**1. t-distributed Stochastic Neighbour Embedding (t-SNE)**

- *feature extraction*

t-SNE is non-linear dimensionality reduction technique which is typically used to visualize high dimensional datasets. Some of the main applications of t-SNE are Natural Language Processing (NLP), speech processing, etc…

t-SNE works by minimizing the divergence between a distribution constituted by the pairwise probability similarities of the input features in the original high dimensional space and its equivalent in the reduced low dimensional space. t-SNE makes then use of the **Kullback-Leiber (KL)** divergence in order to measure the dissimilarity of the two different distributions. The KL divergence is then minimized using gradient descent.

When using t-SNE, the higher dimensional space is modelled using a Gaussian Distribution, while the lower-dimensional space is modelled using a Student’s t-distribution. This is done, in order to avoid an imbalance in the neighbouring points distance distribution caused by the translation into a lower-dimensional space.

**2. Why using 2 layers?**

**3. Why SoftMax?**

Sigmoid’s probabilities produced by a Sigmoid are independent. Furthermore, they are *not*constrained to sum to one: 0.37 + 0.77 + 0.48 + 0.91 = 2.53. The reason for this is because the Sigmoid looks at each raw output value separately. Whereas Softmax’s the outputs are interrelated. The Softmax probabilities will always sum to one by design: 0.04 + 0.21 + 0.05 + 0.70 = 1.00. In this case, if we want to increase the likelihood of one class, the other has to decrease by an equal amount.

**4.**