Locating sources of uncertainty in an EEG classification task

Bachelor's Project Thesis

Martijn Wobbes, s4339312, m.s.wobbes@student.rug.nl,
Supervisor: I.P. de Jong

1 Introduction

Because of their high performance, deep neural networks have become the state-of-the-art method for various machine learning tasks. They are prevalent in computer vision, speech recognition, natural language processing, and the biomedical industry. However, because of their excellent performance, the outputs of these networks are often blindly trusted and presumed to be accurate, which is not always valid as the networks are generally overconfident. Generally, a network has no way of communicating how certain the model is with its predictions. Proper quantification of these uncertainties is becoming increasingly important in the practical field of machine learning. Uncertainty quantification gives the user a better insight into a model's prediction.

1.1 Uncertainty in machine learning

The uncertainty of a model consists of two types: aleatoric and epistemic uncertainty. Aleatoric uncertainty is the noise inherent in the observations, such as sensor noise. Feeding more data to the model will not help resolve this uncertainty. Epistemic uncertainty is the uncertainty in the model parameters itself. Feeding more data into a model can help resolve this type of uncertainty.

We can split aleatoric uncertainty further into two types. Homoscedastic aleatoric uncertainty is constant over all inputs. Heteroscedastic aleatoric uncertainty depends on the given input and differs between episodes. For example, homoscedastic aleatoric uncertainty can be sensor noise, constant across all episodes. In contrast, the heteroscedastic aleatoric uncertainty can only be a disconnected or faulty sensor in a few episodes.

1.2 Uncertainty quantification

Researchers have proposed various methods for the quantification of these uncertainties.

One of the most studied methods for modeling uncertainties is Bayesian Deep Learning Methods (Kendall & Gal, 2017; Chen et al., 2020). These models represent the weights as probability distributions rather than fixed values. The weights are treated as random values with prior distributions. After observing the data, the prior distributions are updated using Bayes' rule to obtain a posterior. The posterior can be used to make predictions, predicting a probability distribution rather than a fixed estimate. The width of this final distribution indicates the uncertainty of the model.

Nix & Weigend (1994) suggested a different approach, initially designed for regression tasks. Their strategy was to use a non-bayesian network with two output heads. One predicts the target mean μ, and one predicts the variance σ with the variance indicating the uncertainty of the model. The model will learn to predict the variance indirectly with the loss function, with erroneous predictions getting punished less if the predicted variance is high. Using a sampling softmax and a different loss function can also be used for classification (Kendall & Gal, 2017).

A different method of capturing predictive uncertainty involved deep ensembles (Lakshminarayanan et al., 2017). The core idea behind deep ensembles is to have multiple models. With each trained independently on the same dataset. The variance between the different models can then be used to estimate the uncertainty. The variance reflects the level of disagreement between the models, which can be used to measure the overall uncertainty.

Monte Carlo Dropout is another method for predicting uncertainty (Gal & Ghahramani, 2016). This method sets random weights to zero, canceling out their contribution. Multiple passes will be

done with random dropouts, resulting in an output distribution with which it is possible to calculate the uncertainty.

These methods can quantify uncertainty, each with its strengths and weaknesses. These methods can all answer the question of how uncertain a model is. However, they need to answer where this uncertainty comes from. Steps towards answering these questions can be made by introducing Explainable AI methods to the model. One approach is to integrate Shapley Values into the network, which approximates the contribution of features to the outcome (Merrick & Taly, 2020).

1.3 Error-related potentials

The medical field is one of the scientific fields that can benefit the most from uncertainty quantification. Generally, tasks such as medical diagnosis are of great importance and significance. However, data collection methods in this field, such as an electroencephalogram (EEG) or electrocardiogram (ECG), are very prone to noise. In the case of EEGs, artifacts, such as signals from blinking or malfunctioning electrodes, can easily corrupt EEG data.

BCIs are systems that use the brain's electrical activity to extract the intention from a user and translate it into computer instructions. These systems are helpful in fields such as prosthetics, where the user's intent can be translated into the movements of said prosthetic. Recently, error-related potentials (ErrP) have become a topic of interest. An ErrP signal is a response to error perception after feedback. These signals spark interest for BCIs since they indicate an erroneous extraction of intention, in which case the system can reconsider the translated action. Alternatively, the erroneous episode can be used as a training example to improve future performance.

Previous research has shown promising results in predicting ErrP signals (Correia et al., 2021). Specialized machine learning architectures designed explicitly for EEG signals have been designed and can be used for these tasks (Lawhern et al., 2018). However, these models all lack a form of uncertainty quantification.

1.4 Proposed method

In this paper, I propose a multi-headed Convolutional Neural Network classifying EEG signals on ErrP signals. The proposed model has two output nodes. One outputting the classification label and one outputting an estimate of the uncertainty. Furthermore, Shaply Values will be integrated into the model to indicate which features contribute the most to the uncertainty estimate, explaining the sources of uncertainty. The model will be trained and tested on the Monitoring Error-Related Potential dataset (Chavarriaga & Millán, 2010).

References

Chavarriaga, R., & Millán, J. d. R. (2010). Learning from eeg error-related potentials in noninvasive brain-computer interfaces. *IEEE transactions on neural systems and rehabilitation engineering*, *18*(4), 381–388.

Chen, W., Zhang, B., & Lu, M. (2020). Uncertainty quantification for multilabel text classification. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *10*(6), e1384.

Correia, J. R., Sanches, J. M., & Mainardi, L. (2021). Error perception classification in brain-computer interfaces using cnn. In *2021 43rd annual international conference of the ieee engineering in medicine & biology society (embc)* (pp. 204–207).

Gal, Y., & Ghahramani, Z. (2016). Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning* (pp. 1050–1059).

Kendall, A., & Gal, Y. (2017). What uncertainties do we need in bayesian deep learning for computer vision? *Advances in neural information processing systems*, *30*.

Lakshminarayanan, B., Pritzel, A., & Blundell, C. (2017). Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in neural information processing systems*, *30*.

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. *Journal of neural engineering*, *15*(5), 056013.

Merrick, L., & Taly, A. (2020). The explanation game: Explaining machine learning models using shapley values. In *Machine learning and knowledge extraction: 4th ifip tc 5, tc 12, wg 8.4, wg 8.9, wg 12.9 international cross-domain conference, cd-make 2020, dublin, ireland, august 25–28, 2020, proceedings 4* (pp. 17–38).

Nix, D. A., & Weigend, A. S. (1994). Estimating the mean and variance of the target probability distribution. In *Proceedings of 1994 ieee international conference on neural networks (icnn'94)* (Vol. 1, pp. 55–60).