Danmarks Tekniske Universitet

DTU

Deep Learning - Project Synopsis

Authors

Aditya Vijendra Girase - s240125
Mads Albert Alkjærsig - s203372
Ivor Baricevic - s232467
Samyak Jain - s232883

November 10, 2024

Intelligent Data Labeling via Neural Network-Driven Active Learning for Enhanced Model Efficiency

This project explores advanced neural network methods for intelligent data labeling, addressing the significant challenge of training models under limited data budgets. Neural networks typically require vast amounts of labeled data to perform accurately, making traditional data labeling methods costly and time-consuming, especially in fields like medical imaging and autonomous driving where expert labeling is essential.

The core innovation of this project lies in implementing active learning strategies, as discussed in *A Survey of Deep Active Learning*, that allow neural networks to selectively choose data points that are most valuable for labeling. By using active learning, the model can actively query labels for data points that maximize learning impact, thus improving performance while minimizing the labeling requirement. This project aims to create a model that balances **exploration** (selecting diverse samples for a broader understanding) and **exploitation** (focusing on highly uncertain samples to enhance decision boundaries), two key challenges identified in the paper.

The project milestones include:

1. **Foundational Analysis**: Investigate the relationship between the number of labeled data points and model accuracy on standard datasets like MNIST and CIFAR-10.

2. **Implementation of Active Learning Techniques**: Develop and implement uncertainty and margin-based methods to identify data points that will yield the most significant learning gain.

3. **Enhanced Data Selection Methods**: Explore advanced techniques, including:

    - **Generative Models**: Use generative networks to propose new, insightful data points for labeling.
    - **Clustering Approaches**: Identify samples near the decision boundary by clustering data points and selecting those that best refine model understanding.
    - **Bayesian Neural Networks**: Improve uncertainty estimates to enhance the reliability of the model's self-selection.

4. **Benchmark and Evaluation**: Compare active learning methods against traditional baselines in terms of labeling efficiency, model accuracy, and robustness to overconfidence issues.

This project aims to develop a flexible active learning framework that adapts dynamically to dataset characteristics, ultimately improving model accuracy with minimal labeled data. By optimizing the selection of data points for labeling, this approach has the potential to advance neural network training efficiency significantly, offering a valuable tool for applications where labeled data is scarce or costly to acquire.