Bachelor Thesis Proposal: Meta-Learning the Number of Clusters for Deep Clustering

Mamdouh Aljoud

February 2026

1 Introduction

Most deep clustering methods require the number of clusters k to be specified before training. In real applications, k is often unknown, and selecting it typically requires running many experiments for different values of k, which is computationally expensive.

This thesis studies whether the number of clusters can be predicted using meta-learning. The idea is to train a meta-model on many datasets so that, given a new dataset, it can estimate a suitable k using cheap dataset statistics and short "probe" training runs of a deep clustering model. These probe runs provide training-dynamics signals (e.g., loss curves) that may contain information about whether the chosen k is too small or too large.

2 Problem Statement

Can a meta-learner predict the correct number of clusters for a new dataset using dataset meta-features and training-dynamics signals from short deep clustering probe runs?

3 Tasks

- Conduct a literature review on methods for estimating the number of clusters and on deep clustering models that require a fixed k
- Design and implement a meta-feature extraction pipeline, including basic dataset statistics and optional probe features from short deep clustering runs (e.g., early loss curves, entropy of assignments, cluster balance)
- Train a meta-model to predict k and evaluate prediction accuracy across datasets

- Compare against simple baselines (e.g., fixed k, heuristics based on dataset size/dimension, or classical internal metrics if desired)

4 Requirements

- Basic knowledge of machine learning and deep learning
- Python programming skills (preferably PyTorch)
- Interest in unsupervised learning and empirical benchmarking

5 Contact

If you are interested, please send an email to `aljoud@dbs.ifi.lmu.de` with your CV and transcript, and briefly describe your interest in this topic.