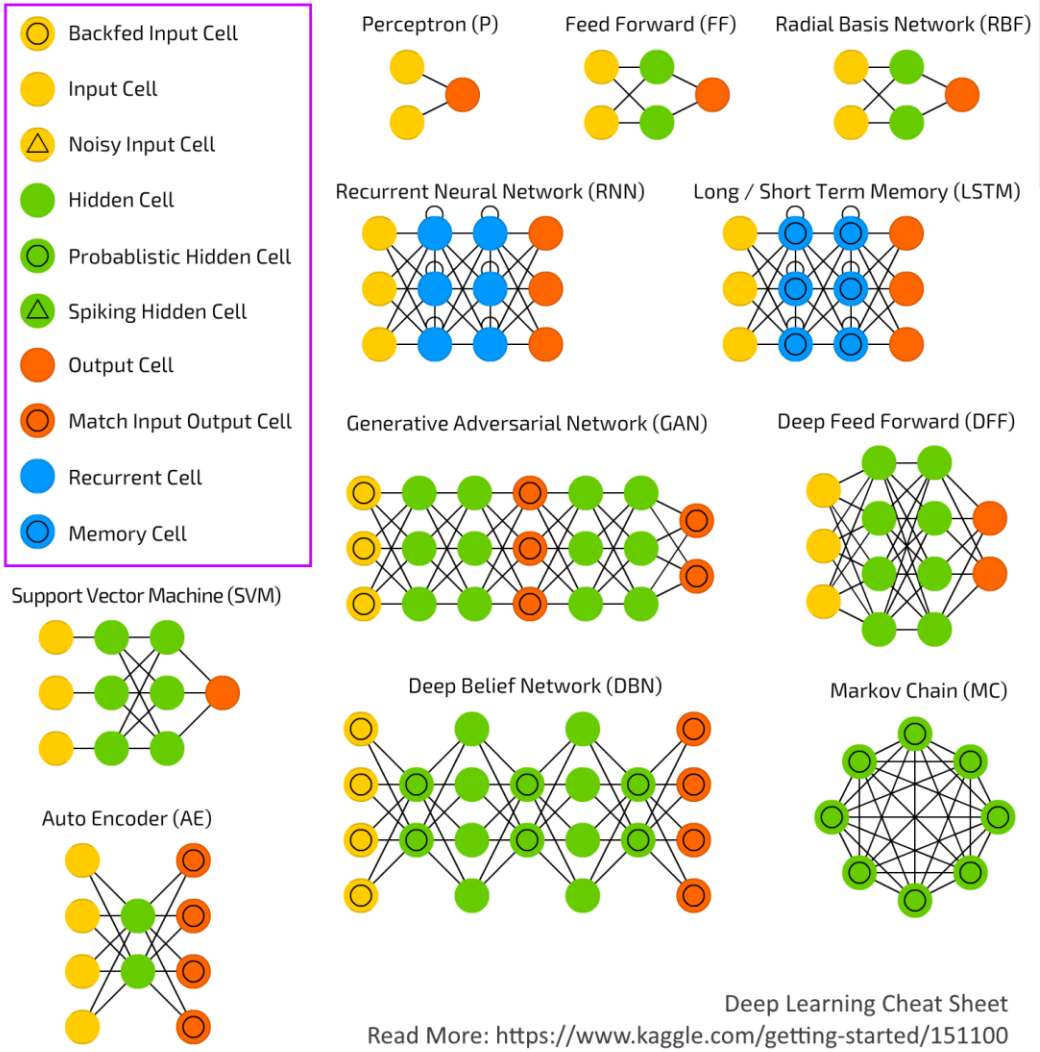
UNIVERSITY OF ARKANSAS

Deep Learning

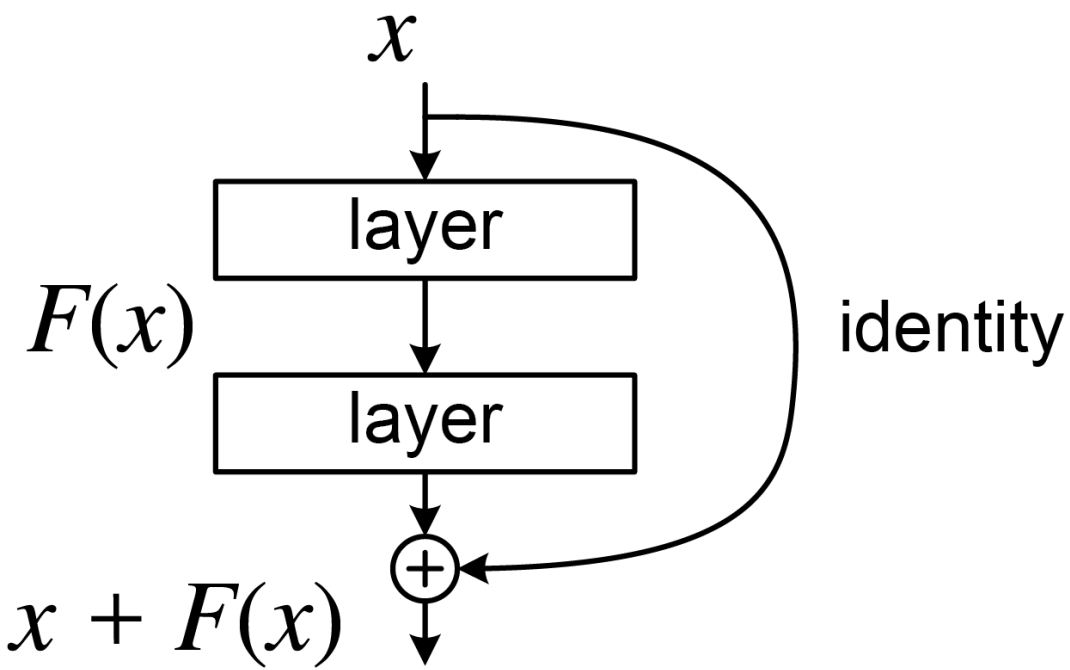
Jiahui Chen
Department of Mathematical Sciences
University of Arkansas

UNIVERSITY OF ARKANSAS

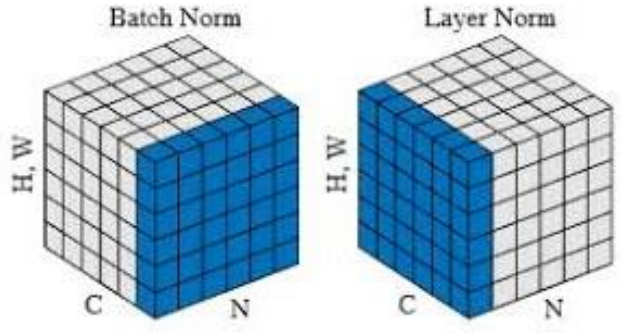


Deep neural network by 2017

Residual Networks (ResNet)



Batch and Layer Normalization



$$\mu_{\mathcal{B}} \leftarrow \frac{1}{m}\sum_{i=1}^{m} x_i$$

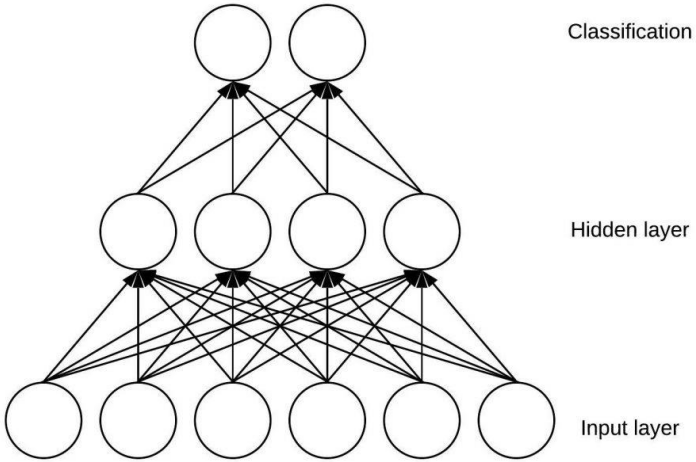
$$\sigma^2_{\mathcal{B}} \leftarrow \frac{1}{m}\sum_{i=1}^{m}(x_i - \mu_{\mathcal{B}})^2$$

$$\widehat{x}_i \leftarrow \frac{x_i - \mu_{\mathcal{B}}}{\sqrt{\sigma^2_{\mathcal{B}} + \epsilon}}$$

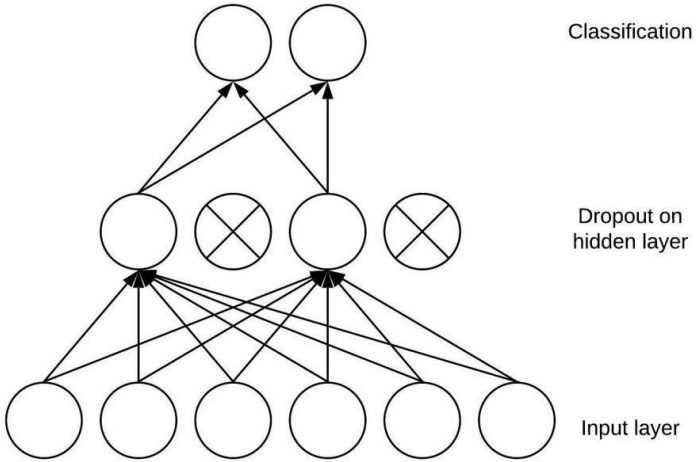
$$y_i \leftarrow \gamma \widehat{x}_i + \beta$$

- N is the batch size
- C is the number of feature channel
- Scaled by γ and shifted by β, which are learnable parameters

DropOut

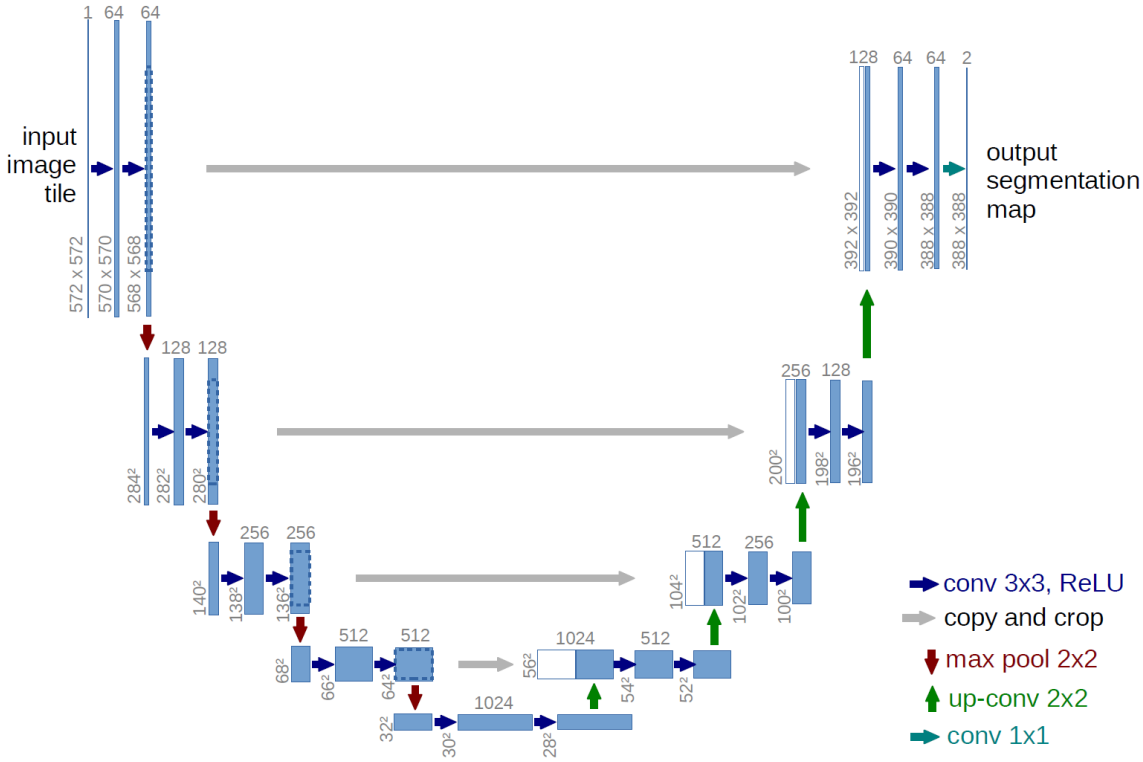


Without Dropout



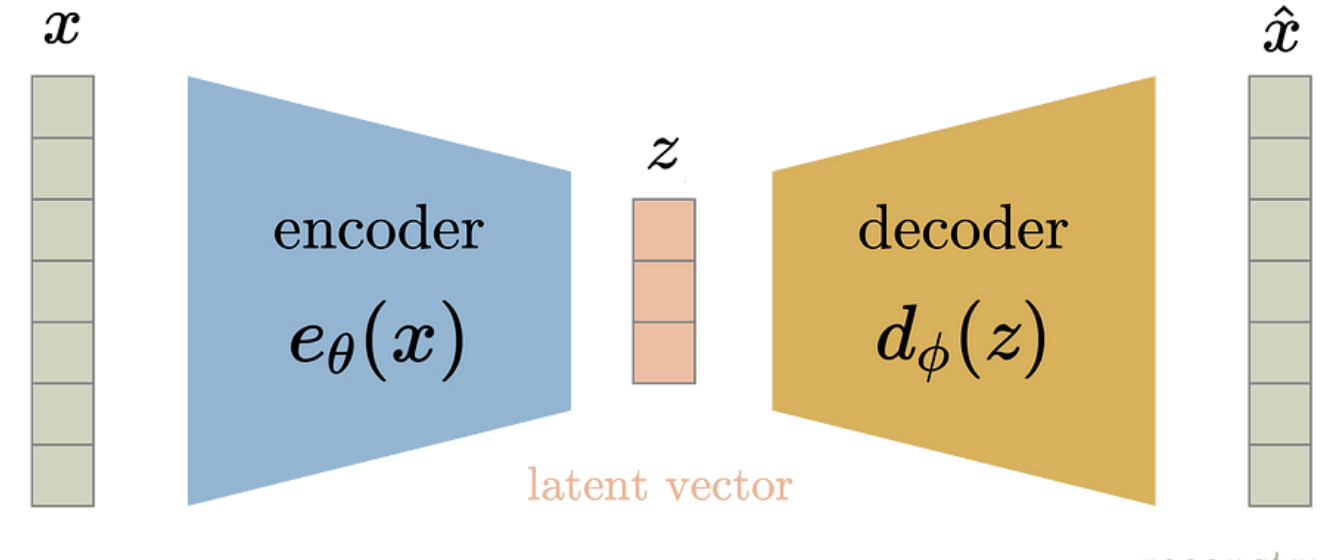
With Dropout

U-Net



[U-Net: Convolutional Networks for Biomedical Image Segmentation]

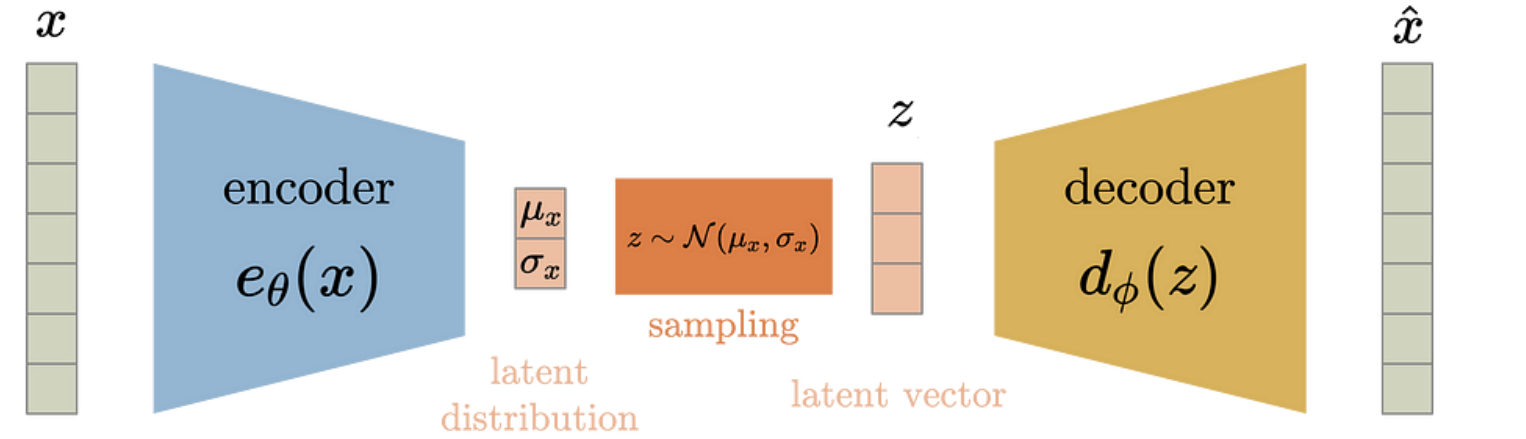
Variational Autoencoder



$$loss = \|x - \hat{x}\|_2 = \|x - d_\phi(z)\|_2 = \|x - d_\phi(e_\theta(x))\|_2$$

[Aqeel Anwar 2021]

Variational Autoencoder



$$\text{reconstruction loss} = \|x - \hat{x}\|_2 = \|x - d_\phi(z)\|_2 = \|x - d_\phi(\mu_x + \sigma_x \epsilon)\|_2$$

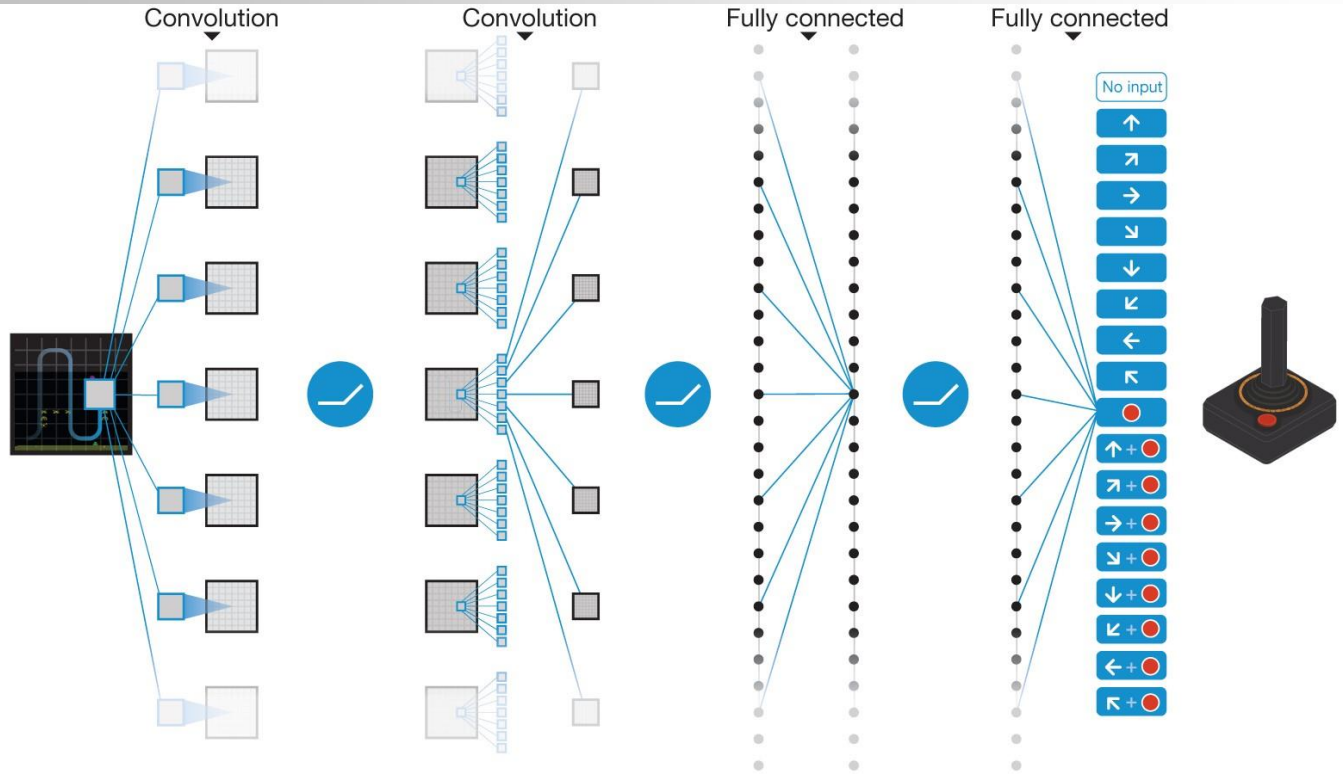
$$\mu_x, \sigma_x = e_\theta(x), \quad \epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$$

$$similarity\ loss = KL\ Divergence = D_{KL}(\mathcal{N}(\mu_x, \sigma_x) \parallel \mathcal{N}(\mathbf{0}, \mathbf{I}))$$

$$loss = reconstruction\ loss + similarity\ loss$$

[Aqeel Anwar 2021]

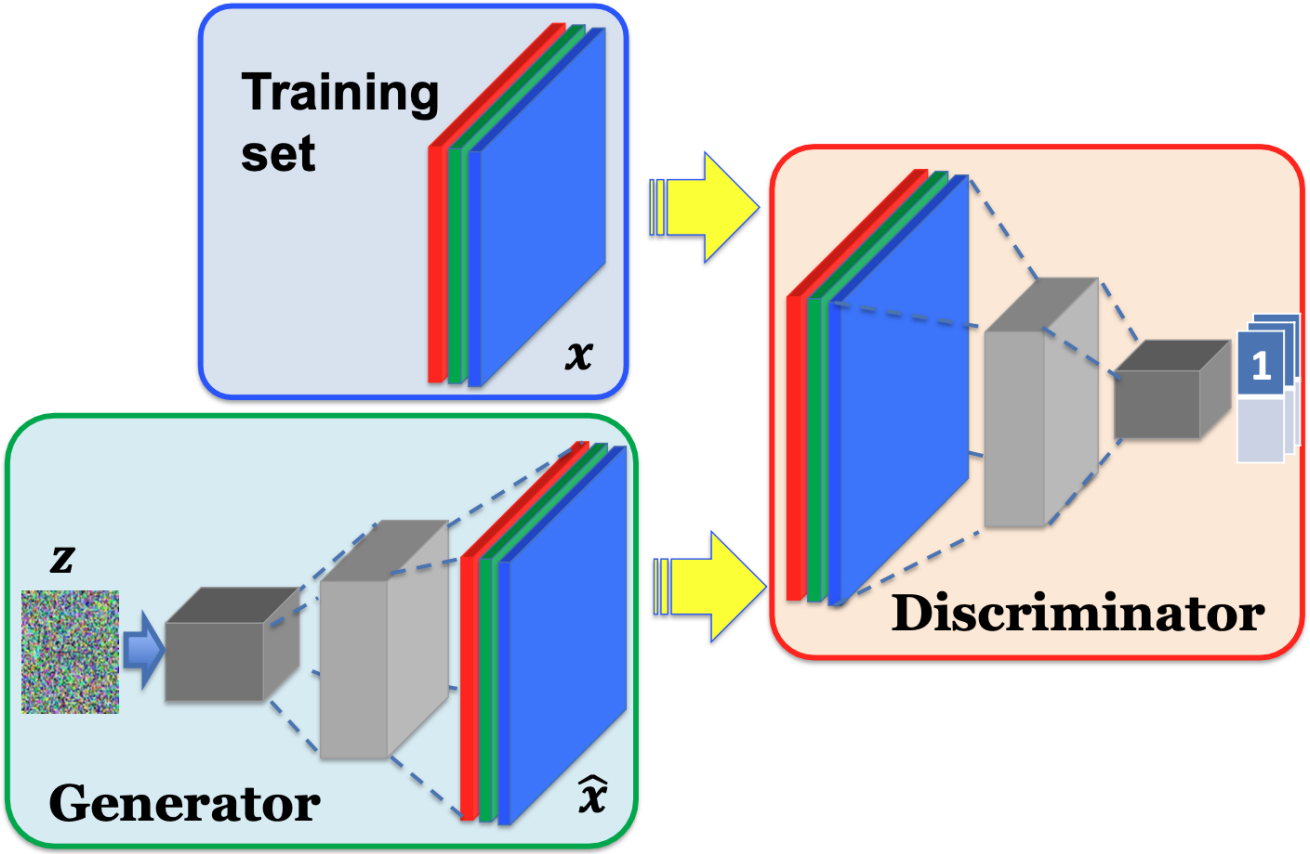
Reinforcement Learning



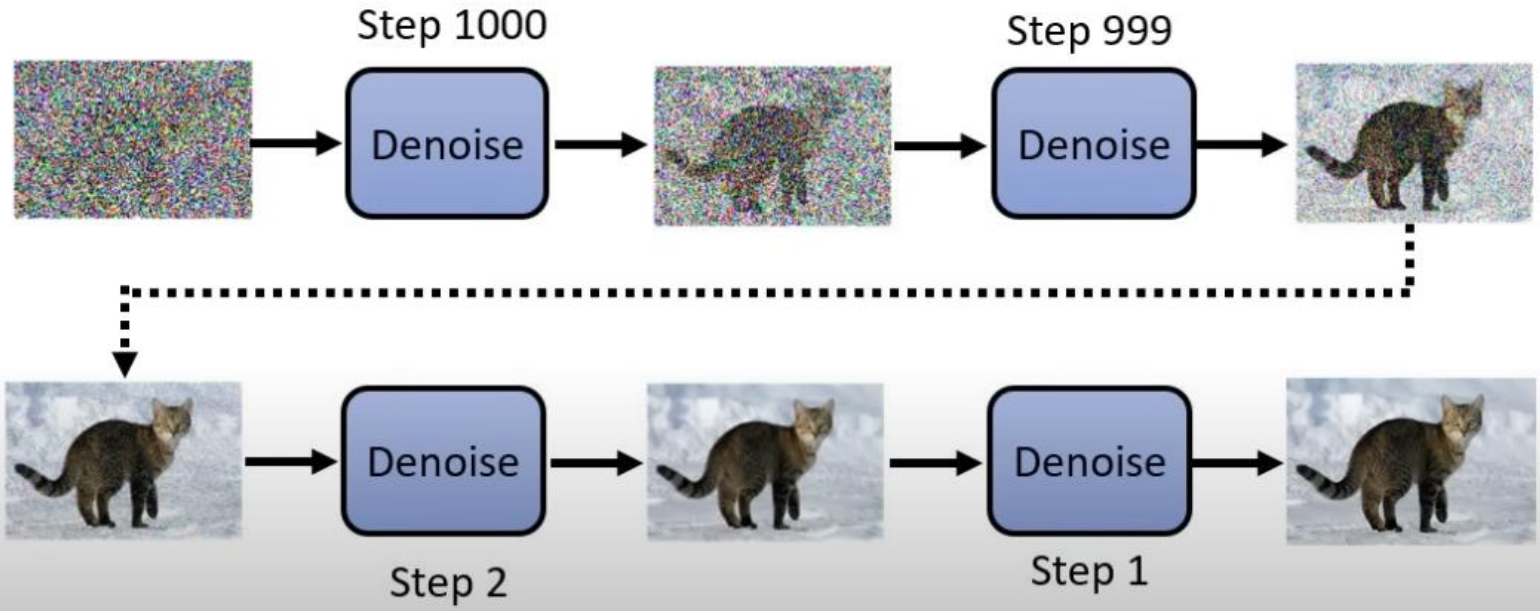
[Mnih, et al 2015]

Generative AI

Generative Adversarial Network

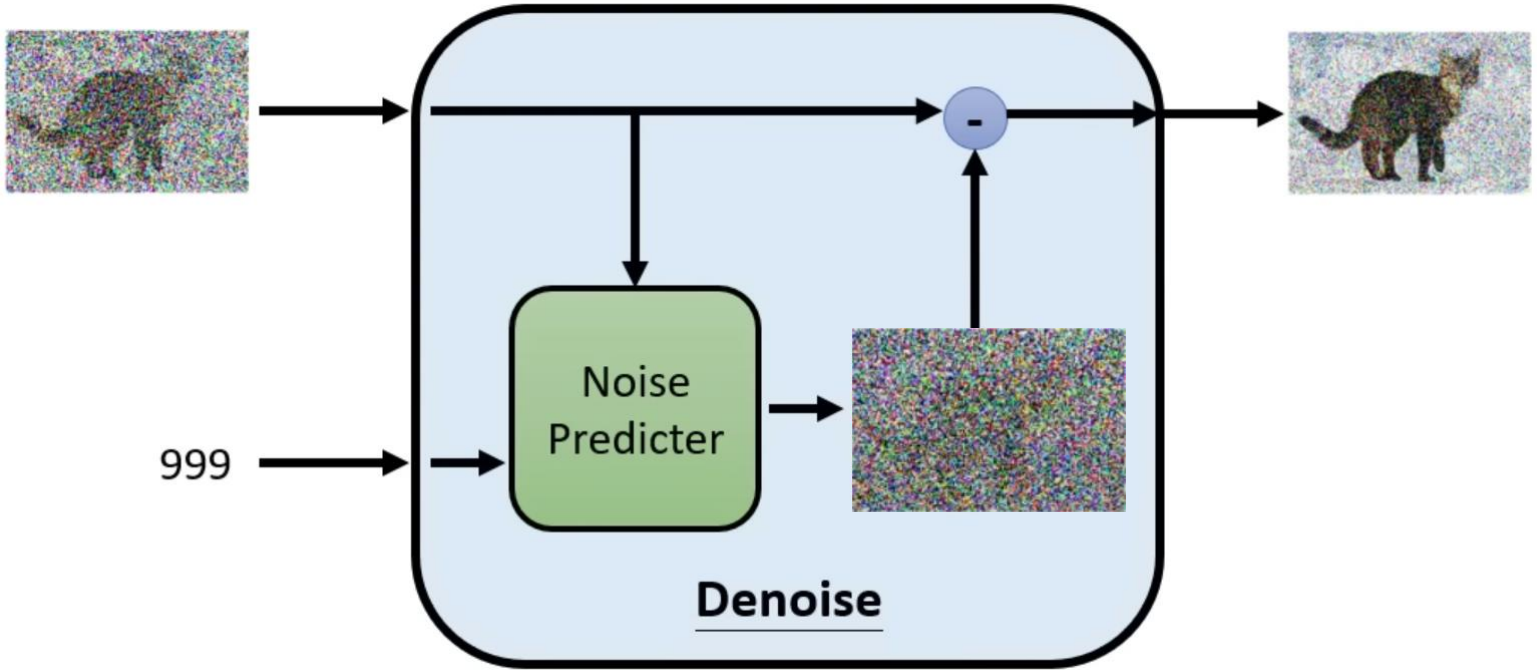


Diffusion Model

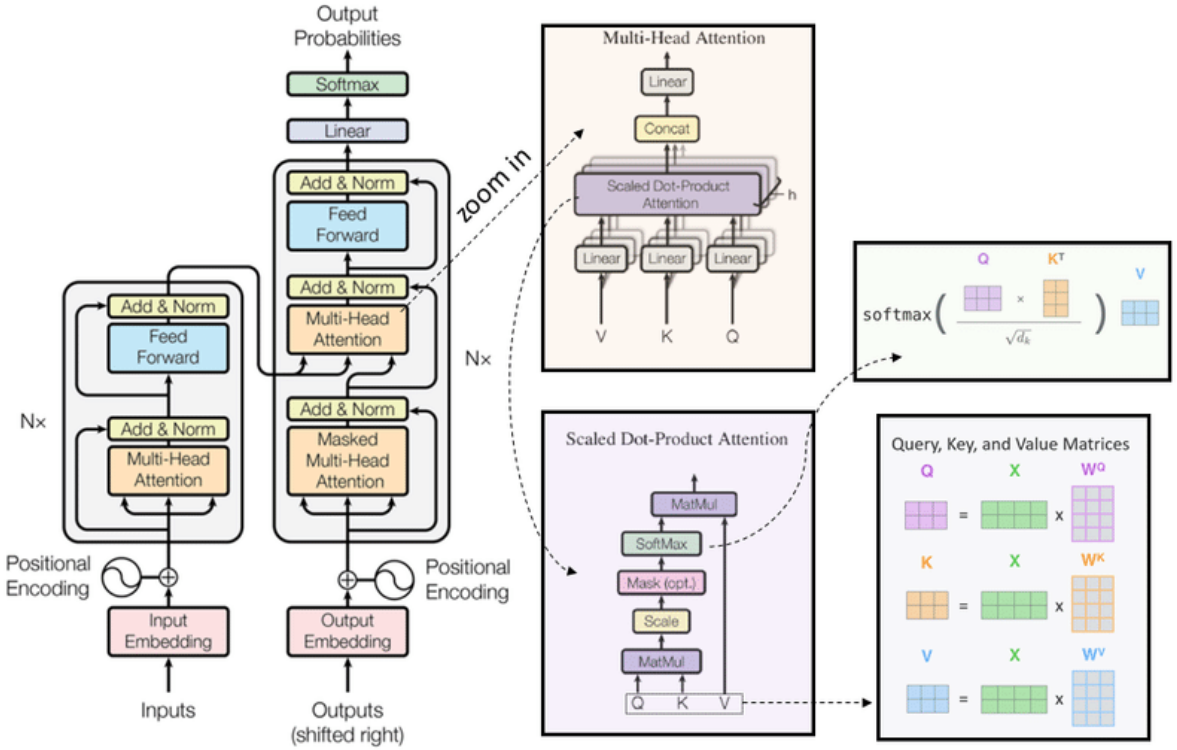


[Denoising Diffusion Probabilistic Models]

Diffusion Model



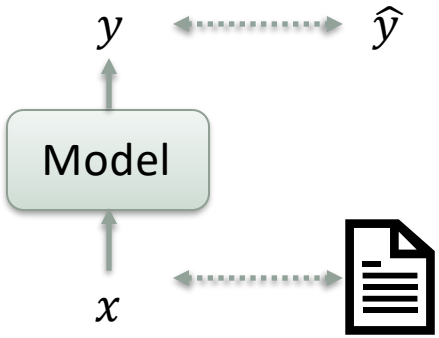
Transformer



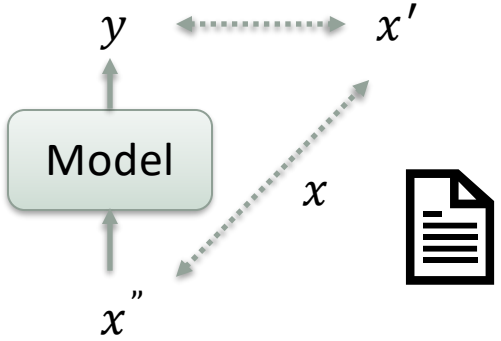
[Jackson Saenz 2022]

BERT– Self-supervised Learning

Supervised



Self-supervised



"I now call it "self-supervised learning", because "unsupervised" is both a loaded and confusing term.
In self-supervised learning, the system learns to predict part of its input from other parts of it input. In other words a portion of the input is used as a supervisory signal to a predictor fed with the remaining portion of the input. – Yan LeCun

Masking Input

Special token

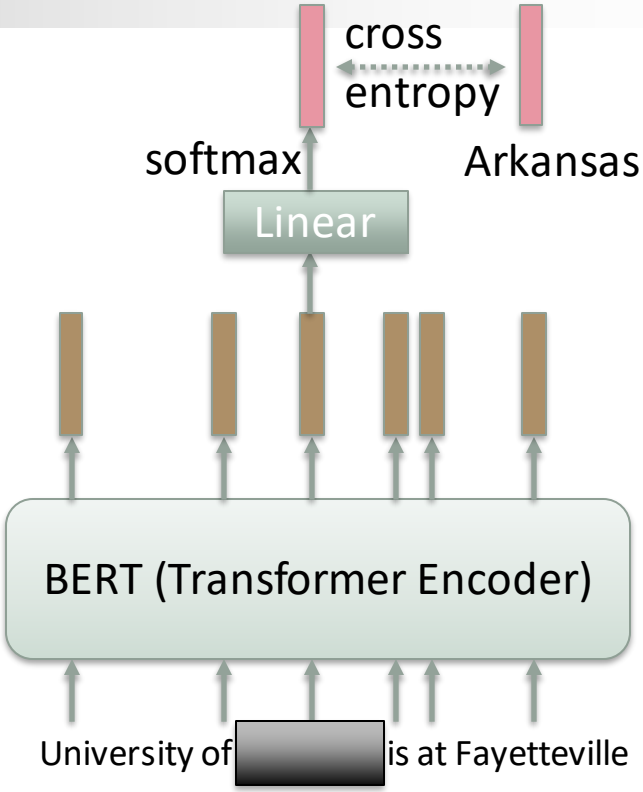
= Mask

or

= Random

Michigan, bird, etc.

Randomly masking some tokens



https://arxiv.org/abs/1810.04805

Mamba: Linear-Time Sequence Modeling with Selective State Spaces

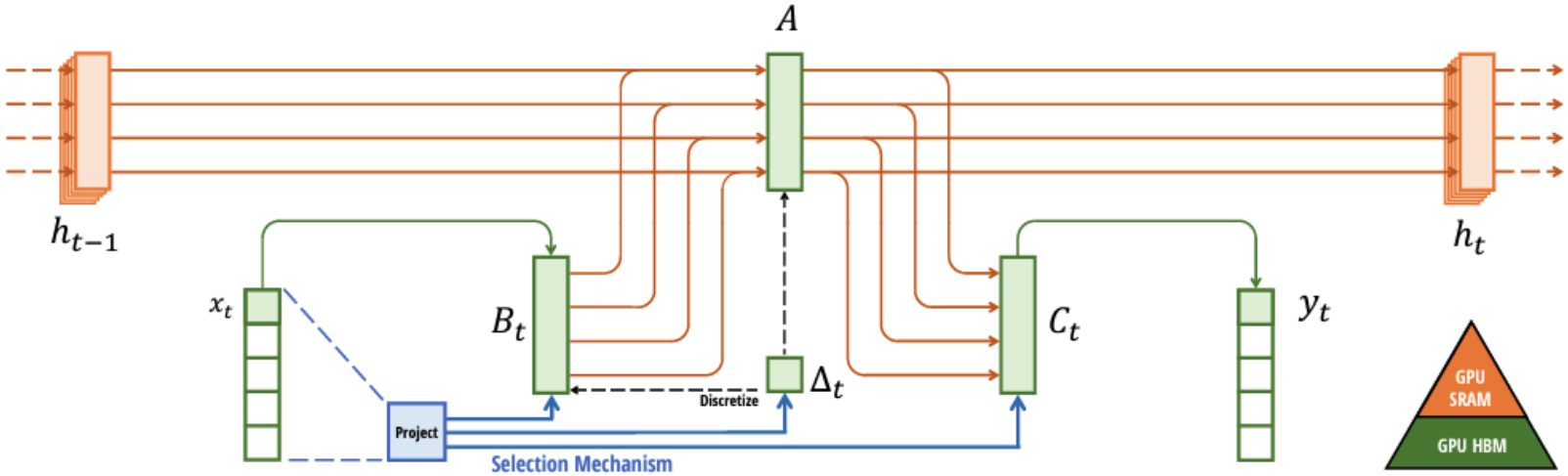
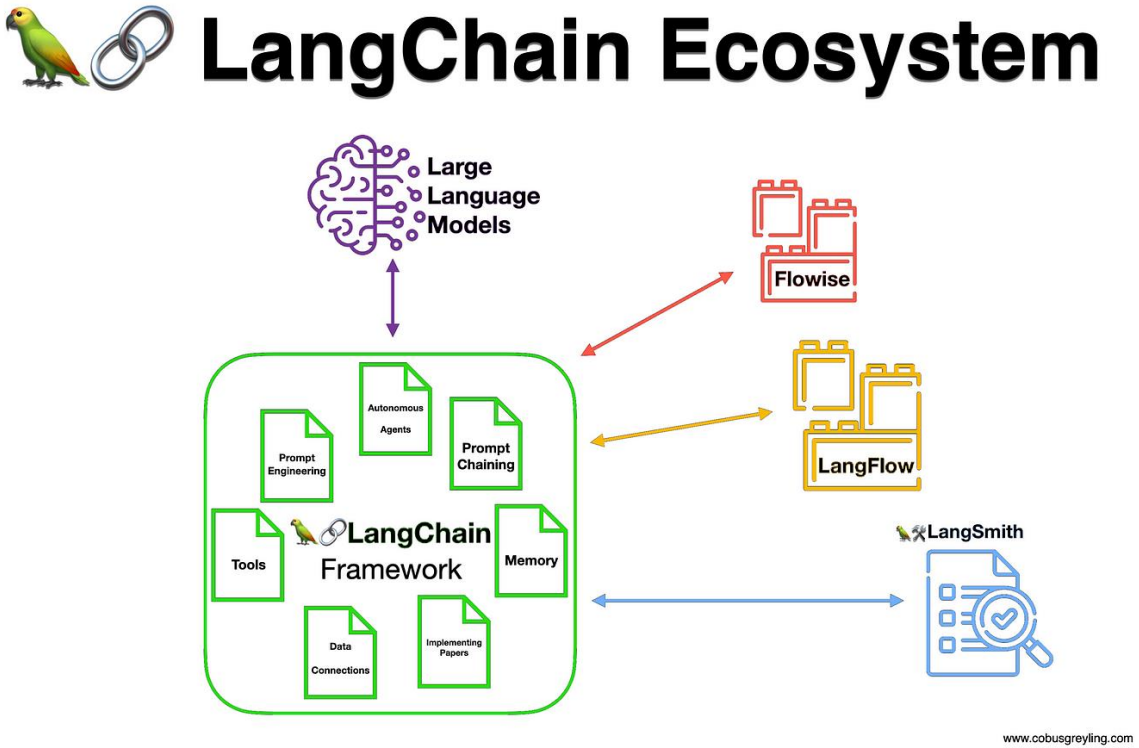


Figure 1: (**Overview**.) Structured SSMs independently map each channel (e.g. $D = 5$) of an input x to output y through a higher dimensional latent state h (e.g. $N = 4$). Prior SSMs avoid materializing this large effective state (DN, times batch size B and sequence length L) through clever alternate computation paths requiring time-invariance: the $(\Delta, \boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C})$ parameters are constant across time. Our selection mechanism adds back input-dependent dynamics, which also requires a careful hardware-aware algorithm to only materialize the expanded states in more efficient levels of the GPU memory hierarchy.

2312.00752v1

LLM – Prompt


LangChain Ecosystem
Large
Language
Models
Flowise
LangFlow
LangSmith
Autonomous
Agents
Prompt
Engineering
Prompt
Chaining
Tools
LangChain
Framework
Memory
Data
Connections
Implementing
Papers
www.cobusgreyling.com