DTU Civil Engineering
Department of Civil Engineering

DEUCE: Distillation of Ensembles for Uncertainty of Classification Estimation

Project Work - Bachelor of Artificial Intelligence and Data, Fall 2024

DEUCE: Distillation of Ensembles for Uncertainty of Classification Estimation

Project Work - Bachelor of Artificial Intelligence and Data, Fall 2024
Date, year

By
Author

Cover photo: Vibeke Hempler, 2012
Published by: DTU, Department of Civil Engineering, Brovej, Building 118, 2800 Kgs. Lyngby Denmark
www.byg.dtu.dk
ISSN: [0000-0000] (electronic version)
ISBN: [000-00-0000-000-0] (electronic version)

ISSN: [0000-0000] (printed version)
ISBN: [000-00-0000-000-0] (printed version)

Approval

This thesis has been prepared over six months at the Section for Indoor Climate, Department of Civil Engineering, at the Technical University of Denmark, DTU, in partial fulfilment for the degree Master of Science in Engineering, MSc Eng.

It is assumed that the reader has a basic knowledge in the areas of statistics.

Author - s123456

.....

Signature

.....

Date

Abstract

Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like “Huardest gefburn”? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language.

Acknowledgements

Author, MSc Civil Engineering, DTU
Creator of this thesis template.

[Name], [Title], [affiliation]
[text]

[Name], [Title], [affiliation]
[text]

Contents

1 Introduction

Semantic classification tasks, such as sentiment analysis, topic categorization, and intent detection, are fundamental components of Natural Language Processing (NLP) applications. The development of deep learning has significantly improved performance in these tasks, largely due to models like BERT [1], GPT [2], and their derivatives. However, deep learning models often lack calibrated uncertainty estimates, which are crucial in applications where understanding the confidence of a prediction is as important as the prediction itself, such as in medical diagnosis, autonomous systems, and risk assessment [3].

Uncertainty estimation in deep learning models has garnered considerable attention in recent years. Techniques such as Bayesian neural networks [4], Monte Carlo Dropout [5], and deep ensembles [6] have been proposed to quantify predictive uncertainty. Among these, deep ensembles have demonstrated superior performance in both predictive accuracy and uncertainty estimation [6]. By training multiple instances of a model with different random initializations or data subsets, deep ensembles effectively capture model uncertainty.

Despite these advantages, deep ensembles are computationally expensive during both training and inference, making them less practical for large-scale or real-time applications. To address this issue, knowledge distillation techniques have been employed to transfer the knowledge from an ensemble of models into a single model [7]. Traditional knowledge distillation focuses on improving the predictive performance of the student model but often neglects the transfer of uncertainty information.

Recent work on uncertainty distillation aims to bridge this gap by enabling the student model to replicate not just the predictions but also the uncertainty estimates of the teacher ensemble [8]. However, most existing approaches have been explored in the context of computer vision, with limited investigation in the NLP domain.

1.1 State-of-the-art

Uncertainty estimation and knowledge distillation are important areas of research in deep learning, particularly within computer vision. However, their integration into NLP, especially for semantic classification tasks, is less mature. This section reviews the existing literature on uncertainty estimation in NLP, ensemble methods, and knowledge distillation.

1.1.1 Uncertainty Estimation in NLP

Uncertainty estimation in deep learning models is critical for applications that require reliable and interpretable predictions. In computer vision, methods such as Bayesian neural networks, Monte Carlo Dropout [5], and Deep Ensembles [6] have been extensively studied. In the NLP domain, these techniques have been adapted to account for the unique challenges posed by language data.

Gal and Ghahramani [5] introduced Monte Carlo Dropout as a Bayesian approximation, which has been applied to recurrent neural networks and transformer architectures in NLP [9]. Their approach enables models to estimate the predictive uncertainty by performing multiple stochastic forward passes during inference.

Desai and Durrett [10] evaluated the calibration of pre-trained transformer models like BERT [11] and found that while these models achieve high accuracy, they often produce

poorly calibrated confidence estimates. They highlighted the need for improved uncertainty estimation methods in NLP models to enhance their reliability in real-world applications.

1.1.2 Ensemble Methods in NLP

Ensemble methods have become integral to advancing state-of-the-art NLP, significantly increasing performance in tasks such as machine translation, text classification, and question answering. Recent strategies involve combining multiple transformer-based models, such as BERT [1], RoBERTa [12], and GPT [2] variants, to leverage their individual strengths and mitigate weaknesses. For example, ensembles of large pre-trained models have achieved top scores on benchmarks like GLUE and SuperGLUE by combining predictions to enhance accuracy and robustness [13][14]. In addition, ensemble techniques have been applied in multilingual settings to improve machine translation quality by incorporating various linguistic patterns captured by different models [15]. These developments show the critical role of ensemble methods in pushing the boundaries of NLP performance.

1.1.3 Knowledge Distillation

Knowledge distillation involves training a smaller ”student” model to replicate the behavior of a larger ”teacher” model or an ensemble of models [7]. This technique has been widely used to compress models and accelerate inference times without a substantial loss in performance.

In NLP, knowledge distillation has been employed to create compact versions of large language models. Sanh et al. [11] introduced DistilBERT, a distilled version of BERT, which retains much of the performance of BERT while being smaller and faster. Jiao et al. [16] proposed TinyBERT, using a two-stage learning framework to distill both the embedding and prediction layers.

1.1.4 Uncertainty Distillation

Although traditional knowledge distillation focuses on transferring predictive performance, recent research has explored distilling uncertainty estimates as well. Uncertainty distillation aims to train a student model that replicates both the predictions and the uncertainty estimates of the teacher ensemble.

Malinin and Gales [17] introduced Prior Networks, which model predictive uncertainty by parameterizing distributions over output probabilities. Gast and Roth [18] proposed a method for uncertainty modeling in deep neural networks through distillation, primarily in computer vision contexts.

1.2 Learning goals

- Better understanding of deep ensemble models and the best way to prepare and augment the dataset to achieve the lowest uncertainty and highest precision.
- How the code for knowledge distillation of a model works in Pytorch
- Training large models on DTUs HPC
- Better time and project management

2 Dataset Description

The Stanford Sentiment Treebank 2 (SST-2) [19] is a refined version of the original Stanford Sentiment Treebank, designed for binary sentiment classification tasks. The dataset is built on movie reviews extracted from the internet and is commonly used to evaluate sentiment analysis models in NLP. The SST-2 dataset provides labels indicating whether a given sentence expresses a positive or negative sentiment.

2.1 Data Structure

- **train.parquet**: Contains training data consisting of sentences with corresponding sentiment labels (positive/negative).
- **test.parquet**: Contains test data used to evaluate the model's performance, without sentiment labels (evaluation is conducted by submitting predictions for this set).
- **validation.parquet**: Contains validation data for model tuning, with the same format as the training set.

2.2 Features

- **sentence**: The text of a movie review sentence or phrase.
- **label**: The sentiment class assigned to the sentence:
  - 0: Negative sentiment
  - 1: Positive sentiment

2.3 Use Case

The SST-2 dataset is used to train, validate and test machine learning models that perform binary sentiment analysis. It serves as a benchmark dataset for tasks involving the classification of the emotional tone in text.

2.4 Size

- **Number of training examples**: $67,349$
- **Number of validation examples**: 872
- **Number of test examples**: $1,821$

2.5 Data Format

- **File format**: Parquet
- **Text encoding**: UTF-8

2.6 Applications

This dataset is often used to train models for:

- Sentiment analysis in natural language processing
- Text classification tasks
- Binary classification model evaluation

2.7 Access and License

The dataset is publicly available at https://huggingface.co/datasets/stanfordnlp/sst2 and is licensed under the

3 Research Questions

1. How does ensemble models improve uncertainty estimation for NLP?
2. What is the impact of distilling an ensemble of models into a smaller, more efficient single model on both predictive performance and uncertainty estimation, and how significant are the changes in these metrics?
3. In what ways does uncertainty-aware distillation contribute to enhancing a model's robustness to out-of-distribution data in NLP tasks, and what factors influence its effectiveness?

4 Methodology

4.1 Overview

This section outlines the methodology used to answer the research questions posed earlier. Our approach uses ensemble methods consisting of Multi-Layer Perceptrons (MLPs) for semantic classification on the SST-2 dataset. We utilize knowledge distillation techniques to transfer both predictive performance and uncertainty estimates from the ensemble to a single, more efficient student model. Evaluation metrics are utilized to assess the effectiveness of this approach in terms of both classification accuracy and uncertainty estimation.

4.2 Large Language Models

Large Language Models (LLMs) such as BERT [1], GPT [2], and their derivatives have significantly advanced the field of NLP. These models are pre-trained on vast datasets and fine-tuned for specific tasks, achieving state-of-the-art performance in areas like sentiment analysis, machine translation, and question answering.

But despite their success, LLMs do have notable limitations such as computational complexity, lack of uncertainty estimation, and deployment challenges. LLMs require substantial computational resources for both training and inference, which can be prohibitive in resource-constrained environments or real-time applications. While LLMs provide high accuracy, they often lack calibrated uncertainty estimates, making it challenging to assess the confidence of their predictions. The size and complexity of LLMs make them difficult to deploy on devices with limited memory and processing capabilities.

To address these challenges, our methodology focuses on using smaller, more efficient models like MLPs, augmented through ensemble methods and knowledge distillation to retain high performance and reliable uncertainty estimates.

4.3 Ensemble Methods with MLPs

Ensemble methods improve model performance by combining the predictions of multiple models, thereby reducing variance and capturing model uncertainty. In our approach, the ensemble consists of several MLPs trained independently, each contributing to the overall predictive capability and uncertainty estimation.

4.3.1 Multi-Layer Perceptrons

MLP is a feedforward artificial neural network that maps input data to appropriate outputs through multiple layers of interconnected neurons. Each neuron applies a non-linear activation function to a weighted sum of its inputs.

Key characteristics of the MLPs used in our ensemble include:

- **Input Layer**: Receives feature representations of input sentences, typically derived from word embeddings or sentence encodings.
- **Hidden Layers**: One or more layers with non-linear activation functions (e.g., ReLU) that capture complex patterns in the data.
- **Output Layer**: Produces the final prediction, using a sigmoid activation function for binary classification tasks.

4.3.2 Ensemble Construction

The ensemble is constructed by training a number independent MLP models, where amount of MLP models is determined based on computational resources. Each MLP is initialized with different random weights to ensure diversity among ensemble members.

4.4 Evaluation Metrics

The effectiveness of the distilled model is evaluated using metrics that assess both predictive performance and the quality of uncertainty estimation.

4.4.1 Predictive Performance

- **Accuracy**: The proportion of correctly predicted instances over the total instances.
- **Precision**: The ratio of true positive predictions to the total positive predictions made by the model.
- **Recall (Sensitivity)**: The ratio of true positive predictions to all actual positive instances.
- **F1 Score**: The harmonic mean of precision and recall, providing a balance between the two.

4.4.2 Uncertainty Estimation

- **Expected Calibration Error (ECE)**: Measures the discrepancy between predicted probabilities and actual outcomes, indicating how well the predicted confidences reflect true likelihoods.
- **Brier Score**: Calculates the mean squared difference between predicted probabilities and actual outcomes, assessing the accuracy of probabilistic predictions.
- **Negative Log-Likelihood (NLL)**: Evaluates how well the predicted probability distributions match the observed data, penalizing overconfident incorrect predictions.

References

[1] Jacob Devlin et al. “BERT: Pre-training of deep bidirectional transformers for language understanding”. In: *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*. Vol. 1. 2019.

[2] Alec Radford. “Improving Language Understanding by Generative Pre-Training”. In: *Homology, Homotopy and Applications* 9.1 (2018). ISSN: 15320081.

[3] Aryan Mobiny, Aditi Singh, and Hien Van Nguyen. “Risk-aware machine learning classifier for skin lesion diagnosis”. In: *Journal of Clinical Medicine* 8.8 (2019). ISSN: 20770383. DOI: 10.3390/jcm8081241.

[4] Alex Kendall and Yarin Gal. “What uncertainties do we need in Bayesian deep learning for computer vision?” In: *Advances in Neural Information Processing Systems*. Vol. 2017-December. 2017.

[5] Yarin Gal and Zoubin Ghahramani. “Dropout as a Bayesian approximation: Representing model uncertainty in deep learning”. In: *33rd International Conference on Machine Learning, ICML 2016*. Vol. 3. 2016.

[6] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. “Simple and scalable predictive uncertainty estimation using deep ensembles”. In: *Advances in Neural Information Processing Systems*. Vol. 2017-December. 2017.

[7] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. “Distilling the Knowledge in a Neural Network”. In: (Mar. 2015).

[8] Steven Landgraf et al. “DUDES: Deep Uncertainty Distillation using Ensembles for Semantic Segmentation”. In: (Mar. 2023). DOI: 10.1007/s41064-024-00280-4.

[9] Yarin Gal and Zoubin Ghahramani. “A theoretically grounded application of dropout in recurrent neural networks”. In: *Advances in Neural Information Processing Systems*. 2016.

[10] Shrey Desai and Greg Durrett. “Calibration of pre-trained transformers”. In: *EMNLP 2020 - 2020 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference*. 2020. DOI: 10.18653/v1/2020.emnlp-main.21.

[11] Victor Sanh et al. “DistilBERT”. In: *arXiv* (2019).

[12] Kian Long Tan, Chin Poo Lee, and Kian Ming Lim. “RoBERTa-GRU: A Hybrid Deep Learning Model for Enhanced Sentiment Analysis”. In: *Applied Sciences (Switzerland)* 13.6 (2023). ISSN: 20763417. DOI: 10.3390/app13063915.

[13] Alex Wang et al. “GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding”. In: *EMNLP 2018 - 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP, Proceedings of the 1st Workshop*. 2018. DOI: 10.18653/v1/w18-5446.

[14] Alex Wang et al. “SuperGLUE: A stickier benchmark for general-purpose language understanding systems”. In: *Advances in Neural Information Processing Systems*. Vol. 32. 2019.

[15] Roee Aharoni, Melvin Johnson, and Orhan Firat. “Massively multilingual neural machine translation”. In: *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*. Vol. 1. 2019. DOI: 10.18653/v1/n19-1388.

[16] Xiaoqi Jiao et al. “TinyBERT: Distilling BERT for natural language understanding”. In: *Findings of the Association for Computational Linguistics Findings of ACL: EMNLP 2020*. 2020. DOI: 10.18653/v1/2020.findings-emnlp.372.

[17] Andrey Malinin and Mark Gales. “Predictive uncertainty estimation via prior networks”. In: *Advances in Neural Information Processing Systems*. Vol. 2018-December. 2018.

[18] Jochen Gast and Stefan Roth. “Lightweight Probabilistic Deep Networks”. In: *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. 2018. DOI: 10.1109/CVPR.2018.00355.

[19] Richard Socher et al. *Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank*. Seattle,Washington,USA, Oct. 2013, pp. 1631–1642.

A Title

Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like “Huardest gefburn”? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language.

Technical
University of
Denmark

Brovej, Building 118
2800 Kgs. Lyngby
Tlf. 4525 1700

www.byg.dtu.dk