Clustering-Oriented Representation Learning with Attractive-Repulsive Loss

Supplemental material

1 Preprocessing AGNews

The AGNews text corpus can be downloaded from `https://www.di.unipi.it/~gulli/AG_corpus_of_news_articles.html`. We preprocess the corpus as follows:

- Titles and descriptions are concatenated into a document D.
- All punctuation, other than dollar signs, is removed.
- We use the standard *spaCy* API (`https://spacy.io/`) to tokenize and then lemmatize every word in D.
- All stop words are removed.
- All numbers are replaced with tokens representing either MONEY if a dollar sign precedes it, a YEAR if the number is four digits and between 1700 and 2200, and otherwise just a NUMBER.

We then filter the topics in order to be of mostly equal size. It was necessary to augment the *Entertainment* category, as many samples marked as *Entertainment* were not clearly "entertainment", reading more like world news articles. We thus created a new entertainment category by selecting articles from all three of *Entertainment*, *Top News*, and *Top Stories* **if and only if** the article came from sources *E! Online*, *TheCelebrityCafe.com*, or any news source that had "entertainment" in its name (e.g., *New York Times Entertainment* or *Reuters Entertainment*). We thus obtained the following corpus statistics:

Topic	**Original**	**Filtered**	**Avg. # of words**
World	81,299	12,000	23.28
Sports	62,151	12,000	21.21
Entertainment	60,314	10,721	22.15
Business	54,432	12,000	22.00
Sci/Tech	41,184	12,000	20.88
Europe	30,889	12,000	19.81
Health	19,910	12,000	21.55
U.S.	13,758	12,000	24.17

Table 1: AGNews corpus statistics in terms of **Original** number of samples and the number of samples after being **Filtered**, along with the average number of words per sample in that topic.

We replace words that are OOV from the Glove 840B corpus with an UNK token and use the same random word embedding for them, with components normally distributed between -0.2 and 0.2. We use similarly distributed different random vectors to randomly initialize the number tokens (MONEY, YEAR, NUMBER) described above. For the purpose of large-scale training and testing, we do not fine-tune word embeddings. We found this to substantially decrease training speed and that doing so did not offer substantial improvements during preliminary testing.

2 λ Tuning

In Figure 1 we visualize each λ tuning across all six settings, as described in our paper.

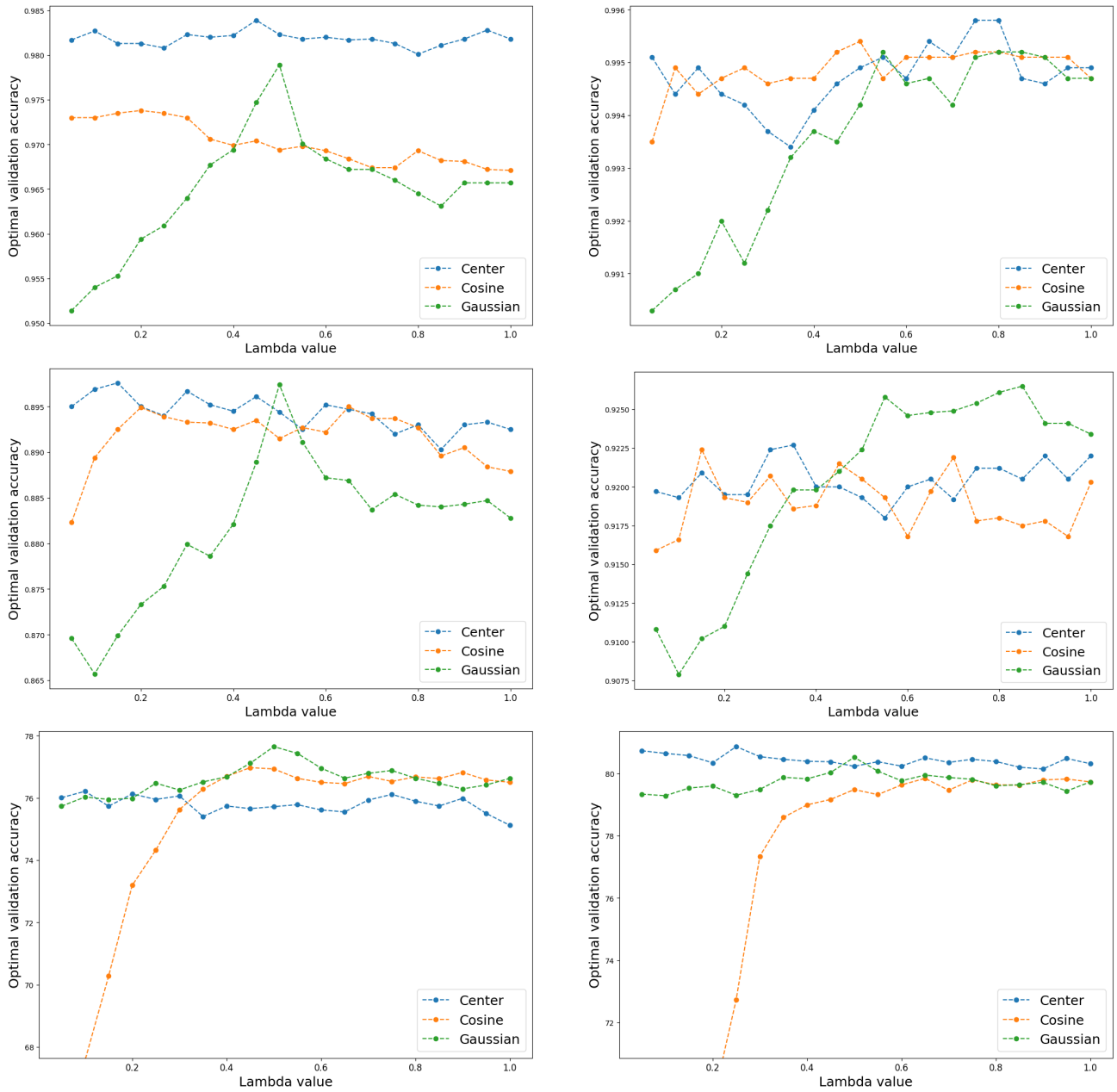

Figure 1: λ tuning results over all six settings. Left side is FFNNs, right side is CNNs; top is MNIST, middle is Fashion-MNIST, bottom is AGNews.