Week 5: A4. Fourth Research / Programming Assignment

Project Proposal: Text Classification Using CNN

This project explores the use of Convolutional Neural Networks (CNNs) for text classification, focusing on the AG's News Topic Classification dataset. CNNs, traditionally utilized for image processing, have shown promising results in natural language processing by effectively capturing local features and hierarchical patterns within text. This project aims to implement various CNN architectures, compare their performance against baseline fully connected networks, and evaluate their effectiveness in classifying news topics. The objective is to determine the potential of CNNs in improving text classification accuracy and efficiency. Previously, Long Short-Term Memory (LSTM) networks were utilized for this task in Assignment 3; now, we intend to investigate the capabilities of CNNs on the same dataset.

Text classification is a fundamental task in natural language processing (NLP), with applications ranging from spam detection and sentiment analysis to news categorization and topic modeling. The AG's News Topic Classification dataset, a widely recognized benchmark, is used to evaluate the performance of different text classification models. This project investigates the use of Convolutional Neural Networks (CNNs) for this task, comparing their performance to traditional fully connected networks.

In Assignment 3, we explored the use of Long Short-Term Memory (LSTM) networks for text classification. LSTMs are designed to handle sequential data, effectively capturing temporal dependencies within text. However, this project shifts focus to Convolutional Neural Networks (CNNs), which are leveraged for their capability to extract local features and patterns from text data. CNNs apply convolutional filters to the input text, enabling the model to learn n-gram features and hierarchical representations. This approach can potentially enhance the model's ability to classify text accurately by capturing essential information within the input sequences.

The primary goal of this project is to implement and evaluate various CNN architectures for text classification using the AG's News Topic Classification dataset. By conducting a series of experiments with different hyperparameters and architectural configurations, the project aims to identify the strengths and limitations of CNNs in handling text classification tasks. The experiments will include variations in the number of convolutional layers, filter sizes, pooling strategies, and activation functions. Additionally, the impact of different preprocessing techniques, such as tokenization methods and word embeddings, will be examined to optimize the input representation for CNNs.

Extensive exploratory data analysis (EDA) and preprocessing will be critical to the success of this project. The following suggestions will guide these preliminary steps:

1. Vocabulary Size: We will tweak the vocabulary size at least at three different levels to observe its impact on model performance. This will help us determine the optimal size that balances model complexity and classification accuracy.

2. Editing the Vocabulary: We will compare models trained on unedited vocabulary (including the most frequent words) against those trained on edited vocabulary (excluding common stopwords such as 'the,' 'a,' etc.). This will help assess the effect of removing high-frequency words on the model's ability to learn meaningful patterns.

CNNs have demonstrated their effectiveness in capturing spatial hierarchies in data, which can be analogously applied to textual data to capture semantic hierarchies. This project will explore the potential of CNNs to learn and generalize patterns that are crucial for distinguishing between different news topics. The comparative analysis with LSTM-based models will highlight the distinct advantages and limitations of each approach, providing a comprehensive understanding of their applicability in various NLP tasks.

The project will also investigate the scalability of CNN models for text classification. Given their parallelizable nature, CNNs can be more computationally efficient than sequential models like LSTMs, which could lead to faster training and inference times. This aspect will be particularly important for applications requiring real-time text classification, such as news aggregation services and social media monitoring.

By the end of this project, we expect to have a detailed evaluation of CNN architectures for text classification, highlighting the best-performing configurations and their practical implications. The findings will provide valuable insights into the design and deployment of CNN-based models for NLP applications, contributing to the broader field of artificial intelligence and deep learning. This comprehensive evaluation will guide future applications and optimizations of CNNs in NLP tasks, ensuring their effective utilization in diverse text classification scenarios.