Review of "Towards a comprehensive visualization of structure in large scale data sets"

Jiaxin Hu

March 3, 2022

This work proposes a parallelized Barnes-Hut t-Stochastic neighbouring Embedding (t-SNE), named pt-SNE, for high-dimensional data visualization. The main contribution of pt-SNE is two-folded: (1) breaking the computational limitations of t-SNE under the large neighbourhood setting with large perplexity (ppx); (2) simplifying the parameterization setup in t-SNE. Here are the comments.

Comments:

1. In line 256, page 10, high sensitivity to ppx is one of the main features of pt-SNE, which may help to reveal the higher level structure in data. However, the high sensitivity of pt-SNE may need extra computational resources and time to find a desired visualization with proper ppx while previous t-SNE methods (e.g. FIt-SNE) are robust to ppx. Also, the chunk&mix protocol introduces extra parameters for parallelization. Therefore, it is unclear whether the tuning process of pt-SNE is indeed simplified compared to other methods, especially when the local structure with small ppx is of interest.

2. In line 209, page 9, authors point out that the ppxs in pt-SNE and FIt-SNE are not equivalent due to the chunk&mix protocol. A larger ppx should be applied for pt-SNE to achieve similar visualization as FIt-SNE. However, authors do not provide a way to judge whether two methods are at the same "perplexity level" with different ppx. Therefore, the fairness of the comparisons between FIt-SNE and pt-SNE in Figure 4 (e) (f) and Figure 6 (e) (f) are suspicious. Because the performances in two methods may not be comparable with the same ppx, and thus the comparisons of accuracy and time may be unfair.

3. Authors implement the experimental comparisons (Figures 3, 4, 5, 6) merely between FIt-SNE and pt-SNE. More discussions or experimental comparisons with MultiCore t-SNE (Ulyanov, 2016) and UMAP (McInnes et al., 2018) should be added. Specifically, MultiCore t-SNE is a multicore modification of Barnes-Hut t-SNE, and MultiCore t-SNE should be the most competitive method with pt-SNE in terms of computation. The UMAP is believed to preserve the global structure better than t-SNE, and it would be interesting to see the comparison between the global structures found by UMAP and pt-SNE.

4. Typo: Should there have an expectation over i for the second term $KL(p_{i|\cdot}, q_{i|\cdot})$ in the right hand side of equation (1)?

Given these limitations, I would be hesitant to decide that this paper is suitable for the venue.

References

McInnes, L., Healy, J., and Melville, J. (2018). Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*.

Ulyanov, D. (2016). Multicore-tsne. https://github.com/DmitryUlyanov/Multicore-TSNE.