Critique Report on the Paper: "Deep Spectral Clustering using Dual Autoencoder Network [1]"

Sumaiya Tabassum Nimi

Background Information and Purpose of the Research

Unsupervised clustering of images is a well-studied problem in Computer Vision research. Unlike the traditional clustering algorithms, deep clustering methods find optimal latent feature space more discriminative than the original input space, prior to clustering and hence leading to better clusterization. In the paper [1], a novel deep clustering method was proposed that employed optimal latent feature space learned using a dual autoencoder.

Methods Proposed

Th proposed method for deep clustering involved the following two steps.

- **Finding Optimal Discerning Feature Subspace:** Dual autoencoder network was at first pretrained to obtain the latent features. Then the decoder was fine-tuned with a proposed adversarial-like loss function for reconstruction, called L_r that enabled learning in a noisy environment. Also, the encoder was fine-tuned with a loss function, called L_e that learned to maximize mutual information between input samples and their reconstructions.
- **Spectral Clustering of Feature Subspace:** Then the obtained feature subspace was clustered using Spectral Clustering, that minimized loss called L_c.

The entire proposed network architecture was trained to minimize loss $L = L_r + L_e + L_c$, in an end-to-end manner, leading to optimal clustering.

Strength of the Work

- Loss functions proposed to train each module of the proposed framework were mathematically rigorous and backed with strong theoretical justifications.
- Reported experimental results clearly demonstrated that the proposed clustering method was more successful compared to the previous approaches on the five datasets called MNIST-*full*, MNIST-*test*, USPS, Fashion-10 and YTF.
- Ablation studies to justify the choices of different learning rules, data transformations and hyperparameters were done and duly reported.

Limitations of the Work

- The proposed approach was tested mostly on monochromatic images, that are considered comparatively easier to learn. Experimental evaluations should have included results on datasets like CIFAR-10, CIFAR-100 and Imagenet, in order to establish the worth of the proposed approach. Its

often the case that some approaches, that are suited to simpler datasets often fail miserably when presented with more complex data. There should have been more experiments conducted to justify that this was not the case with the proposed approach.

- The dimension of the input image, associated with each dataset on which the proposed approach was tested, was quite small. This was a factor that led to success of techniques like minimizing KL divergence or maximizing mutual information. When the dimension of the image becomes large, like $224 \times 224 \times 3$ for Imagenet, the feature space becomes too large to learn with such loss functions and thus leading to worse than expected performance.

- Also, if the image dimension became larger, it seemed that the proposed framework would have become proportionately memory inefficient and hence it would have become eventually harder to train the setup on regular hardware configurations.

- The proposed framework would be difficult to deploy on edge devices, being so heavy on both computation and memory.

Questions Unanswered

- How sensitive will the clustering be to an out-of-distribution sample? If, say, the English letter I was input to a model trained using the proposed approach on MNIST dataset, will the model be able to tell this letter apart from the cluster of digit 1? How closer will this be to the aforementioned cluster?

- The hardware configurations required for training the proposed framework should have been described in the paper. The proposed framework looked big enough and the proposed learning rules seemed complex enough to train on common desktop configurations. Hence, the required configurations should have been explicitly discussed.

Suggested Future Studies

- Instead of pretraining an autoencoder, features could be collected from readily available pretrained deep learning models for image classification. That way, the training could be made faster and easier, with possibly better end results.

- Adversarial samples corresponding to the positive samples could be used as negative samples for training the decoder. This could lead to more robust training.

- Also, a more adversarial learning setup could be designed by using Generative Adversarial Learning for training to generate negative samples and this training could be associated with clustering, leading to even more robust learning of the feature space.

- Also, the proposed model could be made robust to Out-of-Distribution and adversarial input samples through an externally incorporated detector module, since it looked like the proposed learning did not render the resulting model inherently capable of such detection.

References

[1] Xu Yang, Cheng Deng, Feng Zheng, Junchi Yan, and Wei Liu. Deep spectral clustering using dual autoencoder network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4066–4075, 2019.