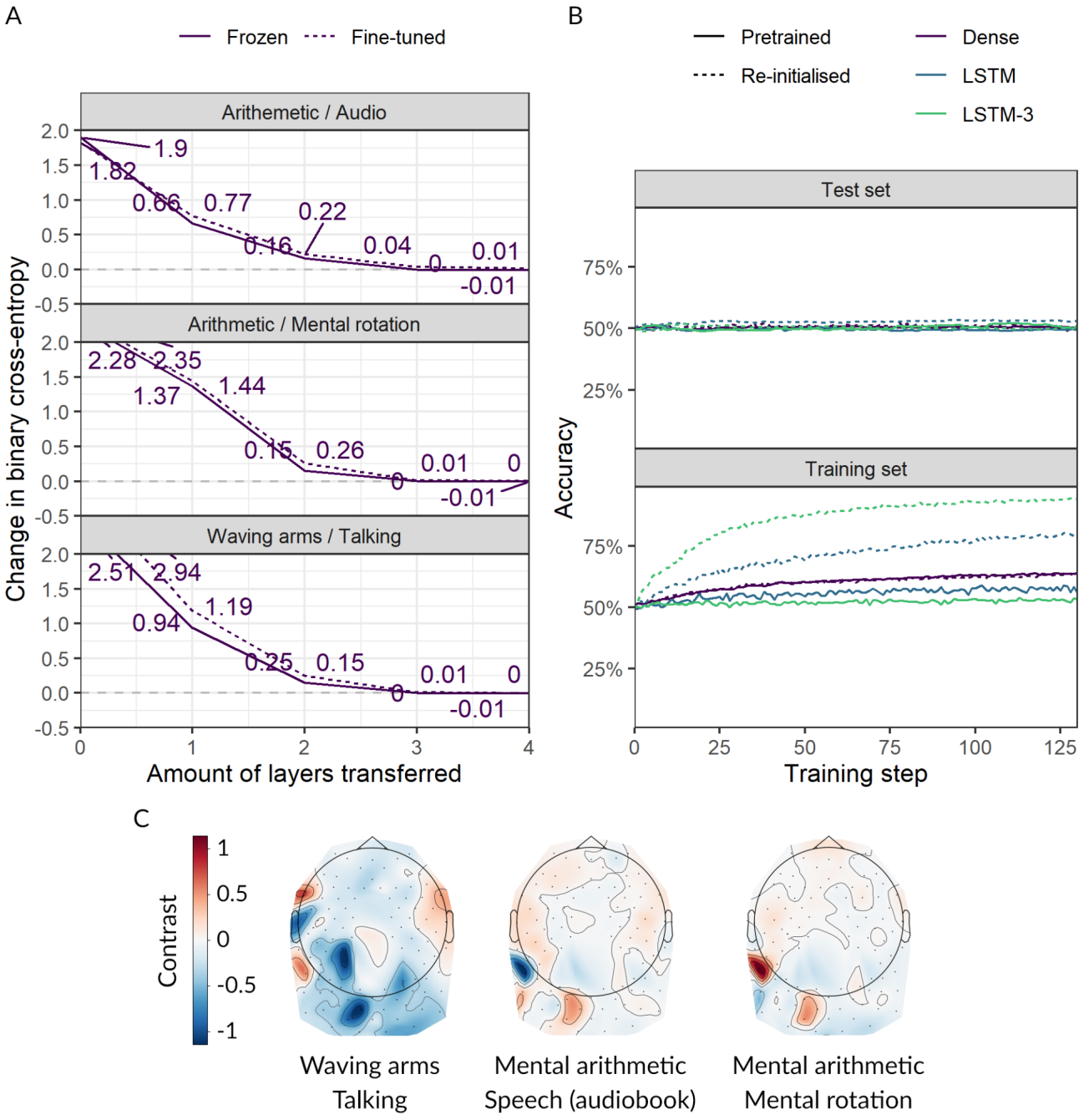

Figure 4: A) Validation loss at the last (250th) subtracted by the first step's validation loss at different amounts of layers transferred with the LSTM-3 model. Positive numbers mean the model has become worse. B) The performance on the training and test set for the different models either pre-trained or not.

C) The brain activity of task 1 subtracted by task 2 for each pair over the optodes (fig. 1). A small difference between the neural activation for each task means a smaller contrast.

FIgure 3 shows the models in table 1 in a live prediction task (fig. 3A) along with their performance (fig. 3B). Additionally, the validation error is shown in fig. 3C and shows that the models achieve a good fit without either over- or underfitting.

We also see that using pre-trained layers reduces the overfitting significantly (fig. 4A) but that the pre-trained models consistently underfit (fig. 4B). Pre-trained layers underfit and re-initialised layers overfit to the training data. Figure 4C shows the neural activation contrasts between the task pairs and shows how small the activation difference is between tasks.

Results discussion

Given the results in figure 3, 4 and table 1, **the LSTMs are able to predict future brain states (1, fig. 2A/B/C) to improve performance on a small-data thought classification task with pre-training (2, fig. 4B) through selective layer transfer learning (3, fig. 4A)** which is according to the hypotheses stated. It is however unclear whether the performance is better in thought classification (mental arithmetic or mental rotation) compared to activity classification (waving arms or talking) as proposed in hypothesis 4.

Despite the fact that nearly all models overfit to the training set compared to the validation set, the pre-training avoids excessive overfitting compared to re-initialised models and even though they severely underfit, their transferred weights avoid learning something wrong.

Predicting brain states from previous brain activity

From the results showcased in figure 3 and table 1, the models are clearly able to predict future brain signals from previous brain activity. This shows that models are able to learn some patterns in the data. Interestingly, we see that there is a limit to how the complexity of the model changes the performance, since the 3-layer stacked bidirectional LSTM does not fare any better than the single-layer unidirectional LSTM in this task. Plausibly, this is because of limitations in the amount of data given such a complex model (Hasson et al., 2020).

Classifying thoughts

The results in figure 4A and B indicate that all the models either underfit or severely overfit in the thought classification task (numbers in appendix 1A). Additionally, fig. 4C shows that the conditions are very similar and the waving arms / talking contrast is the only showing clear differences around the language area in the left inferior frontal cortex called Broca's area (Musso et al., 2003) and generally less activity, probably related to how waving the arms causes increased blood flow as a result of limb movement (Querido & Sheel, 2007).

Overfitting and underfitting are both general problems in ML (Dietterich, 1995; Jabbar & Khan, 2014). Overfitting is often related to too high model complexity capturing noise in the data and too little data. Due to the actual complexity of the multivariable modelling of brain signals (Markram, 2012), the models probably had the proper complexity but were missing the latter, the data. Even with data augmentation, the data points did not fill the state space properly which means the data augmentation augmented the same noise and did not properly embody the complexities of neural activity. The underfitting from the pre-trained models can largely be attributed to the same factors; there was not enough signal in the data for the models to properly fit.

When all that is said, the fact that the pre-trained models caused underfitting instead of overfitting means that information was carried over from the pre-training task which is a major success given the immense complexity of such a knowledge transfer between both tasks, datasets and models. The intention in comparing 4-sample and 16-sample prediction in table 1 was to show how the complex models retain performance while simpler models lose out, i.e. the last-value baseline. From this, we can infer that the complex models retain more "understanding" since they can model the data farther into the future (fig. 3A) and if we transferred the 4-sample predicting models, they might not reduce overfitting to the same degree.

A highly relevant result relating to this is the clear effect we see on the overfitting the more layers are transferred of the LSTM-3 (fig. 4A). Significantly, the two first layers show a much larger effect than the last two (LSTM and dense layer), indicating that they capture some underlying brain dynamics that prevent overfitting compared to the later-stage that might capture signals inferred from the first two

layers' embedding of the signal. This matches earlier work performing the same type of transfer layer analysis (Yosinski et al., 2014).

That speed of learning should decrease, accuracy should increase, and fine-tuning should increase performance (H2a, H2b, and H3b) is not possible to conclude on because of the underfitting. However, as mentioned, that transferring a specific amount of layers will affect performance (H3a) seems true, as it reduces the severe overfitting at different amounts (fig. 4a).

Active BCI compared to reactive BCI

The hypothesis that the mental-only (active) BCI tasks will have a higher benefit from pre-training is also hard to analyse. However, the initial assumption is that mental-only tasks will have smaller differences in neural activity is evident from fig. 4C which is itself interesting. Since the models underfit, the analysis of this difference in the model performance is impossible, however nothing in the results disconfirm this hypothesis which means future research can help enlighten this further.

Implications

With the ability to predict future brain states, the present paper shows the capabilities of LSTMs in learning to understand fNIRS brain signals. Additionally, the advanced transfer learning from a significantly different task to a classification task is shown to avoid overfitting and the possibility of good model performance with more data is a definite plausibility. For research purposes, this enables a precedent for future studies on the possibilities of neural signal classification from pre-trained models, something only previously shown in same-task transfer learning (Makin et al., 2020).

In regards to the main use case of brain-computer interfaces, the results are not definitive enough to conclude that we can use this sort of pre-training/transfer learning framework for BCI yet. However with more study, the framework might be validated in future work. The evidence that brain state prediction can be achieved seems to support this view and enable intuitions about what the next steps are.

Future research and limitations

There are several limitations to this study: 1) The data clearly limits the inferences we are able to make with the models, 2) the models were pre-trained on the first channel's HbO but further study should use a multi-channel prediction method since it would seem more robust, 3) the model evaluation happens on a training/test dataset but would be much more robust with a k-fold validation, 4) much advanced backend functionality in the implementation of the deep learning might limit the models' performance, 5) the LSTMs might have overfit due to focusing too much on noise in the signal which implicates other models and 6) other models such as Transformers and convolutional neural networks (Vaswani et al., 2017) should be compared to test their performances. Several of these were developed but omitted due to time constraints.

From these limitations, future research can focus on replicating the pipeline but adding 1) more BCI task data, 2) more pre-training data to match the LSTM-3 complexity, 3) developing the robustness of the models further and 4) comparing other similarly contemporary deep neural networks. Additionally, investigations into which channels and time steps the LSTM is triggered by might bring more intuition to what the models learn.

The study focused on a single-subject (S1) pre-training and transfer dataset, making it a 2-layer paradigm for transfer learning. For BCI usage, creating a more general, multi-subject pre-trained brain state prediction model that fine-tunes on the single-subject use case would give answers to hypotheses about a 3-layer paradigm. Thus, if possible, a population-level pre-trained model might be fine-tuned on the individual's unstructured brain data while that model is then used on the thought classification. Interesting experiments can be designed for all multiples of this sort of design, e.g. other subjects' transfer models might be a good pre-trained baseline for the single-subject transfer task.

Conclusion

Brain-computer interface technology can use more and more advanced methodologies to analyse brain signals and despite the difficulty of inferring thoughts with neural models, understanding neural signals with functional near-infrared spectroscopy (fNIRS) technology by blood oxygen flow is plausible using time series-specific neural models such as long short-term memory (LSTM) models.

The LSTM models in this paper show performance that indicates they can be used to understand and use patterns in fNIRS-measured neural activity to predict future neural activity. The paper's results also suggest that transferring the weights of these LSTMs and fine-tuning them on a classification task is

possible and reduces overfitting but needs more data than present in the study to reach optimum performance metrics. The results also suggest that specific early layers in deeper LSTM models embed the neural signals in ways that reduce overfitting more than later layers do.

Overall, the results described in this paper show how hard this task is and how much work is needed to properly elaborate the problem space but they are also strong preliminary steps to revolutionising the possibilities of BCI technologies with quicker and better model accuracy, leading to downstream benefits for patients and users alike.

References

Agarap, A. F. (2019). Deep Learning using Rectified Linear Units (ReLU). *ArXiv:1803.08375 [Cs, Stat]*. http://arxiv.org/abs/1803.08375

Aliakbaryhosseinabadi, S., Dosen, S., Savic, A. M., Blicher, J., Farina, D., & Mrachacz-Kersting, N. (2021). Participant-specific classifier tuning increases the performance of hand movement detection from EEG in patients with amyotrophic lateral sclerosis. *Journal of Neural Engineering*, *18*(5), 056023. https://doi.org/10.1088/1741-2552/ac15e3

Arthurs, O. J., & Boniface, S. (2002). How well do we understand the neural origins of the fMRI BOLD signal? *Trends in Neurosciences*, *25*(1), 27–31. https://doi.org/10.1016/S0166-2236(00)01995-0

Audette, W. (2013). *High-Quality, Low-Cost, Multi-Channel EEG System for Non-Traditional Users | SBIR.gov*. https://www.sbir.gov/sbirsearch/detail/408117

Biewald, L. (2020). *Experiment Tracking with Weights and Biases*. W&B. https://wandb.ai/site

Block Imaging. (2021). *MRI Machine Price Guide*. https://info.blockimaging.com/bid/92623/mri-machine-cost-and-price-guide

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., Brynjolfsson, E., Buch, S., Card, D., Castellon, R., Chatterji, N., Chen, A., Creel, K., Davis, J. Q., Demszky, D., … Liang, P. (2021). On the Opportunities and Risks of Foundation Models. *ArXiv:2108.07258 [Cs]*. http://arxiv.org/abs/2108.07258

Buzsaki, G. (2019). *The Brain from Inside Out*. Oxford University Press.

Cadieu, C. F., Hong, H., Yamins, D. L. K., Pinto, N., Ardila, D., Solomon, E. A., Majaj, N. J., & DiCarlo, J. J. (2014). Deep Neural Networks Rival the Representation of Primate IT Cortex for Core Visual Object Recognition. *PLOS Computational Biology*, *10*(12), e1003963. https://doi.org/10.1371/journal.pcbi.1003963

Chen, L.-C. (2016). *Cortical plasticity in cochlear implant users*. https://doi.org/10.13140/RG.2.2.33251.76324

Chen, Z., Zhang, H., Zhang, X., & Zhao, L. (2018). Quora question pairs. *URL Https://Www. Kaggle. Com/c/Quora-Question-Pairs*.

Craik, A., He, Y., & Contreras-Vidal, J. L. (2019). Deep learning for electroencephalogram (EEG) classification tasks: A review. *Journal of Neural Engineering*, *16*(3), 031001. https://doi.org/10.1088/1741-2552/ab0ab5

Davatzikos, C. (2019). Machine learning in neuroimaging: Progress and challenges. *NeuroImage*, *197*, 652–656. https://doi.org/10.1016/j.neuroimage.2018.10.003

Dell, E. (2012, October 22). *Optical Density for Absorbance Measurements | BMG LABTECH*. BMGLabtech.Com. https://www.bmglabtech.com/optical-density-for-absorbance-assays/

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). ImageNet: A large-scale hierarchical image database. *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 248–255. https://doi.org/10.1109/CVPR.2009.5206848

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *ArXiv:1810.04805 [Cs]*. http://arxiv.org/abs/1810.04805

Dietterich, T. (1995). Overfitting and undercomputing in machine learning. *ACM Computing Surveys*, *27*(3), 326–327. https://doi.org/10.1145/212094.212114

Dozat, T. (2016). *Incorporating Nesterov Momentum into Adam*. https://openreview.net/forum?id=OM0jvwB8jIp57ZJjtNEZ

Eagleman, D. (2020). *Livewired: The inside story of the ever-changing brain*. Pantheon.

Elango, V., Patel, A. N., Miller, K. J., & Gilja, V. (2018). *Sequence Transfer Learning for Neural Decoding*. https://openreview.net/forum?id=rybDdHe0Z

Fishburn, F. A., Ludlum, R. S., Vaidya, C. J., & Medvedev, A. V. (2019). Temporal Derivative Distribution Repair (TDDR): A motion correction method for fNIRS. *NeuroImage*, *184*, 171–179. https://doi.org/10.1016/j.neuroimage.2018.09.025

Gagnon, L., Yücel, M. A., Dehaes, M., Cooper, R. J., Perdue, K. L., Selb, J., Huppert, T. J., Hoge, R. D., & Boas, D. A. (2012). Quantification of the cortical contribution to the NIRS signal over the motor cortex using concurrent NIRS-fMRI measurements. *NeuroImage*, *59*(4), 3933–3940. https://doi.org/10.1016/j.neuroimage.2011.10.054

Gramfort, A., Luessi, M., Larson, E., Engemann, D. A., Strohmeier, D., Brodbeck, C., Parkkonen, L., & Hämäläinen, M. S. (2014). MNE software for processing MEG and EEG data. *Neuroimage*, *86*, 446–460.

Gürkök, H., Plass-Oude Bos, D., Bos, B., Laar, F., Nijboer, A., & Nijholt, A. (2011). User Experience Evaluation in BCI: Filling the Gap. *International Journal of Bioelectromagnetism Www.Ijbem.Org*, *13*, 54–55.

Hasson, U., Nastase, S. A., & Goldstein, A. (2020). Direct Fit to Nature: An Evolutionary Perspective on Biological and Artificial Neural Networks. *Neuron*, *105*(3), 416–434. https://doi.org/10.1016/j.neuron.2019.12.002

Hennrich, J., Herff, C., Heger, D., & Schultz, T. (2015). Investigating deep learning for fNIRS based BCI. *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 2844–2847. https://doi.org/10.1109/EMBC.2015.7318984

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). Most people are not WEIRD. *Nature*, *466*(7302), 29–29. https://doi.org/10.1038/466029a

Hill, N. J., & Wolpaw, J. R. (2016). Brain–Computer Interface☆. In *Reference Module in Biomedical Sciences*. Elsevier. https://doi.org/10.1016/B978-0-12-801238-3.99322-X

Hochberg, L. R., Bacher, D., Jarosiewicz, B., Masse, N. Y., Simeral, J. D., Vogel, J., Haddadin, S., Liu, J., Cash, S. S., van der Smagt, P., & Donoghue, J. P. (2012). Reach and grasp by people with tetraplegia using a neurally controlled robotic arm. *Nature*, *485*(7398), 372–375. https://doi.org/10.1038/nature11076

Hochreiter, S., & Schmidhuber, J. (1997). Long Short-Term Memory. *Neural Computation*, *9*(8), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

Homan, R. W., Herman, J., & Purdy, P. (1987). Cerebral location of international 10–20 system electrode placement. *Electroencephalography and Clinical Neurophysiology*, *66*(4), 376–382. https://doi.org/10.1016/0013-4694(87)90206-9

Hong, K.-S., Naseer, N., & Kim, Y.-H. (2015). Classification of prefrontal and motor cortex signals for three-class fNIRS–BCI. *Neuroscience Letters*, *587*, 87–92. https://doi.org/10.1016/j.neulet.2014.12.029

Ismail Fawaz, H., Forestier, G., Weber, J., Idoumghar, L., & Muller, P.-A. (2018). Transfer learning for time series classification. *2018 IEEE International Conference on Big Data (Big Data)*, 1367–1376. https://doi.org/10.1109/BigData.2018.8621990

Izzetoglu, M. (2012). Functional Optical Brain Imaging. In *Biosignal Processing*. CRC Press.

Jabbar, H. K., & Khan, R. Z. (2014). Methods to Avoid Over-Fitting and Under-Fitting in Supervised Machine Learning (Comparative Study). *Computer Science, Communication and Instrumentation Devices*, 163–172. https://doi.org/10.3850/978-981-09-5247-1_017

Johnson, B., & Kernel. (2020). *Kernel: Flow*. https://www.kernel.com/flow

Karlik, B., & Olgac, A. V. (2010). Performance Analysis of Various Activation Functions in Generalized MLP Architectures o. *Journal of Artificial Intelligence and Expert Systems*, 111–122.

Klein, F., & Kranczioch, C. (2019). Signal Processing in fNIRS: A Case for the Removal of Systemic Activity for Single Trial Data. *Frontiers in Human Neuroscience*, *13*, 331. https://doi.org/10.3389/fnhum.2019.00331

Kocsis, L., Herman, P., & Eke, A. (2006). The modified Beer–Lambert law revisited. *Physics in Medicine and Biology*, *51*(5), N91–N98. https://doi.org/10.1088/0031-9155/51/5/N02

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet Classification with Deep Convolutional Neural Networks. In F. Pereira, C. J. C. Burges, L. Bottou, & K. Q. Weinberger (Eds.), *Advances in Neural Information Processing Systems 25* (pp. 1097–1105). Curran Associates, Inc. http://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-netwo

rks.pdf
Lai, G., Xie, Q., Liu, H., Yang, Y., & Hovy, E. (2017). RACE: Large-scale ReAding Comprehension Dataset From Examinations. *ArXiv:1704.04683 [Cs]*. http://arxiv.org/abs/1704.04683
LeCun, Y., & Misra, I. (2021, March 4). *Self-supervised learning: The dark matter of intelligence*. https://ai.facebook.com/blog/self-supervised-learning-the-dark-matter-of-intelligence/
Livezey, J. A., & Glaser, J. I. (2020). Deep learning approaches for neural decoding: From CNNs to LSTMs and spikes to fMRI. *ArXiv:2005.09687 [Cs, q-Bio]*. http://arxiv.org/abs/2005.09687
Ma, T., Wang, S., Xia, Y., Zhu, X., Evans, J., Sun, Y., & He, S. (2021). CNN-based classification of fNIRS signals in motor imagery BCI system. *Journal of Neural Engineering*, *18*(5), 056019. https://doi.org/10.1088/1741-2552/abf187
Makin, J. G., Moses, D. A., & Chang, E. F. (2020). Machine translation of cortical activity to text with an encoder–decoder framework. *Nature Neuroscience*, *23*(4), 575–582. https://doi.org/10.1038/s41593-020-0608-8
Markram, H. (2012). THE HUMAN BRAIN PROJECT. *Scientific American*, *306*(6), 50–55.
Musk, E. & Neuralink. (2019). *An integrated brain-machine interface platform with thousands of channels* [Preprint]. Neuroscience. https://doi.org/10.1101/703801
Musso, M., Moro, A., Glauche, V., Rijntjes, M., Reichenbach, J., Büchel, C., & Weiller, C. (2003). Broca's area and the language instinct. *Nature Neuroscience*, *6*(7), 774–781. https://doi.org/10.1038/nn1077
Naseer, N., & Hong, K.-S. (2015). fNIRS-based brain-computer interfaces: A review. *Frontiers in Human Neuroscience*, *9*. https://doi.org/10.3389/fnhum.2015.00003
NIRx. (2021). *fNIRS Software: NIRS Analysis | NIRS Data Recording | fNIRS Systems | NIRS Devices | NIRx*. NIRx Medical Technologies. https://nirx.net/software
NIRx. (2021). *NIRSport2 | fNIRS Systems | NIRS Devices | NIRx*. NIRx Medical Technologies. https://nirx.net/nirsport
Pereira, F., Mitchell, T., & Botvinick, M. (2009). Machine learning classifiers and fMRI: A tutorial overview. *NeuroImage*, *45*(1, Supplement 1), S199–S209. https://doi.org/10.1016/j.neuroimage.2008.11.007
Querido, J. S., & Sheel, A. W. (2007). Regulation of Cerebral Blood Flow During Exercise. *Sports Medicine*, *37*(9), 765–782. https://doi.org/10.2165/00007256-200737090-00002
Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). *Improving Language Understanding by Generative Pre-Training*. 12.
Ramos, D., Franco-Pedroso, J., Lozano-Diez, A., & Gonzalez-Rodriguez, J. (2018). Deconstructing Cross-Entropy for Probabilistic Binary Classifiers. *Entropy*, *20*(3), 208. https://doi.org/10.3390/e20030208
Repovs, G. (2010). Dealing with noise in EEG recording and data analysis. *Informatica Medica Slovenica*, *15*(1), 18–25.
Rojas, R. F., Romero, J., Lopez-Aparicio, J., & Ou, K.-L. (2021). Pain Assessment based on fNIRS using Bi-LSTM RNNs. *2021 10th International IEEE/EMBS Conference on Neural Engineering (NER)*, 399–402. https://doi.org/10.1109/NER49283.2021.9441384
Schwemmer, M. A., Skomrock, N. D., Sederberg, P. B., Ting, J. E., Sharma, G., Bockbrader, M. A., & Friedenberg, D. A. (2018). Meeting brain–computer interface user performance expectations using a deep neural network decoding framework. *Nature Medicine*, *24*(11), 1669–1676. https://doi.org/10.1038/s41591-018-0171-y
Simon, J. (2021, October 26). *Large Language Models: A New Moore's Law?* https://huggingface.co/blog/large-language-models
Singh, S. P. (2014). Magnetoencephalography: Basic principles. *Annals of Indian Academy of Neurology*, *17*(Suppl 1), S107–S112. https://doi.org/10.4103/0972-2327.128676
Socher, R., Perelygin, A., Wu, J. Y., Chuang, J., Manning, C. D., Ng, A. Y., & Potts, C. (2013). *Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank*. 12.
Solovey, E. T., Girouard, A., Chauncey, K., Hirshfield, L. M., Sassaroli, A., Zheng, F., Fantini, S., & Jacob, R. J. K. (2009). Using fNIRS brain sensing in realistic HCI settings: Experiments and guidelines.

Proceedings of the 22nd Annual ACM Symposium on User Interface Software and Technology, 157–166. https://doi.org/10.1145/1622176.1622207

Subha, D. P., Joseph, P. K., Acharya U, R., & Lim, C. M. (2010). EEG Signal Analysis: A Survey. *Journal of Medical Systems*, *34*(2), 195–212. https://doi.org/10.1007/s10916-008-9231-z

Sun, C., Qiu, X., Xu, Y., & Huang, X. (2019). How to Fine-Tune BERT for Text Classification? In M. Sun, X. Huang, H. Ji, Z. Liu, & Y. Liu (Eds.), *Chinese Computational Linguistics* (pp. 194–206). Springer International Publishing. https://doi.org/10.1007/978-3-030-32381-3_16

Sutskever, I., Vinyals, O., & Le, Q. V. (2014). Sequence to Sequence Learning with Neural Networks. *Advances in Neural Information Processing Systems*, *27*. https://proceedings.neurips.cc/paper/2014/hash/a14ac55a4f27472c5d894ec1c3c743d2-Abstract.html

Varoquaux, G., & Thirion, B. (2014). How machine learning is shaping cognitive neuroimaging. *GigaScience*, *3*(1), 2047-217X-3–28. https://doi.org/10.1186/2047-217X-3-28

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention Is All You Need. *ArXiv:1706.03762 [Cs]*. http://arxiv.org/abs/1706.03762

Weyand, S., Schudlo, L., Takehara-Nishiuchi, K., & Chau, T. (2015). Usability and performance-informed selection of personalized mental tasks for an online near-infrared spectroscopy brain-computer interface. *Neurophotonics*, *2*(2), 025001. https://doi.org/10.1117/1.NPh.2.2.025001

Willett, F. R., Avansino, D. T., Hochberg, L. R., Henderson, J. M., & Shenoy, K. V. (2021). High-performance brain-to-text communication via handwriting. *Nature*, *593*(7858), 249–254. https://doi.org/10.1038/s41586-021-03506-2

Wolpaw, J. R., Birbaumer, N., Heetderks, W. J., McFarland, D. J., Peckham, P. H., Schalk, G., Donchin, E., Quatrano, L. A., Robinson, C. J., & Vaughan, T. M. (2000). Brain-computer interface technology: A review of the first international meeting. *IEEE Transactions on Rehabilitation Engineering*, *8*(2), 164–173. https://doi.org/10.1109/TRE.2000.847807

Wolpaw, J. R., Birbaumer, N., McFarland, D. J., Pfurtscheller, G., & Vaughan, T. M. (2002). Brain–computer interfaces for communication and control. *Clinical Neurophysiology*, *113*(6), 767–791. https://doi.org/10.1016/S1388-2457(02)00057-3

Yoo, S.-H., Woo, S.-W., & Amad, Z. (2018). Classification of three categories from prefrontal cortex using LSTM networks: FNIRS study. *2018 18th International Conference on Control, Automation and Systems (ICCAS)*, 1141–1146.

Yosinski, J., Clune, J., Bengio, Y., & Lipson, H. (2014). How transferable are features in deep neural networks? *ArXiv:1411.1792 [Cs]*. http://arxiv.org/abs/1411.1792

Zhuang, M., Wu, Q., Wan, F., & Hu, Y. (2020). State-of-the-art non-invasive brain–computer interface for neural rehabilitation: A review. *Journal of Neurorestoratology*, *8*(1). https://doi.org/10.26599/JNR.2020.9040001

Appendix

1A: Accuracy for different transfer learning models based on the task pair and if it is pre-trained or not. P-values are compared to the random baseline of 50%. There are no patterns indicating good overall performance.

BCI task pair	Architecture	Pre-trained	Final validation loss lower than start loss (“No” = overfitted)	Validation accuracy
Mental arithmetic (subtraction) and mental rotation (imagining rotating a box)	3-stack LSTM	Yes	Yes	54.86%
		No	No	52.92%
	LSTM	Yes	No	**59.03%**
		No	No	53.47%
	Dense	Yes	Yes	55.14%
		No	No	53.68%
Mental arithmetic (subtraction) and listening to speech	3-stack LSTM	Yes	Yes	55.15%
		No	No	**57.11%**
	LSTM	Yes	No	52.94%
		No	No	56.13%
	Dense	Yes	Yes	54.9%
		No	Yes	53.19%
Waving arms and talking	3-stack LSTM	Yes	Yes	54.43%
		No	No	51.56%
	LSTM	Yes	Yes	**56.51%**
		No	No	53.91%
	Dense	Yes	Yes	56.25%
		No	No	51.56%

1B: Parameters multiplicatively explored during transfer learning training:

1. Model architecture (dense, lstm, lstm-3)
2. BCI task pair (arms-talk, arithmetic-audio, arithmetic-rotation)
3. Layers transferred (0 = not pre-trained, 1, 2, 3, 4 = fully pre-trained)
4. Trainable (fine-tuned or not)

This is a total of 90 models (3 × 3 × 5 × 2).