Eberhard Karls Universität Tübingen

Mathematisch-Naturwissenschaftliche Fakultät

Bachelor Kognitionswissenschaft

Curvature-based step-sizes in deep Neural Networks

Bachelorarbeit

von

David Suckrow

geb. am 01. 01. 2000

in Leipzig

12345

Erstgutachter Prof. Dr.-Ing. Moritz Musterprof

Zweitgutachter: Dipl.-Ing. Manuela Musteringenieurin

Tübingen, Oktober 2024 – Februar 2025

Erklärung

Ich versichere wahrheitsgemäß, die Bachelorarbeit selbständig angefertigt, alle benutzten Hilfsmittel vollständig und genau angegeben und alles kenntlich gemacht zu haben, was aus Arbeiten anderer unverändert oder mit Abänderungen entnommen wurde.

.....

David Suckrow

Leipzig, den 01. 12. 2025

Inhaltsverzeichnis

Abbildungs- und Tabellenverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

1 Einleitung

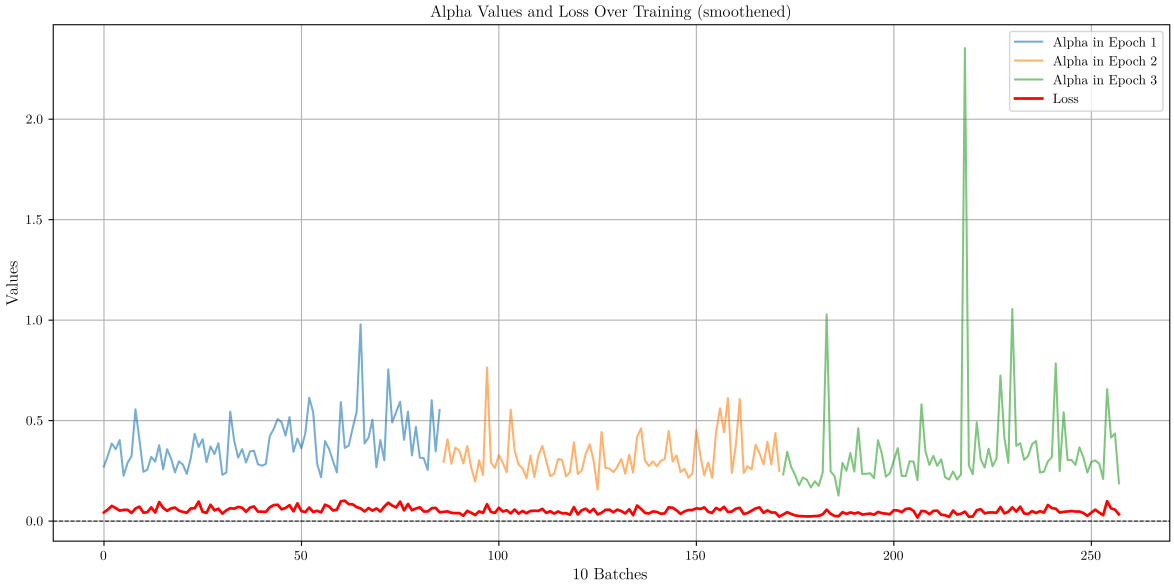
Artificial neural networks (ANNs) have proven to be a useful tool for complex questions that involve large amounts of data. Our use case of predicting soil maps with ANNs is in high demand by government agencies, construction companies, or farmers, given cost and time intensive field work. However, there are two main challenges when applying ANNs. In their most common form, deep learning algorithms do not provide interpretable predictive uncertainty. This means that properties of an ANN such as the certainty and plausibility of the predicted variables, rely on the interpretation by experts rather than being quantified by evaluation metrics validating the ANNs. Further, these algorithms have shown a high confidence in their predictions in areas geographically distant from the training area or areas sparsely covered by training data. To tackle these challenges, we use the Bayesian deep learning approach "last-layer Laplace approximation" , which is specifically designed to quantify uncertainty into deep networks, in our explorative study on soil classification. It corrects the overconfident areas without reducing the accuracy of the predictions, giving us a more realistic uncertainty expression of the model's prediction. In our study area in southern Germany, we subdivide the soils into soil * Corresponding author at: Department of Geoscience, University of Tübingen, Rümelinstraße 19-23, Tübingen 72070, Baden-Württemberg, Germany. E-mail addresses: kerstin.rau@uni-tuebingen.de (K. Rau), katharina.eggensperger@uni-tuebingen.de (K. Eggensperger), f.schneider@uni-tuebingen.de (F. Schneider), philipp.hennig@uni-tuebingen.de (P. Hennig), thomas.scholten@uni-tuebingen.de (T. Scholten). Contents lists available at ScienceDirect Science of the Total Environment journal homepage: www.elsevier.com/locate/scitotenv https://doi.org/10.1016/j.scitotenv.2024.173720 Received 2 February 2024; Received in revised form 31 May 2024; Accepted 31 May 2024 Science of the Total Environment 944 (2024) 173720 2 regions and as a test case we explicitly exclude two soil regions in the training area but include these regions in the prediction. Our results emphasize the need for uncertainty measurement to obtain more reliable and interpretable results of ANNs, especially for regions far away from the training area. Moreover, the knowledge gained from this research addresses the problem of overconfidence of ANNs and provides valuable information on the predictability of soil types and the identification of knowledge gaps. By analyzing regions where the model has limited data support and, consequently, high uncertainty, stakeholders can recognize the areas that require more data collection efforts. 1. Introduction The use of machine learning in science has become incredibly valuable and has significantly transformed many areas of research. The number of studies in which methods from the field of machine learning (ML) are used is constantly increasing (Zhang et al., 2022).

Soil science is one of the pioneers here, where extensive applications in the field of soil mapping were already developed at the beginning of this century (Behrens et al., 2005; McBratney et al., 2003). Today, digital soil mapping is one of the largest areas in which the methods are widely used for all kinds of climatic and geomorphometric regions of the World and in different areas of soil science, which has been demonstrated by numerous papers (Minasny and McBratney, 2016; Rentschler et al., 2022; Scull et al., 2003; Taghizadeh-Mehrjardi et al., 2021b; Zhang et al., 2022). Methodologically, applications of ML in soil science range from linear regression to modelling soil properties and their relationships to complex deep learning methods (Moore et al., 1993; Veres et al., 2015). The increasing use of these methods is not only due to their suitability for soil scientific and geographical questions, but also because producing soil type maps in the traditional way with cartographers surveying the landscape is very costly and time-consuming. This effort can be reduced with machine learning, especially for larger or even difficult to access areas (Behrens et al., 2005; Grunwald et al., 2011; Hewitt, 1993). At the same time, machine learning methods and their source code are becoming more accessible due to open-source software and widely available computational resources (Dramsch, 2020), and with the publication of several large open source datasets containing digital elevation models, climate data and other remote sensing data, especially those describing the vegetation, it is getting more convenient to apply them (Gascon et al., 2017; McBratney et al., 2003). Looking at the properties and functions of soils, for example carbon and water storage and plant nutrition, the soil type as a highly integrated prediction variable has the advantage that we can infer mechanical properties, dynamic processes and general characteristics from it with little effort (Albrecht et al., 2005; Hartemink and Bockheim, 2013). For example, Zhou et al. (2004) showed new spatial patterns in the predicted soil type map with their Bayesian predictive modelling approach. Grinand et al. (2008) uses classification tree analysis, which also supports decision-making in soil map extrapolation using machine learning methods. Adhikari et al. (2014) compared an existing soil map from a field survey with a predicted map calculated using a decision tree model. Artificial neural networks (ANNs) are currently one of the most popular machine learning methods (Taghizadeh-Mehrjardi et al., 2020, 2021a), as they are able to process large amounts of data and compute predictions comparably fast (Haykin, 1998; Schmidhuber, 2015; Silveira et al., 2013). Brungard et al. (2015) predicted soil taxonomy classes using eleven different models and found that the complex models containing neural networks were more accurate. Furthermore, Zhu (2000) found that ANNs can be used to obtain high-resolution soil maps. Although Heung et al. (2016) has also achieved good results with ANNs, they also have to admit that ANNs are difficult to interpret. Despite the results being rich in information, a major drawback of the predicted soil maps, and especially of the survey maps, is that they do not quantify the uncertainty of the

individual soil types at a given geographical location (Hengl et al., 2017). Instead, mostly is only given an overall accuracy statement in the form of a single statistical number. This is usually calculated as a coefficient using cross-validation techniques, where a subset of the training dataset is used to quantify the uncertainty of the overall performance (Wadoux et al., 2020). However, this is not sufficient, especially for regional or global tasks using unbalanced data sets, and that further analysis on uncertainty statements is needed, which was highlighted by (Meyer and Pebesma, 2022). Also, studies considering the uncertainty of predicted classes, like soil or vegetation classes, only looking at the probability of the predicted class or its confidence interval, have been criticized as well (Wadoux et al., 2020). They reported that out of 175 papers, only 30 most were focused on achieving high prediction accuracy and only a handful used machine learning methods for the uncertainty quantification. It is obvious that a better understanding and quantification of the uncertainty of soil maps modelled with ML is needed, especially when extrapolating from the training domain or when transferring the model to other more or less similar domains. In particular, working with ANNs as a black box requires such an assessment, as this model class is also known to be overconfident (Breiman, 2001; Nguyen et al., 2015; Hein et al., 2019). This means that ANNs can predict very reliable results, in our case soil classes, with a probability of up to 100 data is incorrect or uncertain. The lack of uncertainty measurement by the ANNs themselves makes it difficult to assess the reliability of the model predictions, which can lead to misinterpretations and incorrect decisions (Guo et al., 2017). With this study, we apply an ANN that predicts soil types inside and outside the known training domain in a trial study. We quantify the uncertainty of our model at every pixel in the area using last-layer Laplace Approximation (LLLA) (Kristiadi et al., 2020). Our aim is to add this uncertainty measurement to a soil classification problem to identify and correct the overconfidence of ANNs and to be able to spatially analyse and interpret in a following step the prediction of the ANN and its uncertainty derived from the LLLA. Further, we will discuss the transferability of the ANN to adjacent similar areas. Overall, our analyses will help to better understand and interpret results from ML models in soil science to provide new insights into soil processes and the spatial structure of the different domains. 2. Material and methods 2.1 Science of the Total Environment 944 (2024) 173720 3 similarities due to the likewise terrestrial geologic formation including sandstone. This great difference in such a small area naturally influences the vegetation and the processes in the soil, including soil formation. In total, the area comprises five major soil landscapes with different characterization, shown in Fig. 1B. These are areas in which, under similar geological, morphological and climatic conditions and under the influence of human, a landscape-typical association of soils has developed. 2.2. Data Fig. 2A shows the soil types in the study area, with each number representing a soil type and its characterization. The soil types are determined

according to the German soil classification system, which is based on the processes taking place in the soil and their properties (Eckelmann et al., 2005). In our area, there are 40 different soil types and the urban area, which is represented by the number 0. A detailed description can be seen in the Table 1, including the translation from the German into the World Reference Base (WRB) soil systematics (WRB, 2022). In order to preserve the diversity that is lost in this translation, we will stick to the German classification. The soil type map used for our prediction variable was initially provided by Landesamt für Geologie, Rohstoffe und Bergbau (LGRB) Baden-Württemberg as a polygon map (Fig. 2A). We converted this polygon map to a raster file using a rasterization function based on the digital elevation grid. While the original scale of the map is 1:50,000, its rasterization allowed to produce a raster with pixels of 10 $\times$10 m. As covariates for the neural network, exemplified in Fig. 2B, we looked for spatially dense data over the whole region to get as detailed data as possible, which is also important for the performance of the neural network. For this purpose, we use a digital elevation model (Fig. 2B(a)), which was also provided by the LGRB with a resolution of 10 m, based on which topographic indices were calculated, also with a resolution of 10 m. The decision on which of the variables we use as covariates is based on expert geographical knowledge of the region, commonly used variables in the geosciences and by using the SCORPAN model introduced by McBratney et al. (2003), which is based on Jenny (1983). To cover most of the covariates presented in the SCORPAN model, we also included satellite data. Copernicus provides the Sentinel-2 data, available from 2017 in 13 spectral bands with a 5-day repetition frequency. For us, the most important variables are the visible (R, G, B) and near-infrared bands, which have a resolution of 10 m. We use these spectral bands to calculate important indices such as the Normalized Difference Vegetation Index (Fig. 2B(d)) to describe vegetation cover. Finally, we calculate the median value for each index over the time series from March to May 2019. In our analysis, we used the median as the mean over years to mitigate the influence of outliers and to ensure a more robust representation of the data. To capture the influence of geology, we add a geological map with the scale of 1:50,000, provided by the LGRB and rasterized in the same way as the soil type map. We provide an overview of all the covariates used for the ANN and the corresponding references in Table 2. 2.3. Model architecture The origin of Artificial Neural Networks (ANNs) lies in the field of image recognition, especially in the area of classification (Goodfellow et al., 2016). These models are known for their ability to model multiple outcomes quickly and efficiently with a large amount of data, even with absence of prior knowledge about the data. Inspired by the neuronal structure, they look for dependencies and patterns in the given data that include input variables and a responding output variable. ANNS are organized in layers consisting of neurons using a (non-)linear activation function to transform and forward their inputs to the next layer,

allowing the ANN to learn complex patterns. The input layer receives the input data and consists of one neuron per input feature, in our case, one neuron per covariate. The neurons in the hidden layers pass the weighted sum of the outputs from the previous layer to their activation function. The final layer outputs the prediction and consists of one neuron per output variable, in our case, one neuron per soil type. During training the weights of the connections between the layers are learned via stochastic gradient descent to minimize a loss function measuring the error of the predictions. There is a wide variability of different constructs for an ANN for computation or information processing in terms of the architecture of the neural network, the number, types and dimensions of layers, or the activation function chosen. Since the focus of our study is on uncertainty of machine learning models in a soil context rather than on model performance, the simplicity of the model was very important to us. We choose a fully connected multilayer perceptron as described in Table 3. As the activation function for the hidden layer, the rectified linear unit function was chosen, first used by Hahnloser et al. (2000) and defined as



This plot shows alpha values and the corresponding loss:
Each datapoint represents the mean alpha values/loss values over 10 consecutive minibatches during training.

2 Grundlagen

3 Material und Methoden

4 Ergebnisse

5 Auswertung

6 Zusammenfassung und Ausblick

Anhang

Anhangsverzeichnis

A Beispiel 1