# EEP 596: LLMs: From Transformers to GPT ‖ Lecture 12

Dr. Karthik Mohan

Univ. of Washington, Seattle

February 20, 2025

# Deep Learning and Transformers References

### Deep Learning

Great reference for the theory and fundamentals of deep learning: Book by Goodfellow and Bengio et al Bengio et al
Deep Learning History

### Embeddings

SBERT and its usefulness
SBert Details
Instacart Search Relevance
Instacart Auto-Complete

### Attention

Illustration of attention mechanism

# Generative AI References

**Prompt Engineering**

Prompt Design and Engineering: Introduction and Advanced Methods

**Retrieval Augmented Generation (RAG)**

Toolformer

RAG Toolformer explained

**Misc GenAI references**

Time-Aware Language Models as Temporal Knowledge Bases

# Generative AI references

Stable Diffusion

Diffusion Explainer: Visual Explanation for Text-to-image Stable Diffusion

The Illustrated Stable Diffusion

# House Keeping

- MP2 (Mini-Project 2) part 3 assigned
- MP2 Part 3 will focus more on the design aspects of the chatbot
- There is no one best way to do the design - Experiment and see which one helps you meet the criteria mentioned in the part 3
- Any questions/concerns?

# Previous Lecture

- Design of Chatbots

# Today's lecture

- Upcoming **Machine Learning Interview Masterclass Short Course**
- More Design principles for Chatbots
- Intro to Image Foundation Models
- Auto Encoders

# Machine Learning Interview Masterclass Short Course - Heads up

1. 1 credit course
2. 12 hours over 2 weekends only
3. Dates: 4/5, 4/6 and 4/12, 4/13
4. Covers practical tips and hacks for better ML Interview Prep
5. Covers ML Design, ML Depth and Coding, ML Coding rounds + Overall Interviewing tips
6. Theory + Hands-on in-person practice + Mock peer-interview sessions

# ML Interview Masterclass - Target Audience

1. You are a working professional or a full-time student with machine learning background
2. You want to prep for Data Science Interviews but are unsure where to start
3. You would like to benefit from structured interviewing prep material
4. You want to gain comfort with prepping for interviews and getting to know how to ace them
5. You are looking for simulated/mock interview sessions with peers
6. You would like to understand where you stand and get feedback for improvements

# Vision for the Course

## Elements in the Course

1. **Hands-on Prep and Practice** The students will get both theory and practice sessions for interviewing prep
2. **3 hour modules:** There will be four 3 hour modules spanning two weekends.
3. **Presentation + Hands-on Tutorial in every module:** The 3 hour module will be a combination of some background, theory and examples for the first half, followed by in-class exercises and mock interview sessions (with peers) that students can work on.
4. **Peer Learning:** Each 3 hour module will also have couple of break-outs for the class to discuss in-class conceptual questions in small groups.

# Vision for the Course

## Elements in the Course

1. **Take Home assignment:** Take-home assignment will have you track your prep progress across different areas of interviewing.
2. **Optional Presentation:** There will be an optional 1 hour presentation slot to show-case mini-project where students can present their learnings from the class and progress on their interviewing take-home assignment
3. **Pitfalls to avoid:** Keep the course real by discussing common pitfalls to avoid in the interviewing process.

# Pre-requisites

This is a little tricky for a short course - But someone with a familiarity with machine learning concepts and basic natural language processing is a must. Also coding in python is a must.

# 2 Week Course Flow

| **Module (3 hr)** | **Lecture Material** | **Coding Tutorial** |
|---|---|---|
| **4/5 2-5 pm** | Intro to Interview Process | Sample Coding Questions |
| **4/6 2-5 pm** | Coding Deep-dive | Coding tutorial |
| | | |
| **Take-Home** | Interview Prep Assignment | Simulated Interviewing |
| | | |
| **4/12 2-5 pm** | DS/ML Design | ML Breadth/Depth tutorial |
| **4/13 2-5 pm** | Real-world interviewing | Pitfalls to avoid |

# Concepts to be covered

1. **Intro to Interviewing Process**
2. **Interviewing Guidelines**
3. **Coding Round**
4. **ML Breadth round**
5. **ML Depth round**
6. **ML Design round**
7. **ML Coding round**
8. **How to prep for coding interview?**
9. **How to prep for ML interviews?**
10. **Gauging your gaps**
11. **What is the interviewer looking for?**
12. **Tips for better interviewing experience**

# Single Day Flow

| **Time** | **Topic** | **Highlights** |
|---|---:|---|
| **2 pm - 3:30 pm** | Interviewing Concepts | Tutorial |
| **3:30 pm - 3:45 pm** | Breakouts | Peer Brainstorm |
| **3:45 pm - 5 pm** | Hands-on coding | Peer Practice Sessions |

# Next Topic: Design of Chatbots

# Next Topic: Foundation Models for Images

# Foundation Models for Images

### Types

CNNs (e.g. Inception, AlexNet, etc) and Visual Transformers or ViTransformer

# Foundation Models for Images

### Types

CNNs (e.g. Inception, AlexNet, etc) and Visual Transformers or ViTransformer

### Building Blocks

Like legos can be used to build a whole factory - Foundation models can be put together across modes (multi-modal) to create interesting and beautiful applications. Text2Image is one such example that combines multiple foundation models - Transformers, ViTransformers, CNNs and also AutoEncoders.

# Foundation Models for Images

### Types

CNNs (e.g. Inception, AlexNet, etc) and Visual Transformers or ViTransformer

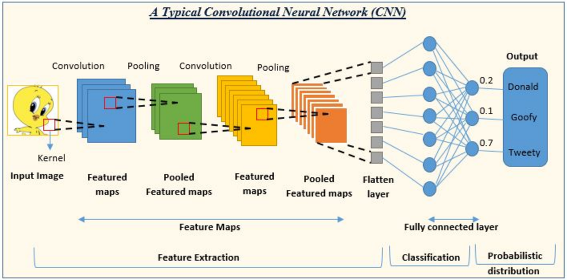
### Building Blocks

Like legos can be used to build a whole factory - Foundation models can be put together across modes (multi-modal) to create interesting and beautiful applications. Text2Image is one such example that combines multiple foundation models - Transformers, ViTransformers, CNNs and also AutoEncoders.

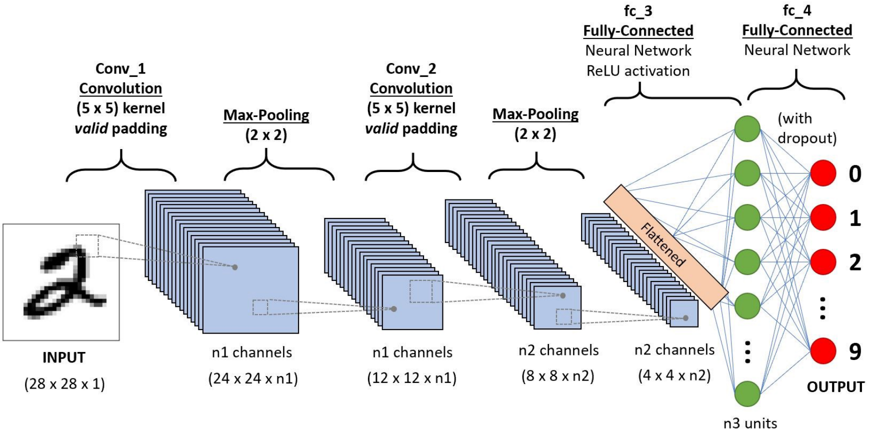
### Applications

Classification (cat or dog?), Image2Text, Text2Image, Image Embeddings, Object Detection, Image Segmentation, etc

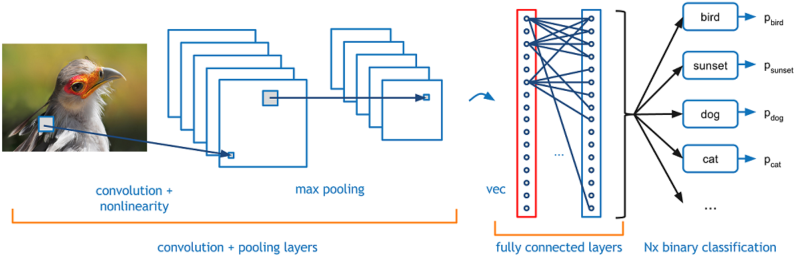
# Foundation Model - CNN (Convolutional Neural Network)



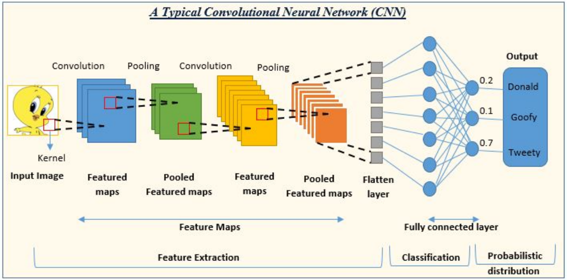
# Foundation Model - CNN (Convolutional Neural Network)



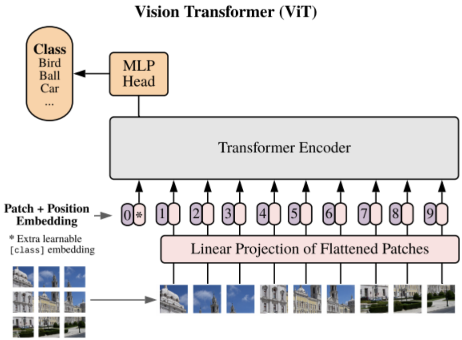
# Foundation Model - CNN (Convolutional Neural Network)



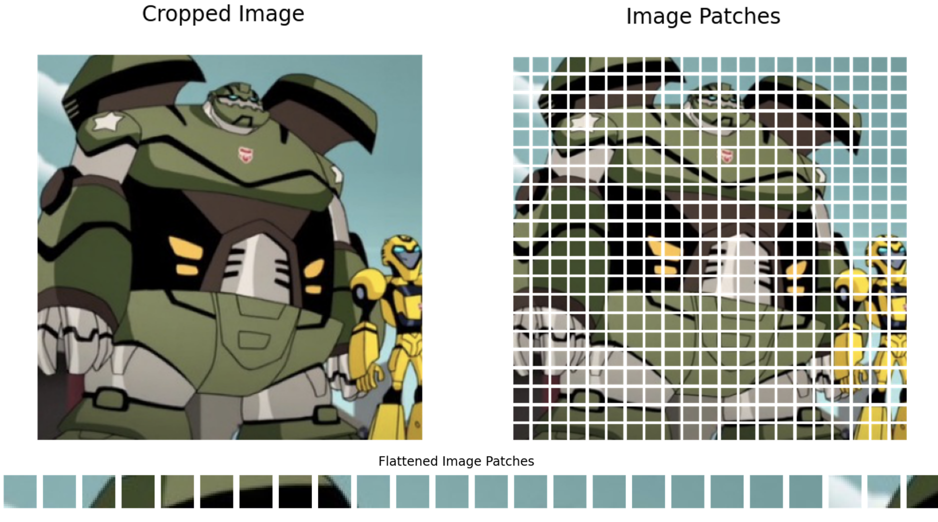
# Foundation Model - CNN (Convolutional Neural Network)



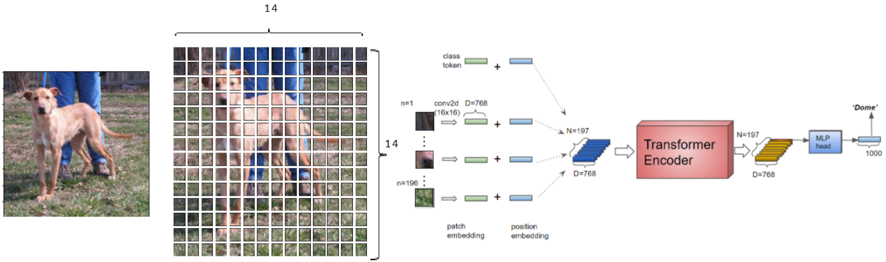
# Foundation Model - Visual Transformers (ViT)



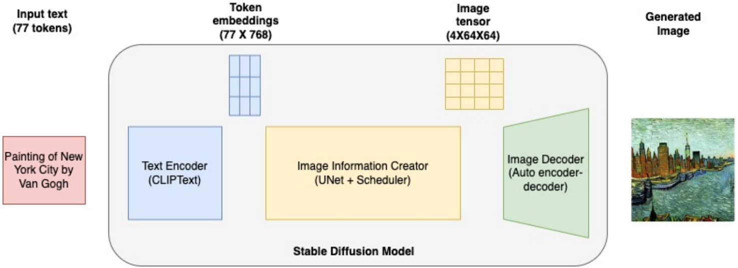
# Foundation Model - Visual Transformers (ViT)



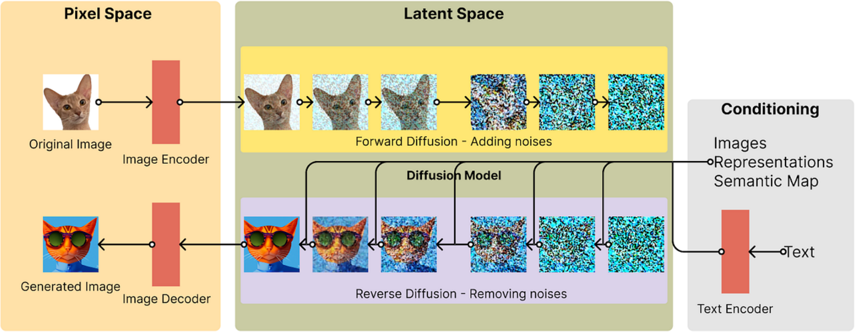
# Foundation Model - Visual Transformers (ViT)



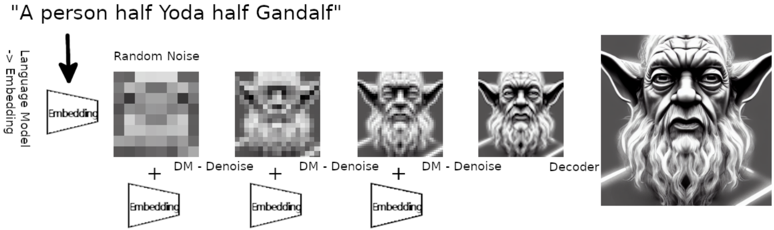
# Foundation Model - Stable Diffusion (Text2Image)



# Foundation Model - Stable Diffusion (Text2Image)



# Foundation Model - Stable Diffusion (Text2Image)

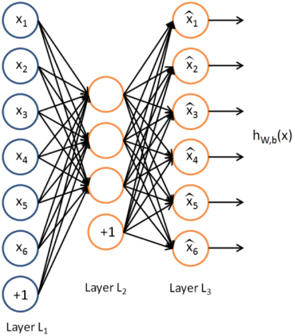



Reference

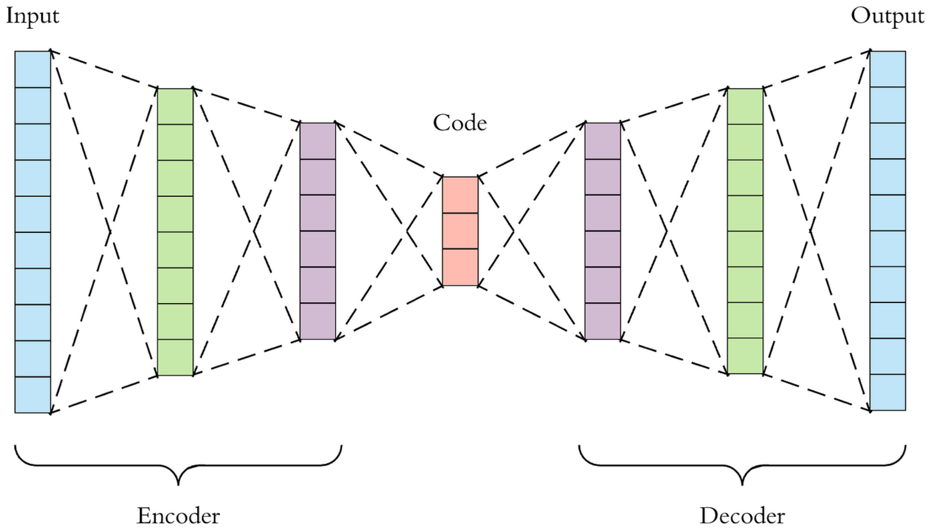
# Stable Diffusion Explained

- Based on the concept of "de-noising auto encoders" and the use of text prompt to *guide the de-noising*
- Stable diffusion is also trained to successfully de-noise and increase the resolution of the image using text guidance

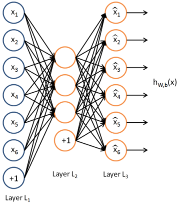
# Auto Encoders

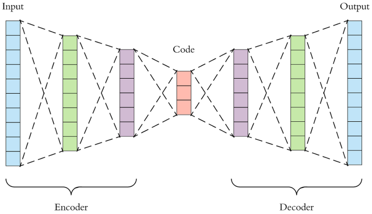


# Deep Auto Encoders



# PCA vs Auto-Encoders



# ICE #1

### PCA vs Auto Encoder

Which of the following statements are true ?

1. Both PCA and Auto Encoders serve the purpose of dimensionality reduction
2. They are both linear models but one uses a neural nets architecture and the other is based on projections
3. PCA is robust to outliers while Auto Encoders are not
4. Auto Encoders can compress images better than PCA

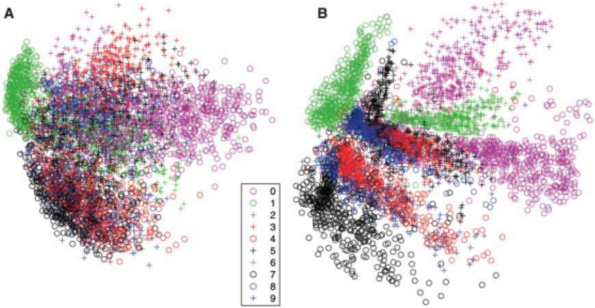
# AutoEncoders and Dimensionality Reduction

Visualization Performance

Auto Encoder Reference Paper

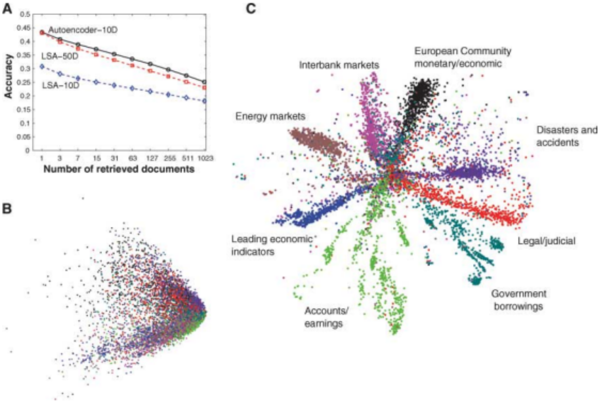
## Reading Reference for AE Dimensionality Reduction

**Fig. 3.** (**A**) The two-dimensional codes for 500 digits of each class produced by taking the first two principal components of all 60,000 training images. (**B**) The two-dimensional codes found by a 784-1000-500-250-2 autoencoder. For an alternative visualization, see (*8*).



## Reading Reference for AE Dimensionality Reduction

**Fig. 4.** (**A**) The fraction of retrieved documents in the same class as the query when a query document from the test set is used to retrieve other test set documents, averaged over all 402,207 possible queries. (**B**) The codes produced by two-dimensional LSA. (**C**) The codes produced by a 2000-500-250-125-2 autoencoder.



# AutoEncders Summary

1. Auto-Encoders are a method for dimensionality reduction and can do better than PCA for visualization

# AutoEncders Summary

1. Auto-Encoders are a method for dimensionality reduction and can do better than PCA for visualization
2. Use Neural Networks architecture and hence can encode non-linearity in the embeddings

# AutoEncders Summary

1. Auto-Encoders are a method for dimensionality reduction and can do better than PCA for visualization
2. Use Neural Networks architecture and hence can encode non-linearity in the embeddings
3. Anything else?

# AutoEncders Summary

1. Auto-Encoders are a method for dimensionality reduction and can do better than PCA for visualization
2. Use Neural Networks architecture and hence can encode non-linearity in the embeddings
3. Anything else?
4. Auto Encoders can learn convolutional layers instead of dense layers - Better for images! More flexibility!!

# ICE #2: Loss Function for Auto-Encoders

What is the loss function used to train Deep Auto-Encoders?

1. Logistic Loss
2. Quadratic Loss
3. Triplet Loss
4. Cross-Entropy Loss

# Removing obstacles in images

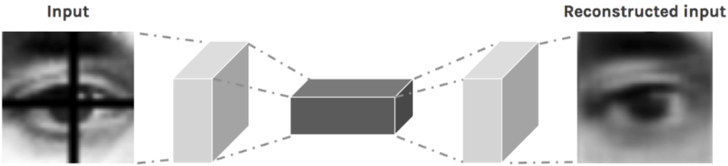



Figure 12: Reconstructed image from missing image [14]

# Removing obstacles in images

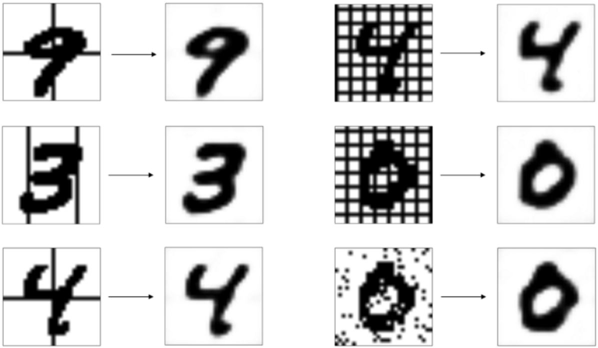
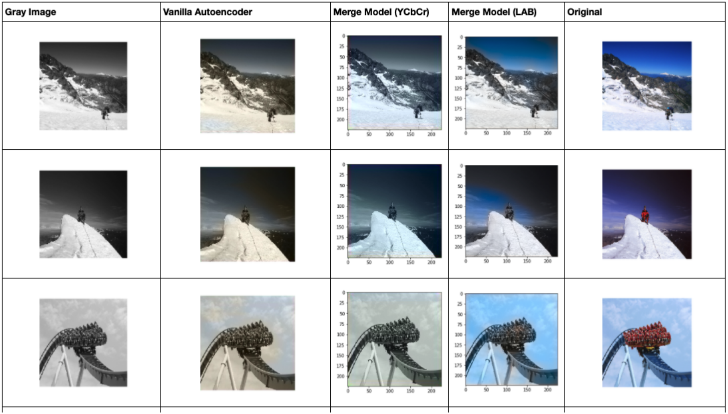
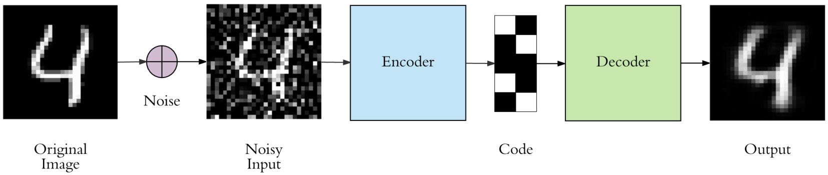


Figure 13: Source [15]

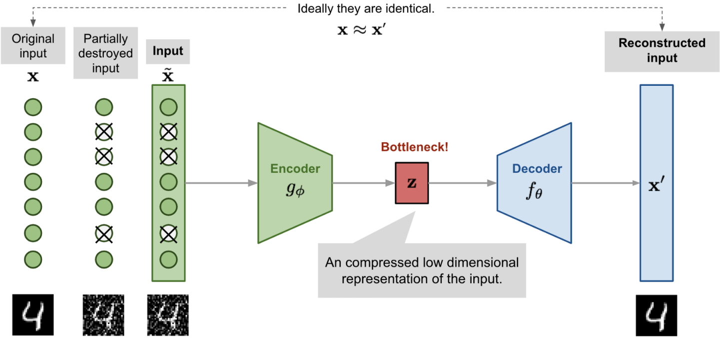
# Coloring Images



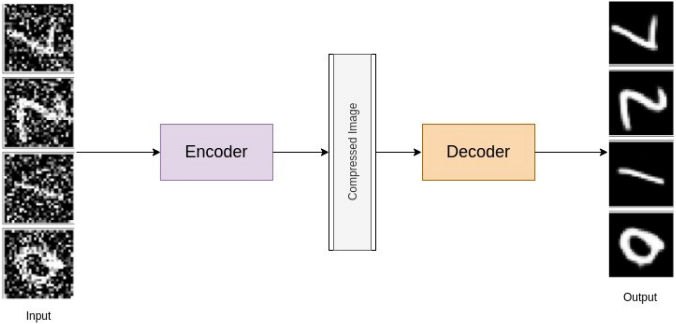
# De-noising Auto Encoders



# De-noising Auto Encoders



# De-noising Auto Encoders



# De-noising Auto Encoders

### Details

- Just like an Auto Encoder

# De-noising Auto Encoders

## Details

- Just like an Auto Encoder
- Difference: Noise is injected in the inputs on purpose but output is a clean data point.

# De-noising Auto Encoders

## Details

- Just like an Auto Encoder
- Difference: Noise is injected in the inputs on purpose but output is a clean data point.
- This forces the Auto Encoder to "de-noise" data, esp. useful for images!

# De-noising Auto Encoders

### Details

- Just like an Auto Encoder
- Difference: Noise is injected in the inputs on purpose but output is a clean data point.
- This forces the Auto Encoder to "de-noise" data, esp. useful for images!
- Esp. useful for a category of objects or images (e.g. digit recognition or face recognition, etc)

# ICE #3

### Unsupervised Learning

Which of these is NOT an example of unsupervised learning?

1. Perceptron
2. Auto Encoder
3. De-noising Auto Encoder
4. K-means++
5. None of the above
6. All of the above

# Breakouts Time #1

5 mins

Discuss in your groups what are some real-world applications of any or many of the Auto Encoder Architectures we discussed so far you can think of in your area of work or in a standard context e.g. images.