Do’s and Don’ts of using t-SNE to Understand Vision Models

Laurens van der Maaten

Interpretable Machine Learning for Computer Vision Workshop
June 18th, 2018
Tutorial goals

• How does t-SNE work?
Tutorial goals

• How does t-SNE work?

• What kind of visualizations can I create using t-SNE?
Tutorial goals

• How does t-SNE work?

• What kind of visualizations can I create using t-SNE?

• What should I be careful about when using t-SNE?
Introduction to t-SNE
Introduction

• We are given a set of high-dimensional data points $x_1, \ldots, x_N$. 
Introduction

- We are given a set of high-dimensional data points $\mathbf{x}_1, \ldots, \mathbf{x}_N$

- We want to know what this data “looks like”? 
Introduction

• We are given a set of high-dimensional data points \( x_1, \ldots, x_N \)

• We want to know what this data “looks like”?

• To do this, we build a “map” of this data in a two- or three-dimensional space
We are given a set of high-dimensional data points $x_1, \ldots, x_N$.

We want to know what this data “looks like”?

To do this, we build a “map” of this data in a two- or three-dimensional space.
PCA

- We are given a set of high-dimensional data points $x_1, \ldots, x_N$

- We want to know what this data “looks like”?

- To do this, we build a “map” of this data in a two- or three-dimensional space
PCA

- PCA learns a linear mapping, which is very restrictive
PCA

- PCA learns a linear mapping, which is very restrictive
- PCA focuses on preserving large pairwise distances
t-SNE

- Compute pairwise similarities between data with normalized Gaussian kernel

\[ p_{ij} = \frac{\exp\left(-\frac{||x_i - x_j||^2}{2\sigma^2}\right)}{\sum_k \sum_{l \neq k} \exp\left(-\frac{||x_k - x_l||^2}{2\sigma^2}\right)} \]
Low-D

- Compute pairwise similarities between data with normalized Gaussian kernel
- Measure normalized Student-t similarities in the t-SNE map

\[
q_{ij} = \frac{(1 + \|\mathbf{y}_i - \mathbf{y}_j\|^2)^{-1}}{\sum_k \sum_{l \neq k} (1 + \|\mathbf{y}_k - \mathbf{y}_l\|^2)^{-1}}
\]
t-SNE

- Compute pairwise similarities between data with normalized Gaussian kernel
- Measure normalized Student-t similarities in the t-SNE map
- Minimize the divergence between both distributions

\[ KL(P||Q) = \sum_i \sum_{j \neq i} p_{ij} \log \frac{p_{ij}}{q_{ij}} \]
t-SNE

• Compute pairwise similarities between data with normalized Gaussian kernel

• Measure normalized Student-t similarities in the t-SNE map

• Minimize the divergence between both distributions

\[
KL(P||Q) = \sum_i \sum_{j \neq i} p_{ij} \log \frac{p_{ij}}{q_{ij}}
\]
Why this loss?

- The Kullback-Leibler divergence preserves local data structure

\[
KL(P||Q) = \sum_i \sum_{j \neq i} p_{ij} \log \frac{p_{ij}}{q_{ij}}
\]
Why this loss?

• The Kullback-Leibler divergence preserves local data structure

\[ KL(P||Q) = \sum_i \sum_{j \neq i} p_{ij} \log \frac{p_{ij}}{q_{ij}} \]

Big \( p \), small \( q \)? Ouch!
Small \( p \), big \( q \)? No worries!
Why this loss?

- The Kullback-Leibler divergence preserves local data structure

- The heavy-tailed distribution corrects volume differences between both spaces
Efficiency

- Naive implementations are quadratic in the number of data points
Efficiency

- Naive implementations are quadratic in the number of data points
- Approximations are possible by grouping interactions between groups of distant points
Efficiency

- Naive implementations are quadratic in the number of data points
- Approximations are possible by grouping interactions between groups of distant points
Do's of using t-SNE
Do:

- Use t-SNE to get some qualitative hypotheses on what your features capture
Do:

- Be creative as to what inputs you use into t-SNE
Do:

- Be creative as to what inputs you use into t-SNE

- Trees, networks, graphs, co-occurrences, and associations naturally take the form of \( p_{ij} \)'s!
Do:

- Be creative in how you visualize the outputs of t-SNE
Do:

- Be creative in how you visualize the outputs of t-SNE
Do:

- Be creative in how you visualize the outputs of t-SNE
Indiana Jones, Final Fantasy,
Raiders of the Lost Ark, Star Wars

Wallace & Gromit: The Curse of the Were-Rabbit
The Simpsons: Season 1
Family Guy, Vol. 1: Seasons 1-2
South Park: Bigger, Longer and Uncut

Team America: World Police

Mission: Impossible II
Mission: Impossible

The World Is Not Enough
Tomorrow Never Dies
GoldenEye
The Spy Who Loved Me

Friends
Star Trek
Don’ts of using t-SNE
Don’t:

- Present a “proof by t-SNE”: your map is not the data!
Don’t:

• Present a “proof by t-SNE”: your map is not the data!

• Forget to consider alternative hypotheses
Don’t:

- Present a “proof by t-SNE”: your map is not the data!
- Forget to consider alternative hypotheses

“The visualization shows that our method learns meaningful transformations.”
(Purple is the base state and red is the transformed base state.)
Don’t:

- Present a “proof by t-SNE”: your map is not the data!
- Forget to consider alternative hypotheses

Repeatingly multiply the base state by a constant >1, and apply t-SNE on the result. You’ll get roughly the same map.

“The visualization shows that our method learns meaningful transformations.”
(Purple is the base state and red is the transformed base state.)
Don’t:

• Assign meaning to distances across empty space
Don’t:

- Assign meaning to distances across empty space
- Beyond the “scale” of the Student-t distribution, all similarities are infinitesimal
Don’t:

• Assign meaning to distances across empty space

• Beyond the “scale” of the Student-t distribution, all similarities are infinitesimal
Don’t:

- Assign meaning to distances across empty space

- Beyond the “scale” of the Student-t distribution, all similarities are infinitesimal

Are ones similar to zeros?
Don’t:

- Think that t-SNE will help you find outliers, or assign meaning to point densities in clusters
Don’t:

- Think that t-SNE will help you find outliers, or assign meaning to point densities in clusters

How input similarities in t-SNE are actually computed
Don’t:

- Think that t-SNE will help you find outliers, or assign meaning to point densities in clusters

**How input similarities in t-SNE are actually computed**

1. Compute *conditional* similarities:

   \[ p_{j|i} = \frac{\exp(-\|x_i - x_j\|^2/2\sigma_i^2)}{\sum_{j' \neq i} \exp(-\|x_i - x_{j'}\|^2/2\sigma_i^2)} \]

   Perform a binary search over \( \sigma_i \) to obtain a target *perplexity*. 
Don’t:

- Think that t-SNE will help you find outliers, or assign meaning to point densities in clusters

---

**How input similarities in t-SNE are actually computed**

1. Compute *conditional* similarities:

   \[
   p_{j|i} = \frac{\exp\left(-\|\mathbf{x}_i - \mathbf{x}_j\|^2/2\sigma_i^2\right)}{\sum_{j' \neq i} \exp\left(-\|\mathbf{x}_i - \mathbf{x}_{j'}\|^2/2\sigma_i^2\right)}
   \]

   Perform a binary search over \(\sigma_i\) to obtain a target *perplexity*.

2. *Symmetrize* the conditionals:

   \[
   p_{ij} = \frac{p_{j|i} + p_{i|j}}{2N}
   \]
Don’t:

• Forget that scale (perplexity) matters
Don’t:

- Forget that scale (perplexity) matters
- You can think of perplexity as the "effective" number of nearest neighbors

* See https://distill.pub/2016/misread-tsne/ for interactive version of this plot.
Don’t:

- Forget that t-SNE minimizes a non-convex objective: there are local minima
Don’t:

- Forget that t-SNE minimizes a non-convex objective: there are local minima
- Local minima generally split a natural cluster into multiple parts

* Figure from the Matlab manual page on t-SNE.
Don’t:

- Forget that t-SNE minimizes a non-convex objective: there are local minima

- Local minima generally split a natural cluster into multiple parts

* Figure from the Matlab manual page on t-SNE.
Don’t:

- Forget that t-SNE minimizes a non-convex objective: there are local minima

- Local minima generally split a natural cluster into multiple parts

* Figure from the Matlab manual page on t-SNE.
Don’t:

• Forget that t-SNE minimizes a non-convex objective: there are local minima

• Local minima generally split a natural cluster into multiple parts

• It is okay to run t-SNE multiple times and pick the best solution

* Figure from the Matlab manual page on t-SNE.
Don’t:

• Forget that low-dimensional metric spaces cannot capture non-metric similarities
Don’t:

- Forget that low-dimensional metric spaces cannot capture non-metric similarities
Don’t:

• Forget that low-dimensional metric spaces cannot capture non-metric similarities
Don’t:

- Forget that low-dimensional metric spaces cannot capture non-metric similarities.
t-SNE is a valuable tool in generating hypotheses and understanding, but does not produce conclusive evidence.

Caveats
- Like clustering techniques, t-SNE has a "scale"
- t-SNE reveals only select parts of the structure of the data
- Certain structure can never be reflected in a low-dimensional map
Thank you!

- Source code: http://lvdmaaten.github.io/tsne