Introduction to Machine Learning IKT466

Final Project

Project Overview

For the final project, you will apply **deep learning methods** to a real-world dataset. Your primary goal is to:

1. Implement and train at least one **deep neural network** (CNN, LSTM/GRU, or Transformer).

2. Explore how architecture choices, preprocessing, and training strategies can affect performance, including experiments with **adding** more layers, novel architectural designs, loss functions, and related techniques.

3. Reflect on model performance, limitations, and trade-offs.

You may also implement **traditional ML baselines** (e.g. Logistic Regression, SVM, Random Forest) if you want to compare - but this is **optional** and not a requirement.

Tracks and Datasets

You will select **one track** (Computer Vision, Text/NLP, Multimodal, or Audio) and **one dataset** from that track. For the order of the datasets I have tried to make it go from “easier” to “harder” - but that is not necessarily true.

I am open for **suggestions** if you have another dataset that is not listed here that you want to do, but it needs to be approved.

Track A: Computer Vision (CV)

Goal: Train a deep network to classify images.

Dataset Options:

1. **CIFAR-100** - 60,000 images, 100 fine-grained classes.
2. **Tiny ImageNet** - 100,000 images, 200 classes, 64×64 px.
3. **Oxford Pets** -
4. **Oxford Flowers** - natural objects, good for transfer learning.

Deep Learning Focus (suggestions):

- CNNs with convolution + pooling layers.
- Explore data augmentation, dropout, batch normalization.
- Transfer learning from pretrained models (ResNet, EfficientNet).

Optional Baseline: Logistic Regression, SVM, or Random Forest on raw/pixel-reduced features.

Track B: Text / NLP

Goal: Train a deep sequence or transformer-based model for text classification.

Dataset Options:

1. **AG News** - 120k news articles, 4-way classification.
2. **SST-5 (Stanford Sentiment Treebank)** - fine-grained 5-class sentiment (very neg → very pos).
3. **IMDb Sentiment** - 50k movie reviews, binary classification.

Deep Learning Focus (suggestions):

- CNN for text (1D convolutions).
- LSTM/GRU sequence models.
- Transformer fine-tuning (DistilBERT or similar).

Optional Baseline: Bag-of-Words or TF-IDF + Logistic Regression / Naive Bayes.

Track C: Multimodal (Text + Vision)

Goal: Build models that learn from both text and images.

Dataset Options:

1. **MM-IMDb** - movie posters + plot summaries → predict genres.
2. **Flickr30k** - 30k images with captions (captioning/retrieval).
3. **Amazon Multimodal Reviews** - product images + review text → predict ratings.

Deep Learning Focus (suggestions):

- Train unimodal models first (text-only, image-only).
- Build a fusion model (concatenate or jointly learn embeddings).
- Experiment with late vs early fusion strategies.

Optional Baseline: TF-IDF + Logistic Regression for text; SVM on CNN features for images.

Track D: Audio

Goal: Train a model to classify audio clips.

Dataset Options:

1. **ESC-50** - 2,000 labeled environmental sounds across 50 categories.
2. **UrbanSound8K** - 8,732 urban sound clips, 10 categories.
3. **Google Speech Commands** - spoken-word commands (e.g. “yes,” “no,” “stop”).

Deep Learning Focus (suggestions):

- Preprocess audio to spectrograms or MFCCs.
- Train CNNs on spectrogram images.
- Try LSTM/GRU models on temporal sequences.

Optional Baseline: Extract MFCCs + train Logistic Regression, Random Forest, or SVM.

Deliverables

- Dataset overview - size, properties, and challenges.
- Deep learning experiments - at least one neural model.
- Optional: classical ML baseline(s) for comparison.
- Results and Reflection - metrics (accuracy/F1), confusion matrices, training behavior.
- Discussion - what worked, what did not, and why.

Report Structure

- Introduction (dataset and task motivation)
- Background (relevant theory for you task and method)
- Methods (deep learning, optional baselines)
- Results (tables, plots, confusion matrices)
- Discussion (analysis, interpretability, limitations)
- Conclusion (summary of insights)

This should follow the template of a MSc thesis. The expected number of pages is ***approximately*** 20-30 pages.