Deep Learning Winter 2025–2026

Homework 4 — Representation Learning

Naaman Kopty naamankoptyta@gmail.com

Due: 28.01.2026

Question 1: (60 Points)

Programming assignment — See the attached Jupyter notebook on Moodle.

Question 3: Representation Geometry and Learning Objectives (20 Points)

Modern deep learning models often learn representations indirectly, through objectives that do not explicitly encode semantic structure. In this question, we explore how such representations arise and what geometric properties they exhibit.

(a) **Distributional representations.** Word embeddings are typically learned from co-occurrence statistics rather than explicit semantic supervision.

Explain why two words that never appear together in the corpus may nevertheless have similar embeddings.

(b) **Geometry of embedding spaces.** In practice, learned embedding spaces are often *anisotropic*, meaning that many embedding vectors concentrate in a narrow region of the space rather than being uniformly distributed.

Explain why anisotropy can be problematic for similarity-based reasoning, and describe one high-level approach for mitigating this issue.

(c) **Contrastive learning and representation structure.** Contrastive learning trains models by pulling together representations of positive pairs and pushing apart representations of negative pairs.

Explain how this objective shapes the geometry of the embedding space. In particular, discuss why contrastive learning encourages *relative* rather than absolute representations, and why such representations often transfer well to downstream tasks.

Question 4: Transformers, Vision Transformers, and Generalization (20 Points)

Transformers are flexible sequence models that rely on self-attention and minimal architectural inductive bias. In this question, we examine how these design choices affect representation learning and generalization.

(a) **Depth and abstraction in Transformers.** Transformer layers all share the same high-level structure, yet deeper Transformers often learn more abstract representations.

Explain how stacking multiple Transformer layers can lead to increasingly abstract features, even though each layer performs a similar type of computation.

(b) **Inductive bias: convolutional networks vs. Vision Transformers.** Vision Transformers process images as sequences of patches and do not explicitly encode locality or translation equivariance.

Explain what is meant by *inductive bias* in neural networks, and describe one important inductive bias present in convolutional networks that is largely absent in Vision Transformers. Then, discuss one consequence of this difference when training on relatively small datasets such as Flickr8k.

(c) **Self-supervised pretraining and data efficiency.** In this homework, self-supervised learning is used to pretrain the Vision Transformer before vision–language alignment.

Explain why self-supervised pretraining can improve generalization and data efficiency, even though no labels or captions are used during this stage. Relate your answer to the quality of representations learned during pretraining and their effect on zero-shot performance.

Good Luck!