CMSC 25300 / CMSC 35300 / STAT 27700 Final Project Proposal (35300)

Yeol Ye

TOTAL POINTS

5 / 5

QUESTION 1

1 Proposal 5 / 5

✓ **- 0 pts** Correct

- 5 pts No submission.

Sounds great.

CMSC 35300: Final Project Proposal

Shawn Shan, Ziyu Ye
Department of Computer Science
University of Chicago
shansixiong, ziyuye@uchicago.edu

1 Introduction and Literature Review

Diagnosis for COVID-19 is an important task recently. However, a lot of critical information used for such diagnosis often consists of high-dimensional data (*e.g.* chest X-ray images, CT scans for Lungs, etc.), which brings about challenges on understanding the underlying patterns to make diagnosis. An common technique is to learn a low-dimensional representation of the data, and use such representation to make predictions.

Current literature has shown that such dimension reduction (or feature compression) techniques help to achieve high accuracy on downstream tasks [4, 3]. Traditional work have proposed different technique for dimension reduction, for examples, principal component analysis (PCA), linear discriminant analysis (LDA), and matrix factorization.

Recent work has shown the promise of using autoencoders (AEs) to reduce data dimension to obtain a more compact representation of data. AEs are known to be able to leverage the large amount of unlabeled data to map high dimension data to embedding vectors. AEs achieve promising performance in several aspect, however, one of the biggest challenge it faces it that the learned features are often not the optimal representation, or sensitive to small input variations. A recent study [1] shows that it is possible to encourage linear AEs to learn the optimal representation (*i.e.* ordered, axis-aligned principal components) by simple regularization.

2 Problem Statement

As mentioned above, we would like to make our own improvement on [1], in order to learn an optimal representation which is more effect and relevant to downstream tasks (*e.g.* COVID-19 diagnosis). Possible directions include encouraging a more discrete latent representation, incorporating label information while training AEs, *etc.*

3 Dataset Description

To evaluation our technique, one major dataset we consider on COVID classification using chest X-ray data [2]. The dataset consists 654 chest-xray images. We plan to use a separate larger scale chest-xray dataset to train a AE and use the AE as a dimension reduction tool for the COVID data. We plan to look at the classification performance comparing directly train a classifier on the chest-xray images, and train a classifier using the embedding.

Besides that, we also consider other datasets including CT scans for Lungs [6, 5].

References

[1] Xuchan Bao, James Lucas, Sushant Sachdeva, and Roger Grosse. Regularized linear autoencoders recover the principal components, eventually. 2020.

[2] Joseph Paul Cohen, Paul Morrison, and Lan Dao. Covid-19 image data collection. *arXiv 2003.11597*, 2020.

[3] Yangqin Feng, Lei Zhang, and Juan Mo. Deep manifold preserving autoencoder for classifying breast cancer histopathological images. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 2018.

[4] Najibesadat Sadati, Milad Zafar Nezhad, Ratna Babu Chinnam, and Dongxiao Zhu. Representation learning with autoencoders for electronic health records: A comparative study. *arXiv preprint arXiv:1908.09174*, 2019.

[5] Hinrich B. Winther, Hans Laser, Svetlana Gerbel, Sabine K. Maschke, Jan B. Hinrichs, Jens Vogel-Claussen, Frank K. Wacker, Marius M. Höper, and Bernhard C. Meyer. A large image dataset for covid-19. 2020.

[6] Jinyu Zhao, Yichen Zhang, Xuehai He, and Pengtao Xie. Covid-ct-dataset: a ct scan dataset about covid-19. *arXiv preprint arXiv:2003.13865*, 2020.

1 Proposal 5 / 5

✓ **- 0 pts** **Correct**

- 5 pts No submission.

Sounds great.