Clusterable Latent Spaces in Neural Networks

Kian Kenyon-Dean, Dr. Jackie Chi Kit Cheung, Dr. Doina Precup
McGill University, Department of Computer Science

Clustering-Oriented Representation Learning

Attractive & repulsive loss signals

$$\mathcal{L}_{COREL} = \sum_{i=1}^{N} -\mathcal{L}^{s}_{attract}(\mathbf{h}^{(i)}, \mathbf{w}_{y_i})\lambda + \mathcal{L}^{s}_{repulse}(\mathbf{h}^{(i)}, \mathbf{W})(1-\lambda)$$

Balance attraction vs. repulsion

- Attractive and repulsive loss defined with **similarity function `s(h, w)`**
  - *Attractive* seeks to **maximize similarity** between latent representations (final hidden layer activations, **h**) and output weight vectors (**w**) for their classes
  - *Repulsive* seeks to **minimize similarity** between representations and the output weights of other classes

- We propose losses based on two different similarity functions:
  - **Cosine-similarity** loss [-1, 1]:

$$s_{cos}(\mathbf{h}, \mathbf{w}) = \frac{\mathbf{h}}{\|\mathbf{h}\|} \cdot \frac{\mathbf{w}}{\|\mathbf{w}\|} \qquad \mathcal{L}^{cos}_{attract} = s_{cos}(\mathbf{h}^{(i)}, \mathbf{w}_{y_i}) \qquad \mathcal{L}^{cos}_{repulse} = \max_{k \neq y_i} s_{cos}(\mathbf{h}^{(i)}, \mathbf{w}_{y_i})^2$$

*Minimize distance to most similar class the sample does not belong to, square to make vectors **orthogonal** (and not 180 degress)*

  - **Gaussian-similarity** loss [0, 1]:

$$s_{gau}(\mathbf{h}, \mathbf{w}) = e^{-\gamma\|\mathbf{h}-\mathbf{w}\|^2} \qquad \mathcal{L}^{gau}_{attract} = \log s_{gau}(\mathbf{h}^{(i)}, \mathbf{w}_{y_i}) \qquad \mathcal{L}^{gau}_{repulse} = \log \sum_{k} s_{gau}(\mathbf{h}^{(i)}, \mathbf{w}_k)$$

Using the intuitions of CCE and softmax, define the objective as maximizing a univariate Gaussian PDF instead of a dot product

- With *any similarity function*, class prediction is:

$$\text{prediction} = \textbf{argmax}_k \, \mathbf{s}(\mathbf{h}, \mathbf{w}_k)$$

CONTRIBUTIONS

- We propose **Clustering-Oriented Representation Learning** (COREL) as a general framework for designing loss functions in neural networks for classification tasks. *Essential components* are:
  1. **Attractive** & **repulsive** loss signals
  2. **Similarity function** between representations and weights
- We redefine categorical cross-entropy (CCE) as a specific case of COREL, and propose **two new loss functions** in our framework, which are *better than CCE* in terms of *clusterability*.

Reinterpreting Categorical Cross-Entropy

- Categorical cross-entropy is defined as:

$$\mathcal{L}_{CCE} = -\sum_{i=1}^{N} \log \frac{\exp(\mathbf{w}_{y_i}^T \mathbf{h}^{(i)})}{\sum_{k=1}^{K} \exp(\mathbf{w}_k^T \mathbf{h}^{(i)})}$$

- COREL generalization redefines it generally with **s**, similarity function:

$$\mathcal{L}_{CCE} = \sum_{i=1}^{N} -s(\mathbf{h}^{(i)}, \mathbf{w}_{y_i}) + \log \sum_{k=1}^{K} e^{s(\mathbf{h}^{(i)}, \mathbf{w}_k)}$$

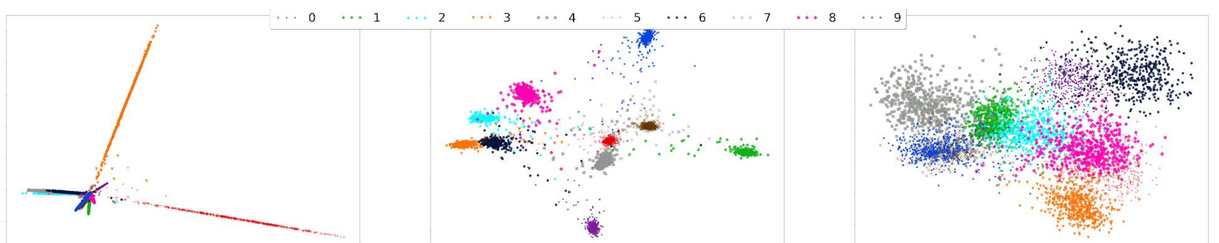
Experiments

- **Objectives**:
  a. Determine if the COREL losses can perform **as well as** CCE
  b. Analyze the ***clusterability*** of the ***latent spaces***
  c. **Qualitatively** analyze the latent representions
- Test on MNIST, **Fashion-MNIST**, and AgNews datasets for classification performance (only Fashion-MNIST results presented here from CNN)
- Test *unsupervised* ***clustering algorithms*** (**K-Means**, Gaussian mixture models) on test set **representations**, determining intrinsic expressivity

*Below we look at **smallest** to **largest norm** for **cosine-COREL** representations.*

Fashion-MNIST Test set results	**Prediction acc.** *(sup.)*	**K-Means acc.** *(unsup.)*	**K-Means density** *(unsup.)*
CCE	0.9124	0.7246	0.2808
Cosine COREL *(lambda = 0.15)*	0.9092	0.9072	**0.8473**
Gaussian COREL *(lambda = 0.85)*	**0.9164**	**0.9127**	0.7382

MNIST Feature Visualization *(PCA-compressed from 128 dimensions)*



Contact Information

Kian Kenyon-Dean. M.Sc. Computer Science, McGill

kian.kenyon-dean@mail.mcgill.ca

https://kiankd.github.io/