Universiteit Leiden
The Netherlands

Learning from small samples

Kocaman, V.

Chapter 1

Introduction

1.1 Background

Machine Learning (ML) and Deep Learning (DL) have proven to be powerful tools for solving complex real-world problems. These techniques have been applied to a wide range of domains, including speech recognition, computer vision, natural language processing, recommendation systems and more. However, the performance of these models relies heavily on the availability of large amounts of annotated data, which is often a key barrier for companies to adopt machine learning. Labeling data gets expensive, and the difficulties of sharing and managing large datasets for model development make it a struggle to get machine learning projects off the ground in practical applications.

Labeled data is a group of samples that have been tagged with one or more labels. After obtaining a labeled dataset, machine learning models can be applied to the data so that new, unlabeled data can be presented to the model and a likely label can be guessed or predicted for that piece of unlabeled data. When the amount of available labeled data is limited, ML models tend to overfit on the training data, leading to poor generalization performance on unseen data whereas human can learn new concepts with just a few example and can often generalize successfully

from just a single example. This is particularly problematic in fields where data collection is time-consuming and expensive, such as medicine and biology. In these domains, the limited availability of labeled data restricts the development of powerful models that can provide meaningful insights and predictions.

In addition to the challenges associated with small datasets, ML models also face the problem of imbalanced data. Imbalanced data refers to a scenario in which the number of observations belonging to one class is significantly lower than those belonging to the other classes. This is a common problem in business contexts where accurate prediction is crucial, such as detecting fraudulent transactions, identifying rare diseases, and predicting customer churn. Standard ML algorithms may not produce accurate results when applied to imbalanced datasets, as they are designed to reduce error rather than consider the class distribution or balance of classes.

In short, imbalanced data concept refers to the situation where the number of examples from each class in a dataset is highly imbalanced. This can lead to biased models that perform poorly on the underrepresented class. Anomaly detection is another important challenge associated with small datasets, where the goal is to identify data instances that deviate from the normal behavior.

In order to overcome these challenges, various approaches have been proposed to effectively learn from small datasets in ML. These include data selection and preprocessing, incorporating domain, prior and context knowledge, ensemble methods, transfer learning, parameter initialization, loss function reformulation, regularization techniques, data augmentation, synthetic data generation, and more. The use of these techniques enables the development of models that are capable of effectively learning from limited amounts of data and generalizing to unseen data.

1.2 Objectives

The main objective of this dissertation is to explore various approaches for effectively learning from small datasets in ML, and to address the challenges of imbalanced data and anomaly detection. This will contribute to the development of more robust and efficient models that can be applied to a wide range of real-world

problems where the availability of labeled data is limited. The specific objectives are as follows:

- To identify and analyze the problems associated with small datasets and how they affect the performance of ML models.
- To review and evaluate different techniques for handling small datasets, including data selection and preprocessing, incorporating prior knowledge, and the use of ensemble methods.
- To examine transfer learning and how it can be used to tackle small data problems.
- To investigate various optimization techniques, such as parameter initialization and loss function reformulation, and how they can be used to improve the performance of ML and DL models on small datasets.
- To study regularization techniques and how they can be used to prevent overfitting in ML and DL models.
- To explore data augmentation and synthetic data generation as potential solutions for small dataset problems.
- To evaluate the performance of self-supervised, semi-supervised, and unsupervised learning techniques on small datasets.
- To investigate the potential of using physics-informed neural networks, meta learning, and active learning to handle small data problems.
- To analyze the problem of imbalanced data and how it can be addressed in small sample settings.
- To examine the challenges of anomaly detection as a small data problem and potential solutions.

The proposed research will provide a comprehensive overview of the challenges and solutions associated with learning from small datasets in ML, and will contribute to the development of more effective and efficient models for this task.

1.2.1 Research Questions

RQ1 (Chapter 3) What are the current methods and techniques used to effectively learn from small datasets in Machine Learning and overcome the challenges posed by small data and extreme imbalance?

RQ2 (Chapters 4) Being one of the most effection regularization technique, how does adding a batch normalization layer just before the softmax output layer in modern CNN architectures affect the training time and test error for minority classes in highly imbalanced datasets, and what is the impact of this technique on the overall performance of the model in terms of recall for the minority class?

RQ3 (Chapters 5) How does the use of salient image segmentation as an augmentation policy in Self-Supervised Learning (SSL) impact the representation and generalization capabilities of images in downstream tasks such as image segmentation, and how does this method compare to other commonly used augmentation policies in SSL?

1.3 Outline of the Dissertation

Chapter 3 provides an overview of the challenges that arise in machine learning when dealing with small data. It starts by discussing the problem of overfitting and generalization, and how they can be mitigated when the data is limited. Then, the chapter delves into the various approaches that have been proposed to effectively learn from small datasets. Handling small data includes a detailed discussion of various techniques that can be used to overcome the limitations posed by small datasets. These techniques are grouped under several headings, such as data selection and preprocessing, incorporating domain, prior, and context knowledge, ensemble methods, transfer learning, parameter initialization, loss function reformulation, regularization techniques, data augmentation, synthetic data generation, problem reduction, optimization techniques, and more.

This chapter also includes a discussion of the various approaches that have been proposed to tackle small data problems, with a focus on how they can be used to effectively learn from limited amounts of data. These approaches include, but

are not limited to, using physics-informed neural networks, unsupervised learning techniques, semi-supervised learning, self-supervised learning, zero-shot, one-shot, and few-shot learning, meta-learning, harnessing model uncertainty, active learning, self-learning, multi-task learning, symbolic learning, hierarchical learning, and knowledge distillation based learning.

Chapter 4 covers the problem of learning from data that is highly skewed towards one class, and the various techniques that have been proposed to mitigate this problem. This chapter mainly focuses on the impact of batch normalisation on learning from small datasets. It provides experimental evidence of the positive impact of adding an additional batch normalisation layer just before the softmax output layer on reducing the training time and test error for minority classes in a highly imbalanced dataset. The results show that this approach can lead to a significant improvement in the performance of the model, particularly when a high recall is desired for the minority class.

Chapter 5 focuses on the impact of self-supervised learning on learning from small datasets. The chapter provides experimental evidence of the positive impact of using salient image segmentation as an augmentation policy in self-supervised learning, when the downstream task is image segmentation. The results indicate that using this augmentation policy leads to better image representations, as compared to using default augmentations or no augmentations at all. The chapter concludes with a discussion of the potential of self-supervised learning in mitigating the limitations posed by small datasets.

1.4 Author's Contributions

The main contributions of the author of this dissertation are the following:

[1] Veysel Kocaman, Ofer M Shir, and Thomas Bäck. Improving model accuracy for imbalanced image classification tasks by adding a final batch normalization layer: An empirical study. In *2020 25th International Conference on Pattern Recognition (ICPR)*, number 10.1109/ICPR48806.2021.9412907, pages 10404–10411. IEEE, 2021.

[2] Veysel Kocaman, Ofer M Shir, and Thomas Bäck. The unreasonable effectiveness of the final batch normalization layer. In *International Symposium on Visual Computing*, volume 13018, pages 81–93. Springer, 2021.

[3] Veysel Kocaman, Ofer M Shir, Thomas Bäck, and Ahmed Nabil Belbachir. Saliency can be all you need in contrastive self-supervised learning. In *International Symposium on Visual Computing*, pages 119–140. Springer, 2022.

1.5 Other Work by the Author

(Out of topic research activities over the course of PhD)

[4] Veysel Kocaman and David Talby. Biomedical named entity recognition at scale. In *International Conference on Pattern Recognition*, pages 635–646. Springer, Cham, 2021.

[5] Veysel Kocaman and David Talby. Improving clinical document understanding on covid-19 research with spark nlp. *AAAI-21 Workshop on Scientific Document Understanding*, (arXiv preprint arXiv:2012.04005), 2020.

[6] Veysel Kocaman and David Talby. Spark nlp: natural language understanding at scale. *Software Impacts*, 8:100058, 2021.

[7] Veysel Kocaman and David Talby. Accurate clinical and biomedical named entity recognition at scale. *Software Impacts*, 13:100373, 2022.

[8] Veysel Kocaman, Bunyamin Polat, Gursev Pirge, and David Talby. Biomedical named entity recognition in eight languages with zero code changes. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2022)*, volume 3202, 2022.

[9] Veysel Kocaman, Youssef Mellah, Hasham Haq, and David Talby. Automated de-identification of arabic medical records. In *Proceedings of ArabicNLP 2023*, pages 33–40, 2023.

[10] Veysel Kocaman, Hasham Ul Haq, and David Talby. Beyond accuracy: Automated de-identification of large real-world clinical text datasets. In *Machine Learning for Health (ML4H) 2023–Findings track*, 2023.

[11] Veysel Kocaman, Hasham Ul Haq, and David Talby. Beyond accuracy: Automated de-identification of large real-world clinical text datasets. In *Proceedings of The Professional Society for Health Economics and Outcomes Research (ISPOR) Europe 2023*, Copenhagen, Denmark, 2023. Poster presentation.

[12] Sutanay Choudhury, Khushbu Agarwal, Colby Ham, Pritam Mukherjee, Siyi Tang, Sindhu Tipirneni, Veysel Kocaman, Suzanne Tamang, Robert Rallo, and Chandan K Reddy. Tracking the evolution of covid-19 via temporal

comorbidity analysis from multi-modal data. In *AMIA*, 2021.

[13] Syed Raza Bashir, Shaina Raza, Veysel Kocaman, and Urooj Qamar. Clinical application of detecting covid-19 risks: A natural language processing approach. *Viruses*, 14(12):2761, 2022.

[14] Khushbu Agarwal, Sutanay Choudhury, Sindhu Tipirneni, Pritam Mukherjee, Colby Ham, Suzanne Tamang, Matthew Baker, Siyi Tang, Veysel Kocaman, Olivier Gevaert, et al. Preparing for the next pandemic via transfer learning from existing diseases with hierarchical multi-modal bert: a study on covid-19 outcome prediction. *Scientific Reports*, 12(1):1–13, 2022.

[15] Hasham Ul Hak, Veysel Kocaman, and David Talby. Deeper clinical document understanding using relation extraction. In *https://arxiv.org/abs/2112.13259*. Scientific Document Understanding workshop at AAAI 2022, 2021.

[16] Hasham Ul Hak, Veysel Kocaman, and David Talby. Mining adverse drug reactions from unstructured mediums at scale. In *W3PHIAI workshop at AAAI-22*. https://arxiv.org/abs/2201.01405, 2022.

[17] Murat Aydogan and Veysel Kocaman. Trsav1: A new benchmark dataset for classifying user reviews on turkish e-commerce websites. *Journal of Information Science*, 1:https–journals, 2022.

[18] Vikas Kumar, Lawrence Rasouliyan, Veysel Kocaman, and David Talby. Using natural language processing to identify adverse drug events of angiotensin converting enzyme inhibitors. International Forum on Quality and Safety in Healthcare EUROPE 2021, 2021.

[19] A. Emre Varol, Veysel Kocaman, Hasham Ul Hak, and David Talby. Understanding covid-19 news coverage using medical nlp. In *5th International Workshop on Narrative Extraction from Texts (Text2Story)*, 2022.

[20] Hasham Ul Haq, Veysel Kocaman, and David Talby. Connecting the dots in clinical document understanding with relation extraction at scale. *Software Impacts*, 12(100294), 2022.

[21] Vikas Kumar, Lawrence Rasouliyan, Veysel Kocaman, and David Talby. Detecting adverse drug events in dermatology through natural language processing of physician notes. In *36th International Conference on Pharmacoepidemiology & Therapeutic Risk Management*, volume 36, pages https–www, 2022.

1.5. Other Work by the Author

[22] Juan Martinez, Veysel Kocaman, Hasham Ul Haq, and David Talby. Zero-shot information extraction for clinical nlp. [under review].

[23] Arshaan Nazir, Thadaka Kalyan Chakravarthy, David Amore Cecchini, Rakshit Khajuria, Prikshit Sharma, Ali Tarik Mirik, David Talby, and Veysel Kocaman. Langtest: A comprehensive evaluation library for custom llm and nlp models. [under review].

[24] Julio Bonis, Veysel Kocaman, and David Talby. Social determinants of health in clinical narratives: A comprehensive review of pubmed clinical case reports from 1975 to 2022. Available at SSRN: https://ssrn.com/abstract=4590921 or http://dx.doi.org/10.2139/ssrn.4590921, 2022.