FROM t-SNE TO UMAP WITH CONTRASTIVE LEARNING

Damrich, S; Bohm, J. K.; Hamprecht, F. A.; Kobak, D.

Heidelberg University

Sep 11, 2024

Presented by Scott Sun from Duke B&B

ChatGPT artwork

prompt: "generate a picture about the process of projecting 3D balls to 2D space, where balls are unequally spaced out"

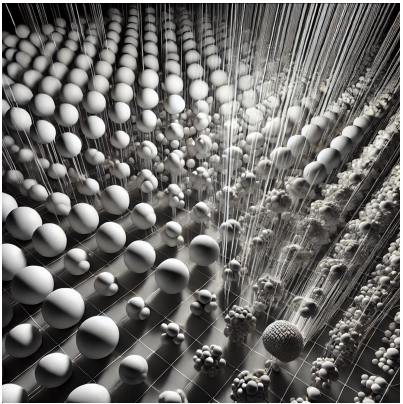


Figure: ChatGPT art

Introduction

The authors investigate the connection among various neighbor embedding (NE) methods for dimension reduction under the framework of contrastive learning.

Goal: explain the mathematical underpinnings of the relationship between t-SNE (t-distributed stochastic neighbor embedding) and UMAP (uniform manifold approximation and projection)

- both t-SNE (NCE) and UMAP (NEG) can produce insightful embeddings, but the distinctions between their results and relationship between their loss functions still remain unclear
- further explore connection between NCE & NEG, which has been discussed since the introduction of skip-gram for word2vec (Mikolov et. al., 2013)
- address the question of how to leverage contrastive methods to compute neighbor embeddings
- explain numerical stability problem of vanilla UMAP

Background: Noise-contrastive estimation (NCE)

Assume we want to fit pdf q_θ, which should takes form $q_\theta(x) = \frac{1}{Z_\theta} \exp(-E_\theta(x))$

In a regular MLE approach, we have to calculate the computationally expensive partition function Z_θ. As an alternative, NCE circumvent the problem by converting density estimation to supervised prediction.

Let observed real data $s \sim p$ with sample size N and random noise $t \sim \xi$ with sample size mN. Then, θ can be estimated by minimize the loss

$$\mathcal{L}^{\text{NCE}} = -\mathbb{E}_{s\sim p} \log \left(\underbrace{\frac{q_\theta(s)}{q_\theta(s) + m\xi(s)}}_{\mathbb{P}(data|s)} \right) - m\mathbb{E}_{t\sim\xi} \log \left(1 - \frac{q_\theta(t)}{q_\theta(t) + m\xi(t)} \right)$$

Equiv. we can solve θ by fitting a logistic regression with

- input: $\log \frac{q_\theta(x_i)}{m\xi(x_i)}, x = (s, t)^\top$
- output: label for being s or t
- fix slope and intercept at (1, 0) but optimize for θ

Background: Noise-contrastive estimation (NCE)

In the previous density estimate example, we approximate p with q_θ. In NCE, we also want to introduce a learnable normalizing parameter Z s.t.

$$q_{\theta,Z} = \frac{q_\theta}{Z}$$

Therefore,

$$\mathcal{L}_Z^{\text{NCE}} = -\mathbb{E}_{s\sim p} \log\left(\frac{q_\theta(s)}{q_\theta(s) + Zm\xi(s)}\right) - m\mathbb{E}_{t\sim\xi} \log\left(1 - \frac{q_\theta(t)}{q_\theta(t) + Zm\xi(t)}\right)$$

Note that, if there is some q_θ able to match p we can also drop Z.

Background: Negative sampling (NEG)

Original, NEG is typically used to train skip-gram model for word2vec. Consider a target (I) and context (O) word pair (w_I, w_O)

$$
\begin{aligned}
l_w &= -\log \sigma(v_{w_I}^\top v_{w_O}) + \sum_{\tilde{w} \sim P(w)} \log \sigma(-v_{w_I}^\top v_{\tilde{w}}) \\
&= -\log \frac{\exp(v_{w_I}^\top v_{w_O})}{\exp(v_{w_I}^\top v_{w_O}) + 1} + \sum_{\tilde{w} \sim P(w)} \log \frac{1}{\exp(v_{w_I}^\top v_{w_O}) + 1}
\end{aligned}
$$

Background: Neighbor embedding (NE)

NE methods are referring to some of the commonly used dimension reduction techniques. They define a notion of similarity over pairs of input data points. The similarity encodes the **neighborhood structure** and informs the **low-dimensional** embedding.

As a result, we would see clusters in the t-SNE and UMAP plots after dim reduction.

Regular t-SNE in a nut-shell:

1. obtain a matrix of (scaled) similarity scores between data points in the original space, calculated using normal distribution with μ being the reference points in pairs
2. obtain a matrix of (scaled) similarity scores between data points in a dimension-reduced space, calculated using t-distribution; make the t-distribution-based sim mat approach the one calculated in the original space

Background: Neighbor embedding (NE)

Consider a much simpler similarity score in the original high-dimensional space. Define a binary symmetric k-nearest-neighbor (skNN) graph, where k is a hyperparameter, s.t. the similarity score in the original space is given by

$$p(ij) = \frac{1}{|\text{skNN}|} \mathbb{1}_{\{ij \in \text{skNN}\}}$$

The low-dimensional similarity measure depends on some distance-based kernel $\phi(d_{ij}), d_{ij} = \|e_i - e_j\|$, where the embeddings depend on θ. In this paper, they pick the Cauchy kernel s.t. $\phi(ij) := \phi(d_{ij}) = \frac{1}{d_{ij}^2+1}$. Then, a normalized model (with partition function) is

$$q_\theta(ij) = \frac{\phi(ij)}{Z}, Z = \sum_{ij} \phi(ij)$$

Then, a **t-SNE** model is optimized by cross-entropy (equiv. to MLE)

$$\mathcal{L}^{\text{t-SNE}}(\theta) = -\mathbb{E}_{ij \sim p} \log\left(q_\theta(ij)\right)$$

Background: Neighbor embedding (NE)

NC-t-SNE uses NCE loss to fit t-SNE with learnable partition function

$$\begin{aligned}\mathcal{L}^{\text{NC-t-SNE}}(\theta, Z) &= -\mathbb{E}_{ij\sim p}\log\left(\frac{q_{\theta,Z}(ij)}{q_{\theta,Z}(ij)+m\xi(ij)}\right) - m\mathbb{E}_{ij\sim\xi}\log\left(1-\frac{q_{\theta,Z}(ij)}{q_{\theta,Z}(ij)+m\xi(ij)}\right)\\ &= -\mathbb{E}_{ij\sim p}\log\left(\frac{q_\theta(ij)}{q_\theta(ij)+Zm\xi(ij)}\right) - m\mathbb{E}_{ij\sim\xi}\log\left(1-\frac{q_\theta(ij)}{q_\theta(ij)+Zm\xi(ij)}\right)\end{aligned}$$

NEG-t-SNE uses NEG loss to fit t-SNE without learnable partition function

$$\mathcal{L}^{\text{NEG-t-SNE}}(\theta) = -\mathbb{E}_{ij\sim p}\log\left(\frac{q_\theta(ij)}{q_\theta(ij)+1}\right) - m\mathbb{E}_{ij\sim\xi}\log\left(1-\frac{q_\theta(ij)}{q_\theta(ij)+1}\right)$$

UMAP is fitted using the following loss (which can be shown as a instance of NEG)

$$\mathcal{L}^{\text{UMAP}}(\theta) = -\mathbb{E}_{ij\sim p}\log\left(q_\theta(ij)\right) - m\mathbb{E}_{ij\sim\xi}\log\left(1-q_\theta(ij)\right)$$

Method: Unification

Generalized t-SNE with plugged-in $\bar{Z}$, and let $|X|$ bet the size of the sample space i.e. $\binom{n}{2}$.

$$\mathcal{L}_{\bar{Z}}(\theta) = -\mathbb{E}_{ij\sim p} \log \left(\frac{q_\theta(ij)}{q_\theta(ij) + \bar{Z} m \xi(ij)} \right) - m\mathbb{E}_{ij\sim \xi} \log \left(1 - \frac{q_\theta(ij)}{1 + \bar{Z} m \xi(ij)} \right)$$

with default $\bar{Z} = \frac{|X|}{m}$ s.t. we get NEG-t-SNE.

Relating NEG to NCE
If $\bar{Z} = Z^{\text{NC-t-SNE}}$, learned from NC-t-SNE, NC-t-SNE = NEG-t-SNE.

Relating UMAP to NEG-t-SNE
Let parametric model $\tilde{p}_\theta(ij)$ denote this kernel. Then, Cauchy kernel $q_\theta(ij) = \frac{1}{1+d_{ij}^2} = \frac{\tilde{p}_\theta(ij)}{\tilde{p}_\theta(ij)+1}$, and UMAP $\approx$ default NEG-t-SNE with kernel $\tilde{q}_\theta$.

Experiment: MNIST & NEG-t-SNE spectrum

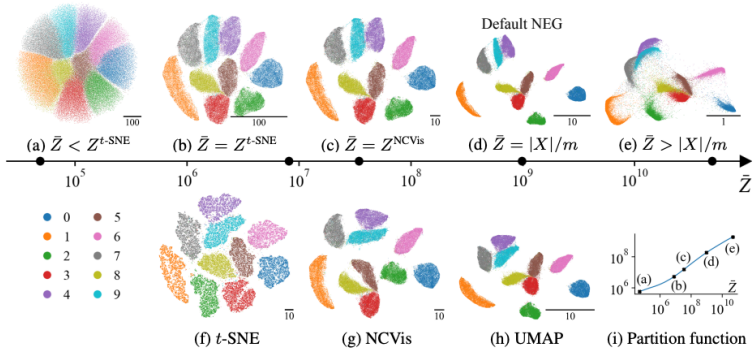
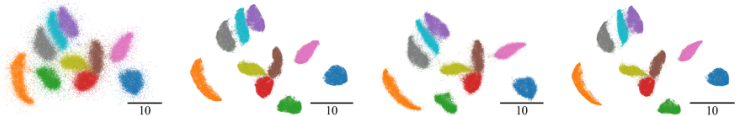


Figure 1: **(a – e)** Neg-t-SNE embedding spectrum of the MNIST dataset for various values of the fixed normalization constant $\bar{Z}$, see Sec. 5. As $\bar{Z}$ increases, the scale of the embedding decreases, clusters become more compact and separated before eventually starting to merge. The Neg-t-SNE spectrum produces embeddings very similar to those of **(f)** t-SNE, **(g)** NCVis, and **(h)** UMAP, when $\bar{Z}$ equals the partition function of t-SNE, the learned normalization parameter Z of NCVis, or $|X|/m = \binom{n}{2}/m$ used by UMAP, as predicted in Sec. 4–6. **(i)** The partition function $\sum_{ij}(1+d_{ij}^2)^{-1}$ tries to match $\bar{Z}$ and grows with it. Here, we initialized all Neg-t-SNE runs using $\bar{Z} = |X|/m$; without this 'early exaggeration', low values of $\bar{Z}$ yield fragmented clusters (Fig. S11).

Experiment: MNIST & NEG-t-SNE spectrum



(a) UMAP, no annealing (b) UMAP with annealing (c) Neg-t-SNE, no ann. (d) Neg-t-SNE with ann.

Figure 2: Embeddings of the MNIST dataset with UMAP and Neg-t-SNE with and without learning rate annealing in our implementation. UMAP does not work well without annealing because it implicitly uses the diverging $1/d_{ij}^2$ kernel in NEG, while Neg-t-SNE uses the more numerically stable Cauchy kernel (Sec. 6). UMAP's reference implementation also requires annealing, see Figs. S1a, d.

NE	attractive term	repulsive term
UMAP	$\log(1 + d_{ij}^2)$	$\log \frac{1+d_{ij}^2}{d_{ij}^2}$
NEG-t-SNE	$\log(2 + d_{ij}^2)$	$\log \frac{2+d_{ij}^2}{1+d_{ij}^2}$

Table: Comparing UMAP & NEG-t-SNE

Results

- This paper provides a unified framework that puts NE methods such as t-SNE and UMAP together through a contrastive learning perspective. The numerical instability of UMAP arises from the repulsive term between pairs of elements that are too close to each other.
- Combined with InfoNCE, t-SNE can be used as a self-supervised learning method. There are potentially some relationship between contrastive NE and SimCLR.

Recommendation

Is it worth reading? Yes.

- the math/framework is clean and elegant; the interpretation is straightforward
- the paper gives clear illustration about the connections between NCE and NEG and relationships between different NE methods

Is it worth implementing? Yes.

- they have a public github repo
- looks interesting to see how these methods can be combined with some energy-based models and generate images/texts through MCMC/Langevin dynamics