Question 2 : Paper Review

Title : Nonlinear Non-negative Matrix Factorization using Deep Learning
Authors : Hui Zhang, Huaping Liu, Rui Song, Fuchun Sun
Published in: [2016 International Joint Conference on Neural Networks (IJCNN)]
Link : https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=7727237

1) **Why Nonlinear NMF**

The NMF method has been used extensively in domains like signal processing, machine learning and text mining because of the promising results. The results are good because the factorization exhibits many properties like sparsity and interpretability. But one major limitation of NMF is that it allows only linear combination of columns of W for estimation. This does not give good results for complex signals.

2) **New formulation**

In this paper, Deep Autoencoders are used to learn a latent space representation of input data points and then NMF is applied on the learnt domain. They propose a modified loss function and optimization algorithm for the same.
Quite intuitively , the new cost function is :

$$\min_{f,D,X} \boldsymbol{E}_{N-NMF} = \|\boldsymbol{f}(\boldsymbol{V}) - \boldsymbol{DX}\|_F^2 \quad s.t. \boldsymbol{f}(\boldsymbol{V}), \boldsymbol{D}, \boldsymbol{X} \geq 0$$

Where **f** represents the nonlinear mapping to a low dimensional space which is deep network in our case, **D** and **X** represent the factorization. The paper uses **deep autoencoders** for the non linear mapping.

3) **Why deep autoencoders only ?**

We could have just used a normal CNN to go from high dimensional space to a low dimensional one. But we need to ensure that the learnt embedding preserves the properties of original data and there structure in space. Hence deep autoencoders are used because the decoder network learns the original samples from the transformed latent domain and hence ensures the data structure in maintained in the latent space.
Reconstruction Error:

$$\boldsymbol{Er}^r = -\sum_i v_i^r \log \hat{v}_i^r - \sum_i (1 - v_i^r) \log(1 - \hat{v}_i^r)$$

Where **υ** is the original data point and v(cap) the corresponding reconstructed one.

4) **Optimization**

Final loss function for deep network :

$$J(\boldsymbol{W}, \boldsymbol{B}, \boldsymbol{D}, \boldsymbol{X}) = \boldsymbol{E}_{N-NMF} + \lambda \sum_{r=1}^{m} \boldsymbol{E}\boldsymbol{r}^{r}$$

A **2** step approach is used, first the deep network is optimized by standard back propagation assuming **D** and **X** as constant and then Multiplicative Update (the same as taught in class) is used to update **D** and **X** assuming **W** and **B** as constant.
This whole process is repeated 100 (variable) times.

5) **Results on Clustering of Images**

Metric for clustering accuracy = fraction of samples clustered correctly
In the paper they have shown extensive comparisons over several datasets. I have included image of one. N-NMF and N-GNMF are the proposed methods.