E9: 309 Advanced Deep Learning

9-11-2020

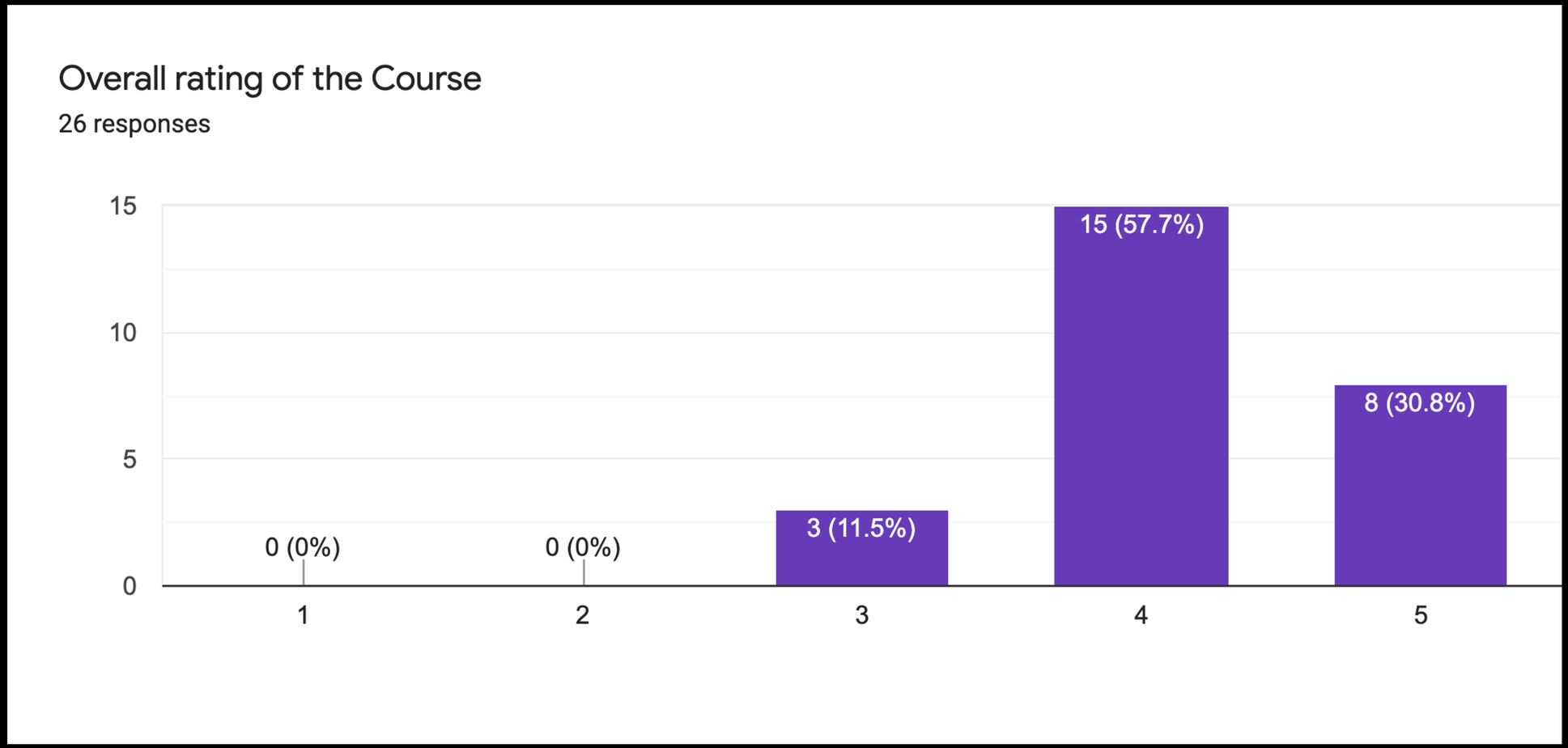
http://leap.ee.iisc.ac.in/sriram/teaching/ADL2020/

Housekeeping

- 1st mini-project
  - ✓ Deadlines
    - ★ Presentation on Nov19 and Nov20
      - ★ Your date allocation has been finalized
      - ★ Presentation and report template will be sent out this week.
        - ★ Report 1 page + references and tools
        - ★ Slides 4 slides for individual project and 6 slides for 2-member.

Feedback received

- Overall rating



Feedback - Good, bad and ugly.

Clarifying of doubts during lectures need to be a bit more tactful. Questions related to explanations are fine but a lot of off topic questions are taking too much time. It would be better to answer them at the end of the lecture rather than during.

Either one of the midterm or final exam should have been abolished.

The course structure can be improved by my suggestion about the projects above. I wish that the lectures were more clear on the Math behind the models. The lectures give a lot of intuition but I feel concreteness is somewhat lacking. I do understand that Deep Learning has a lot of heuristics and it is difficult to make the subject matter concrete, but if possible it would improve the overall understanding of students.

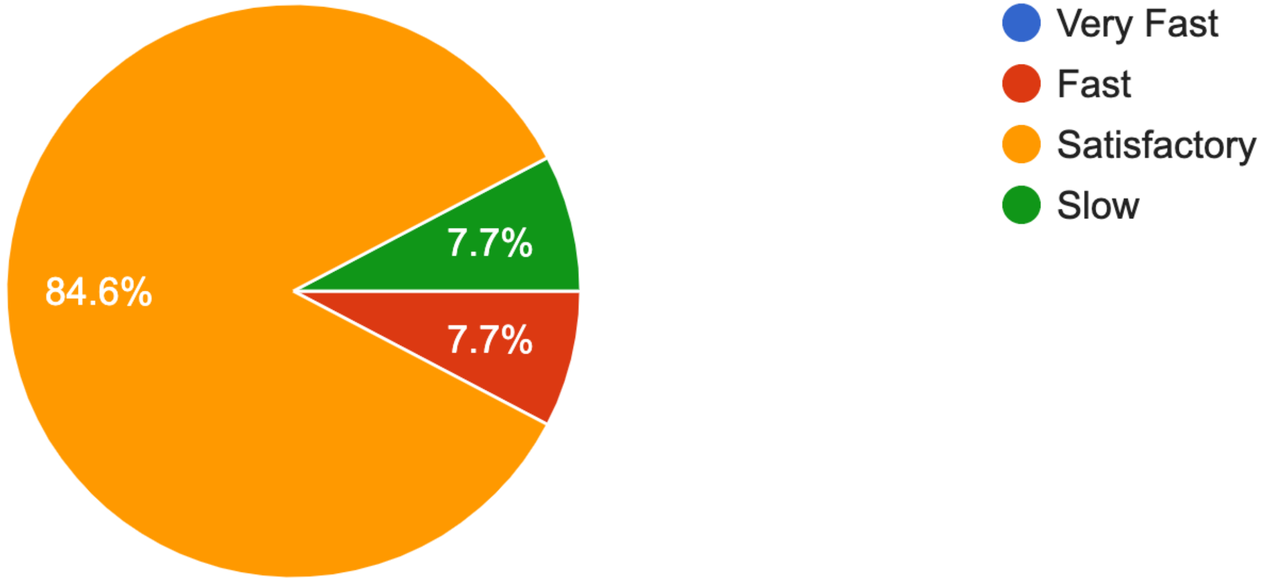
Pace: Most of the time I feel the pace is too slow. I had done a Deep Learning course last semester and this being an "advanced" course I expected advanced topics. But most of the topics covered in this course (till now) was already covered in the DL course.

The structure of the course is awesome, although the instructor's mic has been troublesome in some of the lectures(low volume).

Feedback

Pace of the Teaching / lecturers

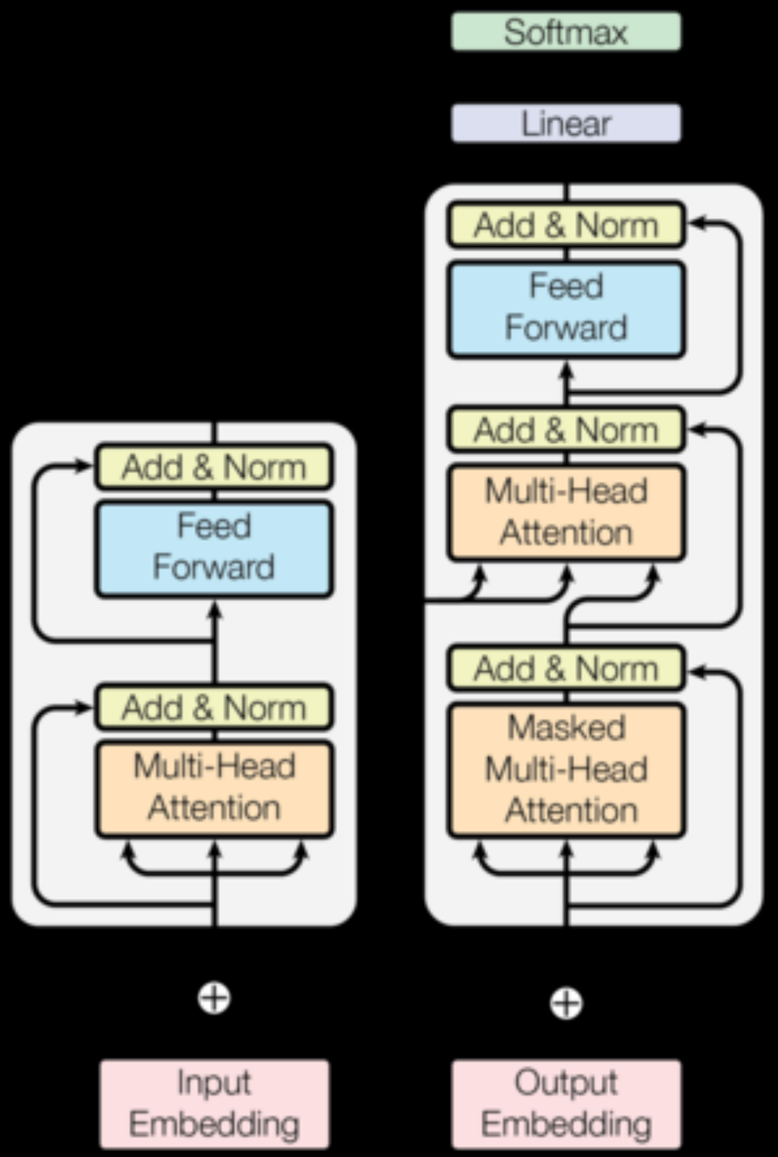
26 responses



Recap of previous class

Transformer - training

- Input sequence and the output sequence shifted right by 1
-



Topics thus far …

Visual and Time Series Modeling: Semantic Models, Recurrent neural models and LSTM models, Encoder-decoder models, Attention models.

Representation Learning, **Causality And Explainability**: t-SNE visualization, Hierarchical Representation, semantic embeddings, gradient and perturbation analysis, Topics in Explainable learning, Structural causal models.

Unsupervised Learning: Restricted Boltzmann Machines, Variational Autoencoders, Generative Adversarial Networks.

New Architectures: Capsule networks, End-to-end models, Transformer Networks.

Applications: Applications in in NLP, Speech, Image/Video domains in all modules.

Representation learning/data-visualization

- Learning a lower dimensional representation
  - Unsupervised representation learning can be useful in visualization
    - ✓ Example: Principal component analysis (PCA)
    - ✓ Explaining deep layers and information propagation in deep learning.
- Techniques which generate transformation applicable to newer data versus dataset specific dimensionality reduction.
  - Question - can we perform neighborhood preserving representation learning.

Neighborhood preserving dimensionality reduction

- Dimensionality reduction problem

$$\mathbf{x}_1, \mathbf{x}_2 \ldots \mathbf{x}_N \in \mathcal{R}^D \longrightarrow \mathbf{y}_1, \mathbf{y}_2 \ldots \mathbf{y}_N \in \mathcal{R}^d \qquad D > d$$

- Neighborhood - based on some distance metric

Example, $d_{i,j} = \dfrac{||\mathbf{x}_i - \mathbf{x}_j||^2}{2\sigma^2}$

- Neighborhood preservation

  - ✓ If two samples are close in the original space, this has to be preserved in the lower dimensional representation as well.

Stochastic neighborhood embedding

- Define a probability distribution based on neighborhood

$$p_{j|i} \triangleq \frac{exp(-d_{ij}^2)}{\sum_{k \neq i} exp(-d_{ik}^2)} \quad ; \quad p_{i|i} = 0$$

- define a distribution on the lower dimensional space. fixed variances $\sigma^2 = 1/2$

$$q_{j|i} \triangleq \frac{exp(-||\mathbf{y}_i - \mathbf{y}_j||^2)}{\sum_{k \neq i} exp(-||\mathbf{y}_i - \mathbf{y}_k||^2)}$$

- Distance preservation is formulated as making the distributions similar.

Error function and optimization

- Error function

$$E = \sum_j \sum_i KL(p_{j|i} || q_{j|i}) = \sum_j \sum_{i \neq j} p_{j|i} log \frac{p_{j|i}}{q_{j|i}}$$

- unknowns are the lower dimensional embeddings $\mathbf{y}_1, \mathbf{y}_2 \ldots \mathbf{y}_N \in \mathcal{R}^d$

  - ✓ can be solved using gradient descent

$$\frac{\partial E}{\partial \mathbf{y}_i} = 2 \sum_j (p_{j|i} - q_{j|i} + p_{i|j} - q_{i|j})(\mathbf{y}_i - \mathbf{y}_j)$$

Iterative process

SNE

$\{x_1 \dots x_N\} \in \mathbb{R}^D$. x_{N+1}

Initialize : $\{y_1^0 \dots y_N^0\} \in \mathbb{R}^d$ from a Gaussian with zero mean.

$E. = KL(p||q)$

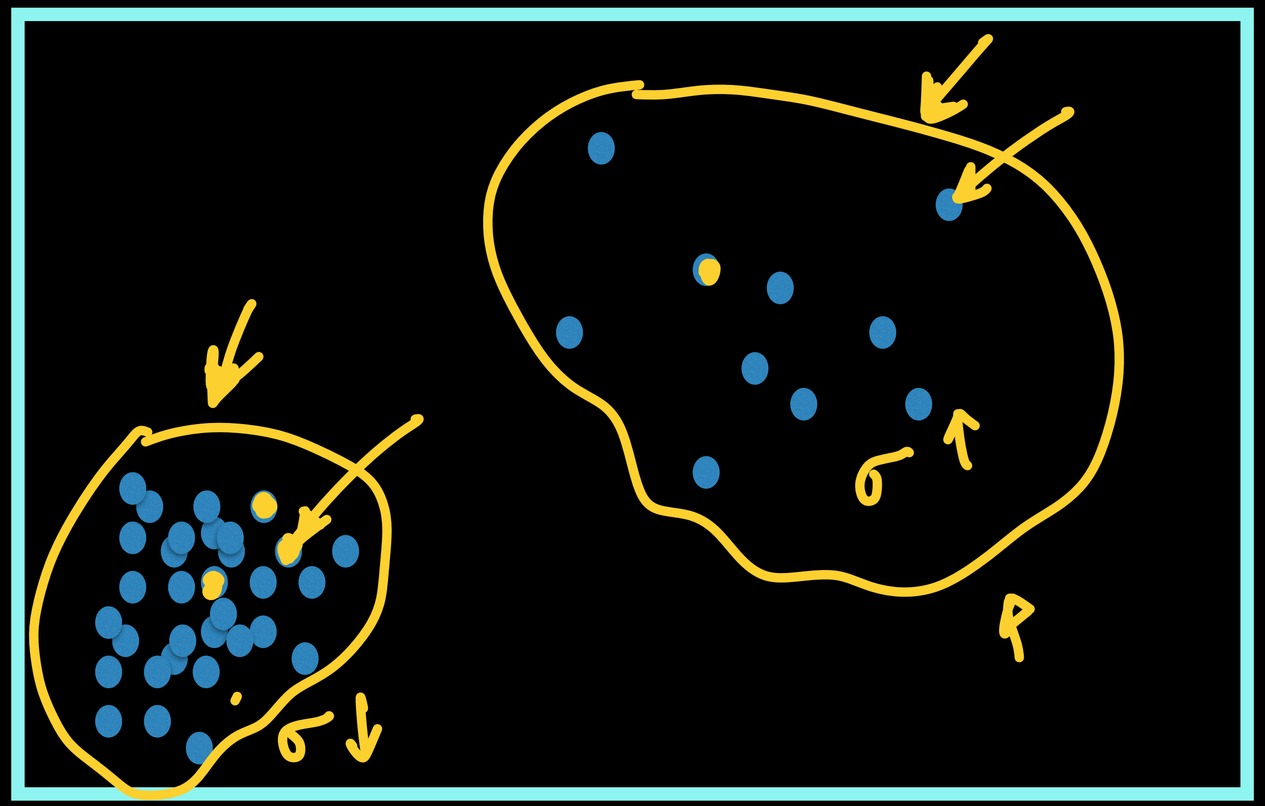
$y_i^{n+1} \rightarrow y_i^n - \eta \left\{ \frac{\partial E}{\partial y_i} \right\}$ { Including momentum }

$\Downarrow$

$\{y_1^* \dots y_N^*\}$ y_{N+1}

Stochastic neighborhood embedding

- Dense versus sparse data regions



$$d_{i,j} = \frac{||\mathbf{x}_i - \mathbf{x}_j||^2}{2\sigma^2}$$

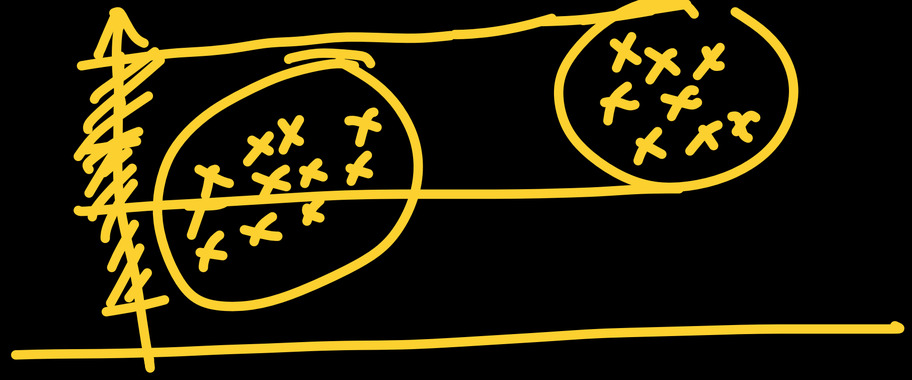
Having a uniform variance may be harmful .
Using a data point specific variance

$$d_{i,j} = \frac{||\mathbf{x}_i - \mathbf{x}_j||^2}{2\sigma_i^2}$$

Variance can be chosen to have uniform entropy

$$H_i = -\sum_j p_{j|i} log(p_{j|i})$$

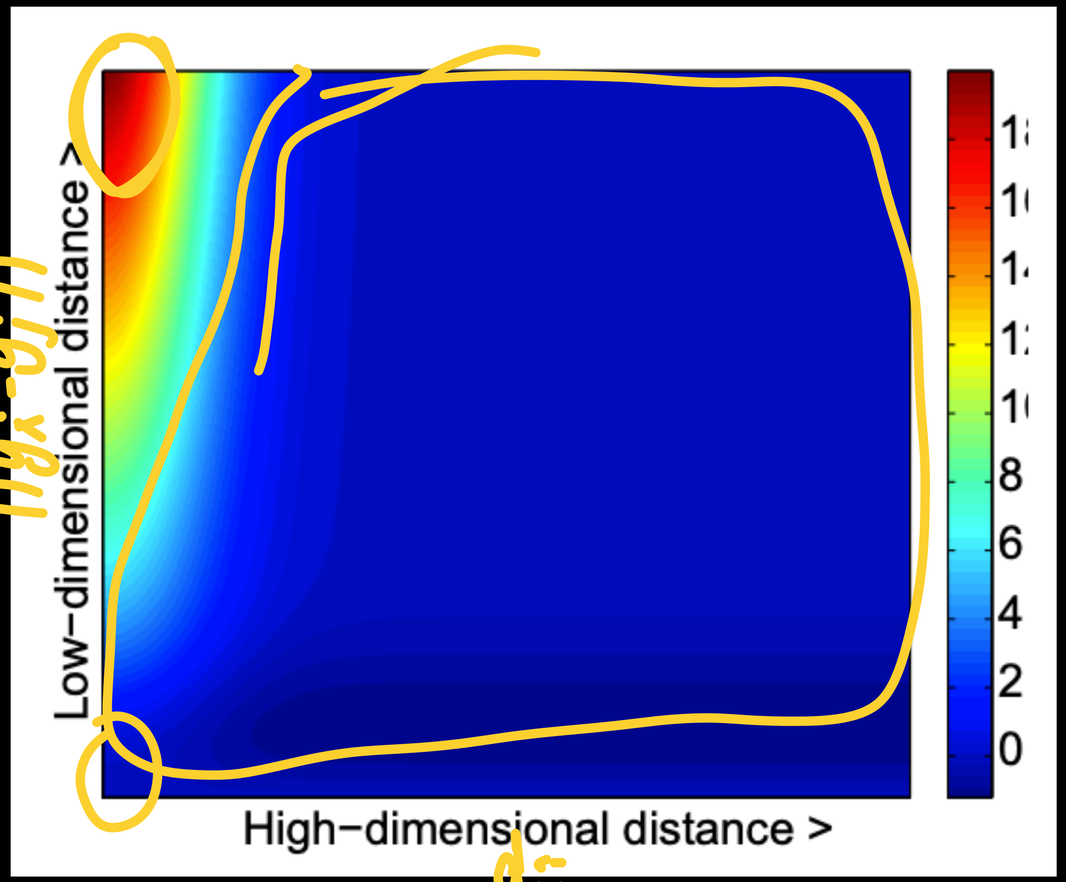
Crowding problem in SNE



- When going form high dimensions to lower
  - ✓ Data in high dimensions tend to loose their spread and therefore tend to crowd in lower dimensions.
    - ★ the area of the lower dimensional map that is available to accommodate moderately distant datapoints will not be nearly large enough compared with the area available to accommodate nearby datapoint.
    - ★ clusters may not be visible in lower dimensional space.
    - ★ partly due to the Gaussian assumptions on the lower dimensional space.

Crowding problem in t-SNE

- Gradient behavior is also unintended



Plot of the gradient

t-SNE

- Student t-distribution with 1-degree of freedom

$$q_{j|i} = \frac{(1 + ||\mathbf{y}_i - \mathbf{y}_j||^2)^{-1}}{\sum_{k \neq i}(1 + ||\mathbf{y}_i - \mathbf{y}_k||^2)^{-1}}$$

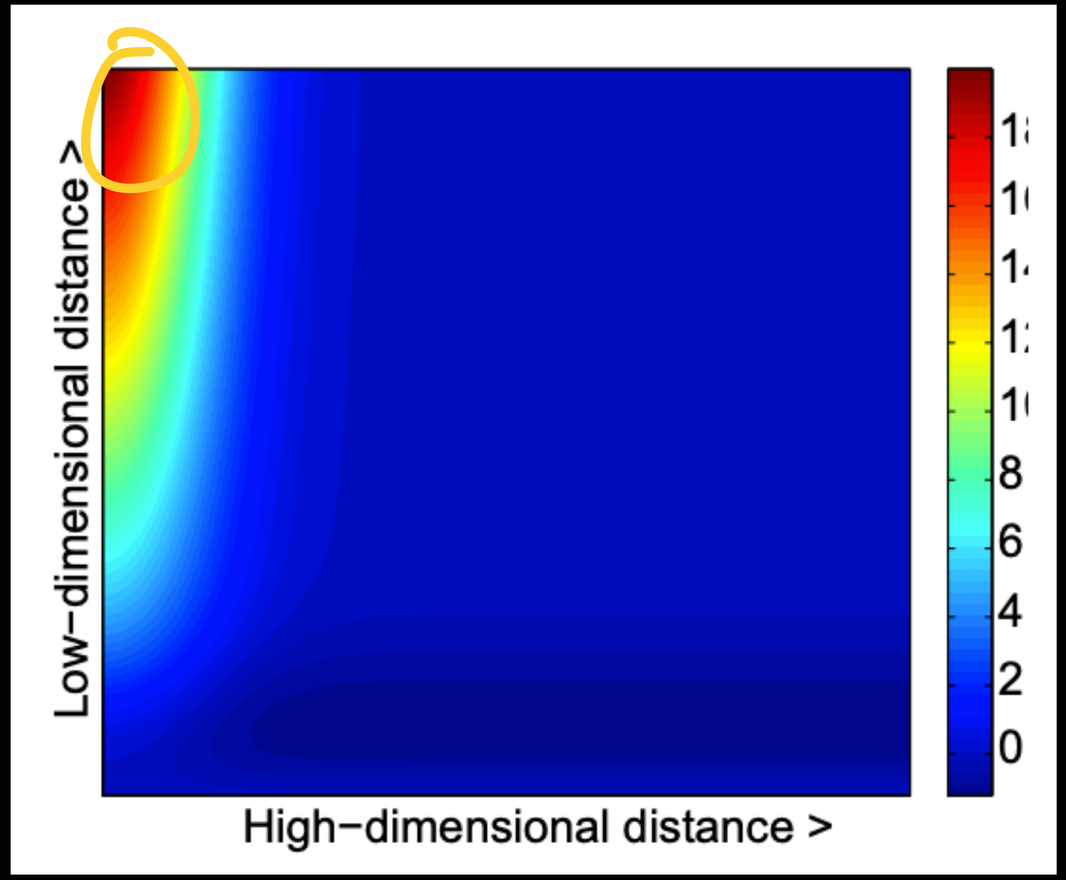
- The gradients are better behaved with this modification (follow and inverse square law).

$$\frac{\partial E}{\partial \mathbf{y}_i} = 2 \sum_j (p_{j|i} - q_{j|i})(\mathbf{y}_i - \mathbf{y}_j)(1 + ||\mathbf{y}_i - \mathbf{y}_j||^2)^{-1}$$

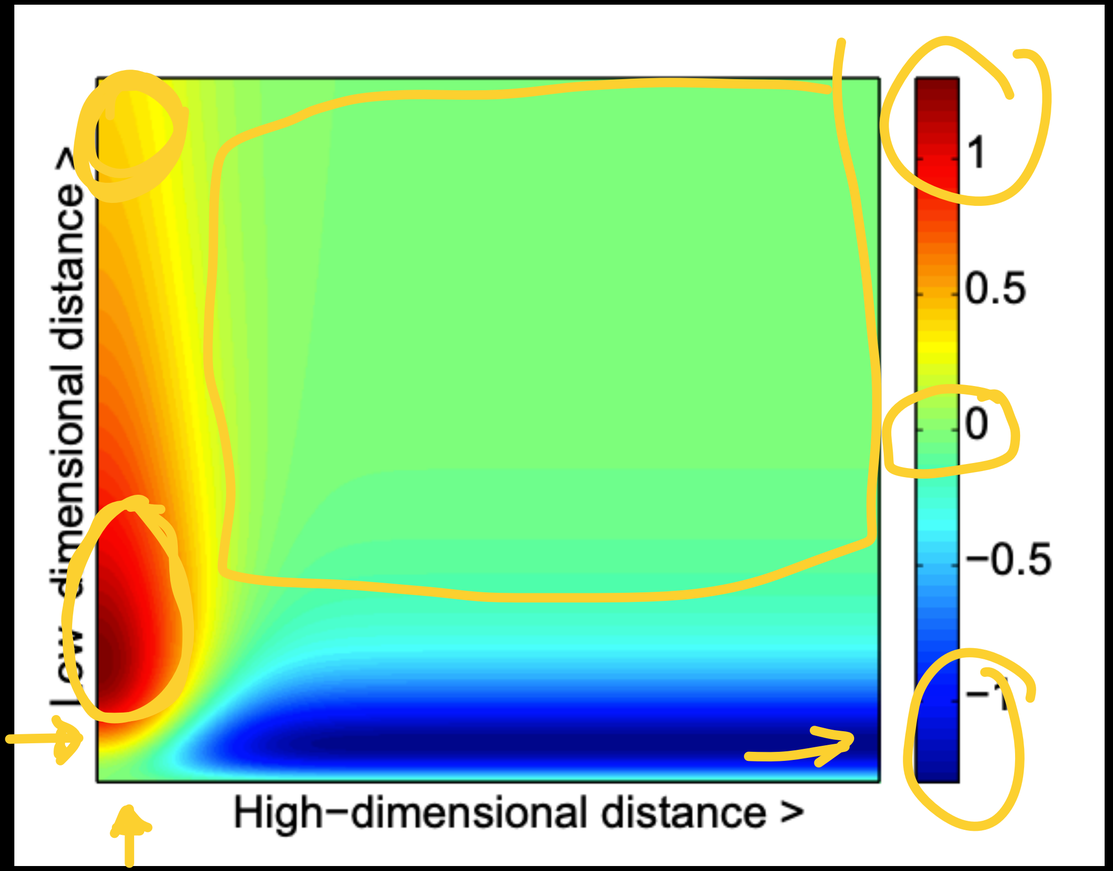
Reading assignment - "Visualizing Data using t-SNE", van der Maaten and Hinton, Journal of Machine Learning Research, 2008.

Crowding problem improved in t-SNE

- Gradient behavior is improved with the modification



SNE

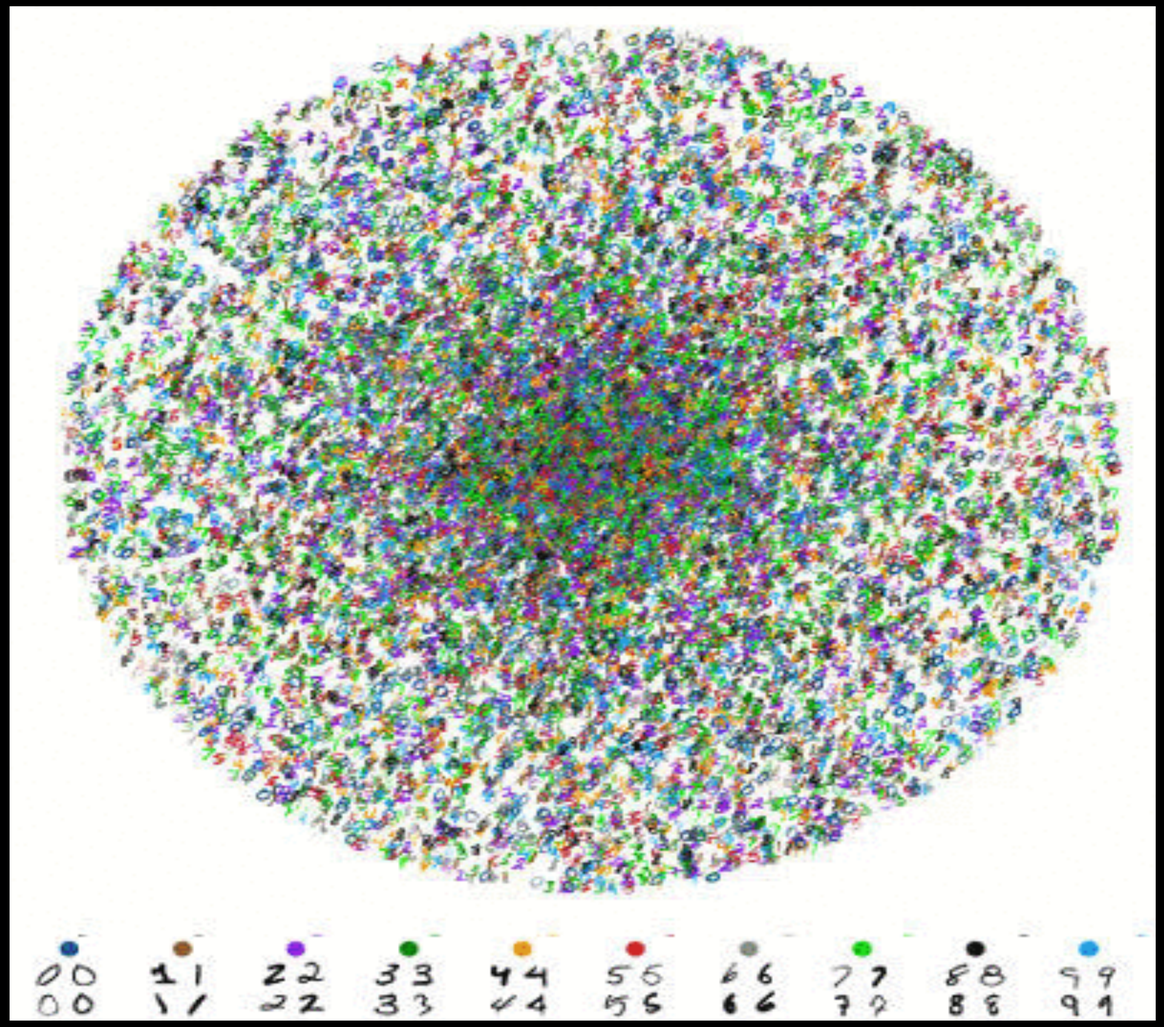


t-SNE

Embeddings of M-NIST dataset

$D = 784$
$d = 2$

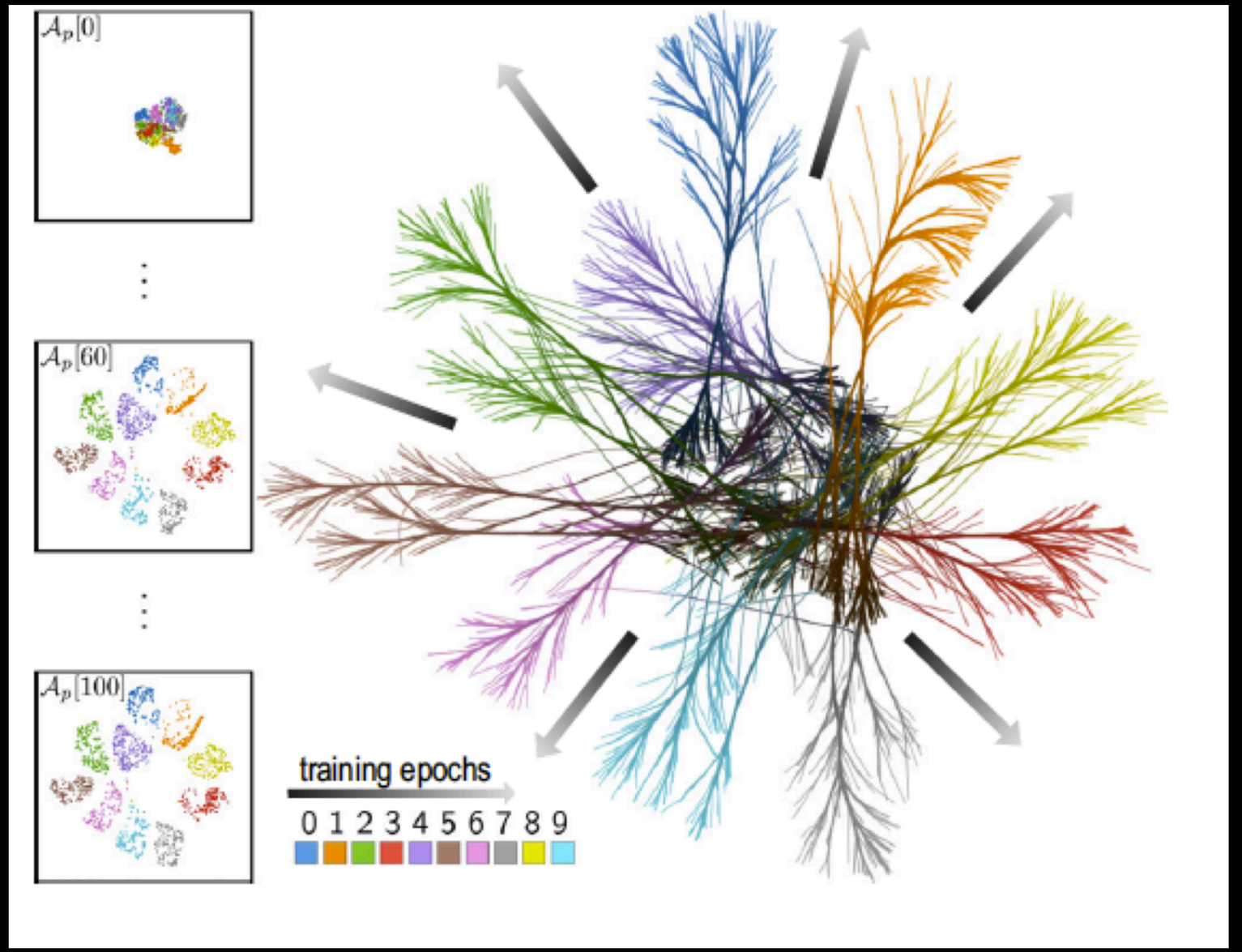
- Handwritten digits of 784D projected onto 2-D using t-SNE



Source : https://ai.googleblog.com/2018/06/realtime-tsne-visualizations-with.html

Visualizing hidden layer activations in CNNs using t-SNE

Inter-epoch evolution of CNN hidden activations



Rauber, Paulo E., et al. "Visualizing the hidden activity of artificial neural networks." *IEEE transactions on visualization and computer graphics* 23.1 (2016): 101-110.

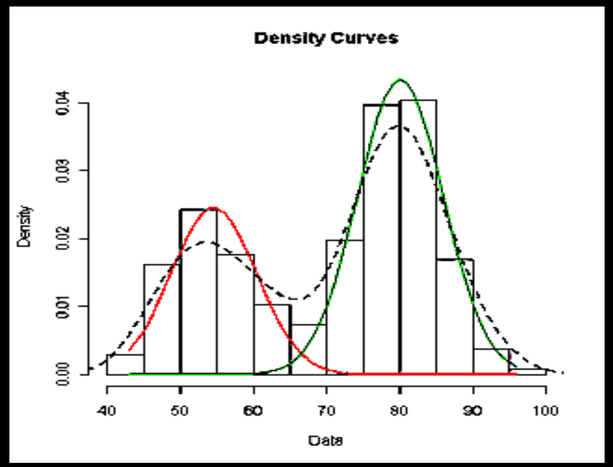
Practical considerations

- Hyper-parameters
  - ✓ Perplexity, learning-rate and number of iterations.
- Pros
  - ✓ Data neighborhood preserving and relatively intuitive in visualizing data.
- Cons
  - ✓ Iterative learning, does not provide a transform applicable on unseen data.

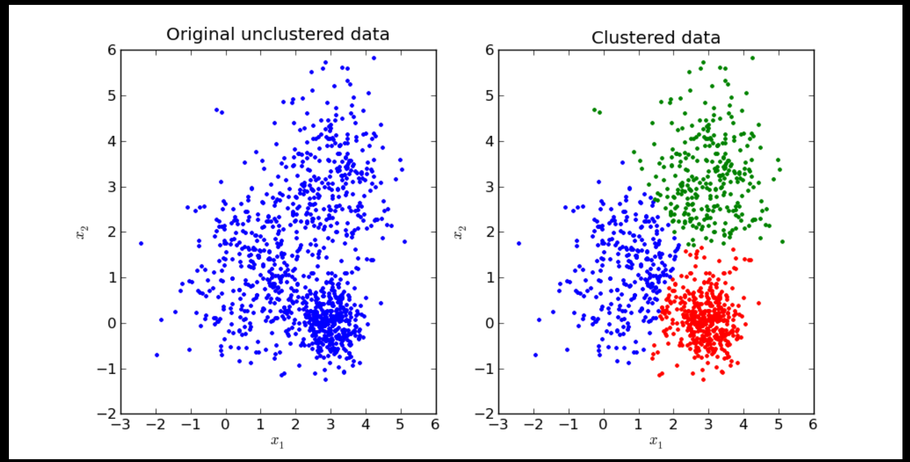
Unsupervised learning

- Developing models that do not need labels
  - May model the generation of data.
  - May allow generation of new data samples
- Broad strategies for unsupervised learning

Based on maximizing likelihood



Based on clustering



Boltzmann machine