Scalable Language Technologies Lab

Summer Term 2026

Exercise From Last Week

- Download these prompt logs and identify search sessions
- Create a dataset that contains search sessions from the AOL Query Log and the prompt logs
- Train a classifier on the joint dataset that predicts whether two queries are from the same search session

$$(\text{cute catd}, \text{cute cats}) \mapsto \text{Yes} \qquad (\text{cute cats}, \text{new york}) \mapsto \text{No}$$

- Encode the texts into vectors using DistilBERT (sentence embeddings)
- Implement a “fine-tuning” of DistilBERT: higher dot product means higher probability that two texts appear within the same search session
- Optional: You can use PyTorch Lightning for training
  PyTorch Lightning removes a lot of the boilerplate involved in Deep Learning. But you should write training code from scratch at least once to know how it works (its not hard; just tedious).
- Work on the assignment individually but exchange ideas and problems with each other
- Short 5–10 min presentations

Agenda

- Your presentations
- Hierarchical Embeddings

Don’t forget to hand in your presentations and code

Hierarchical Embeddings

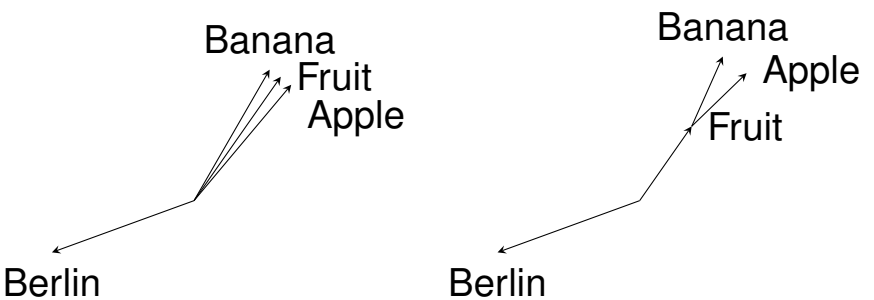
Order Embeddings [Vendrov et al. (2016)]

- We often need to represent text as numbers to apply deep learning methods

➜ Map words into a vector space (“embeddings”)

- Semantically similar words lie in similar directions from the origin
- But: No hierarchical information

➜ Order Embeddings: Distance from the origin signals specificity



Hierarchical Embeddings

Poincaré Embeddings [Nickel and Kiela (2017)]

- Problem:
  - Hierarchies are trees; node count grows exponentially as depth increases
  - The surface area of an n-dimensional sphere grows only polynomially with the radius
  - ➜ The embedding space does not have enough space
- Solution: do not use euclidean space but a hyperbolic manifold
  The Poincaré disk model is the prefered (but not the only) model of hyperbolic space
- Properties:
  - Riemannian manifold: Smooth manifold endowed with an inner product
    Gradient descent is possible; it just requires a manifold-aware optimizer
  - Conformal: Angles on the euclidean space are the same as on the manifold
- Dataset with hierarchical sentence information: HierarCaps [Alper et al. (2024)]
  Hierarchical image captions [tau-vailab.github.io]

Exercise

- Download and prepare WordNet and WordNet Mammals and HierarCaps
- Modify your embedding model from last week to learn a hierarchical sentence representation:
  - What should training samples look like?
  - What loss objective?
  - How can this later be applied to the query/prompt logs?
  - Discuss this within the group
- Encode the texts into vectors using DistilBERT (sentence embeddings) then map these into hyperbolic space
- Visualize results: draw a scatter plot (train 2D embeddings in the first place)
- Work on the assignment as a group but split it into sub-responsibilities
- Use this repo to develop your code together
- Short 5–10 min presentations
- Deadline in **two** weeks (June 01) with two online consultations (TBD)

Hints

- Use geoopt for the optimization:
  - Hyperbolic manifold: PoincareBall(c=1)
  - PoincareBall.expmap0 maps a point from euclidean space into the manifold
  - Manifold-aware optimizer: RiemannianAdam
  - Take the definition of a linear layer on the manifold from this example
- Use PyTorch Lightning for training
- PyTorch Lightning automatically logs various metrics. Use TensorBoard to interactively view them
  Alternatively you can have a look at WeightsAndBiases
- Use early stopping to avoid overfitting