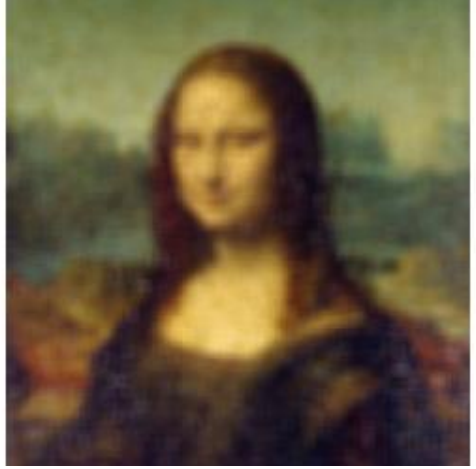
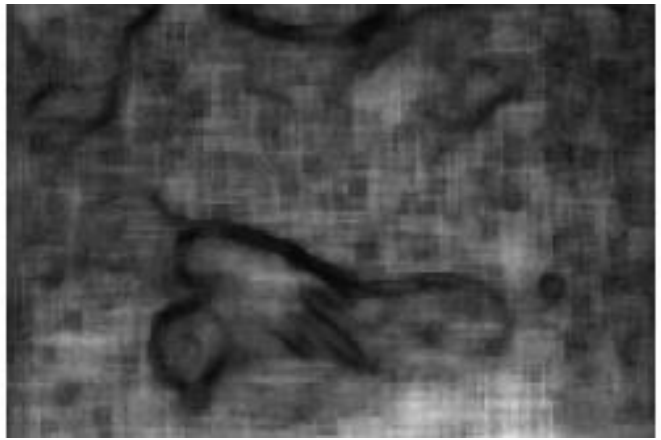
# Computer Vision
and
# Image Processing







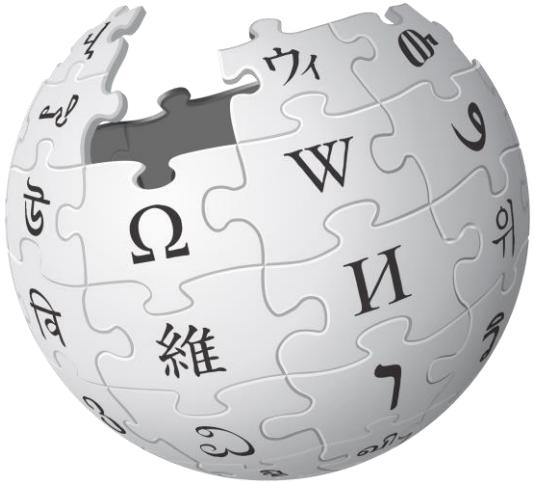


Sanparith.Marukatat@nectec.or.th

IPU / AINRG / NECTEC | AIAT

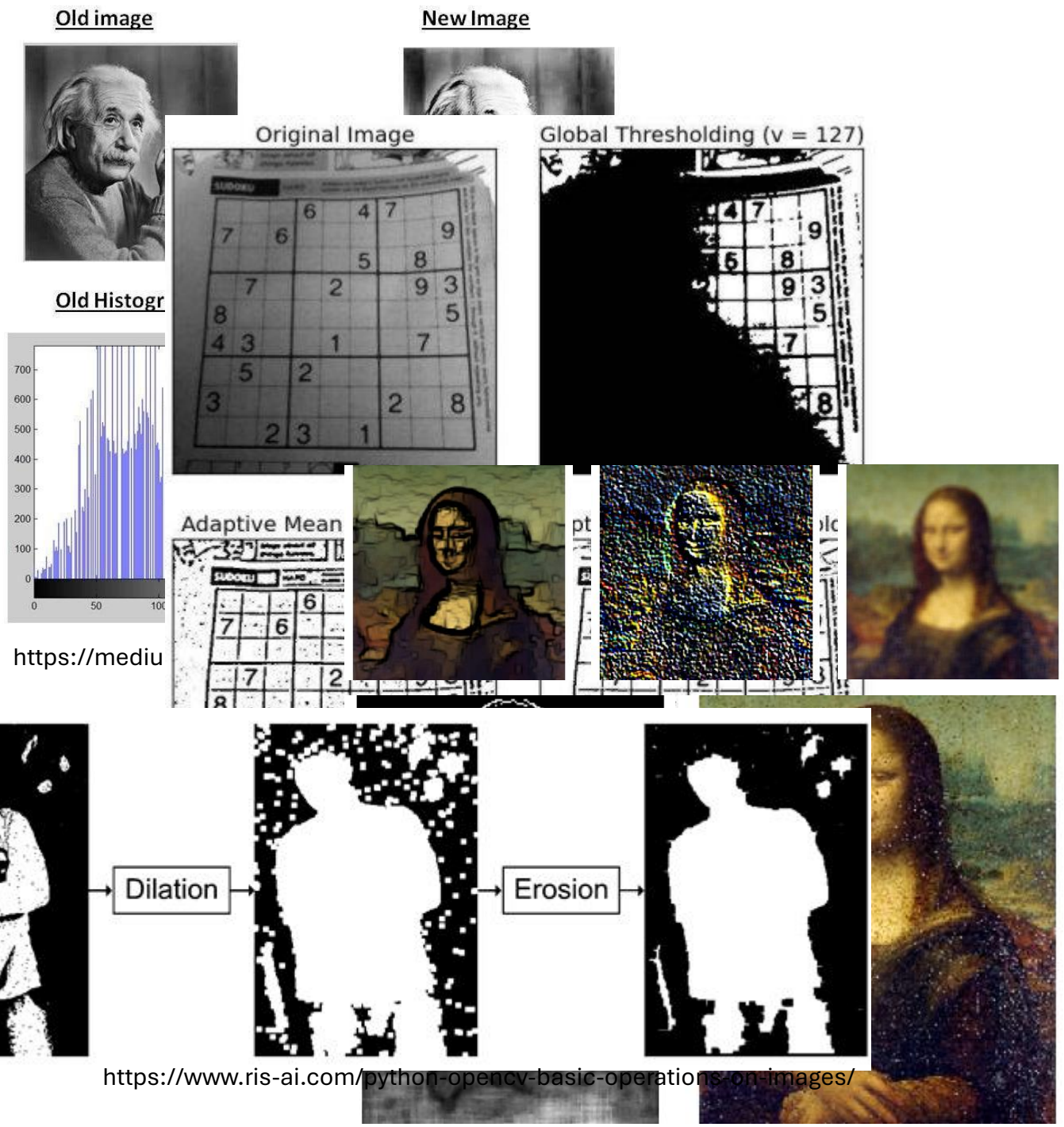
# Definition

- An interdisciplinary field that deals with how computers can be made to gain **high-level understanding** from digital images or videos.
- From the perspective of engineering, it seeks to automate tasks that the human visual system can do.



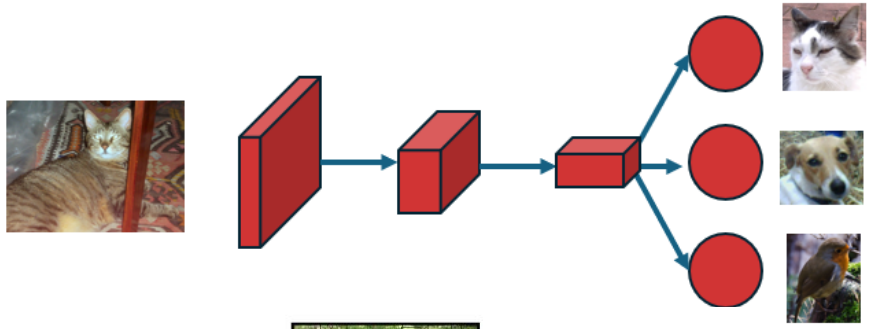
# Tasks (low level)

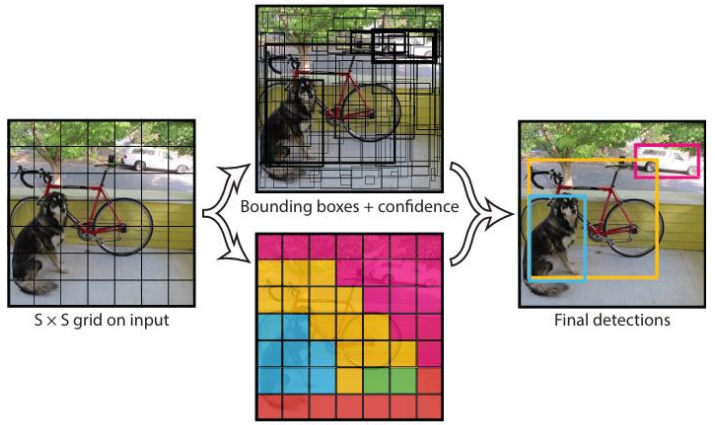
- Intensity transformation
  - Histogram equalization
  - Contrast stretching
- Image binarization (thresholding)
- Image filtering
  - Smoothing
  - Edge detection
  - Sharpening
- Morphological operations
  - Erosion and Dilation
  - Opening and Closing



# Tasks (high level)

- Image classification
- Object detection
  - Tracking
  - Pose estimation







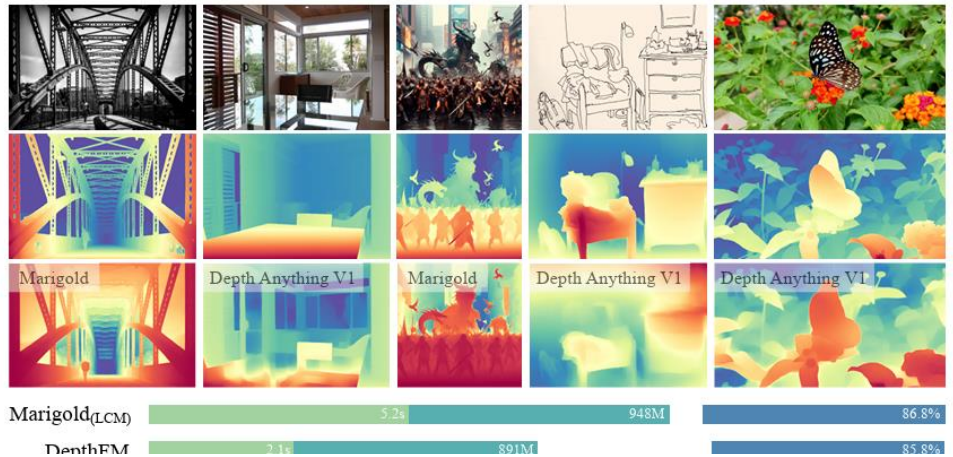
https://opencv.org/blog/multiple-object-tracking-in-realtime/

https://ar5iv.labs.arxiv.org/html/2001.08095

# Tasks (high level)

- Image segmentation
  - Semantic
  - Instance
  - Panoptic
- Depth estimation
- Structure from motion



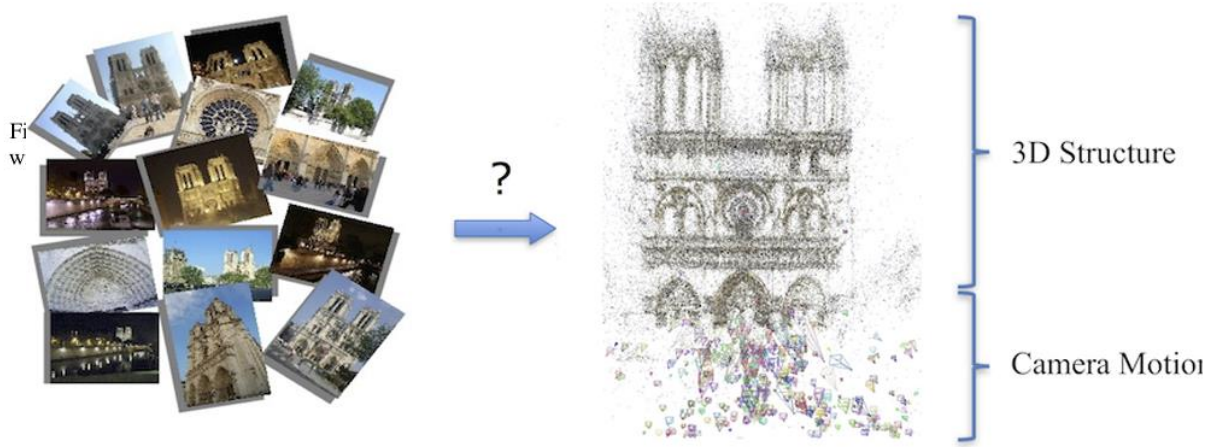




*Figure 1: The structure from motion (SfM) problem.*

https://arxiv.org/abs/1701.08493

# Tasks (high level)

- Image Search
  - Content-Based Image Retrieval
  - Object Search
  - Image Search from text query
- Multi-Modal
  - Image Captioning
  - Visual Question Answering
  - Image Generation from text prompt

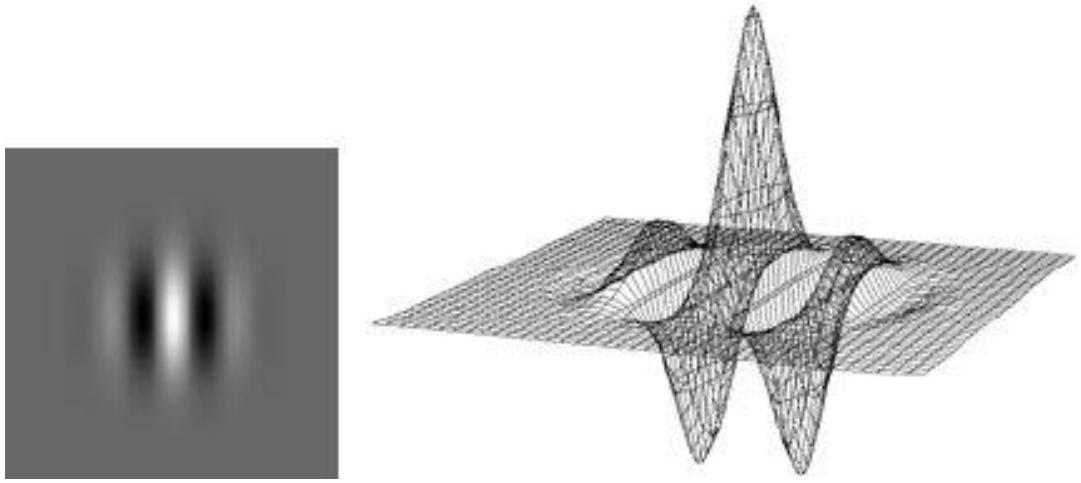
# Tasks (More advanced)

- Training framework
  - Self-supervised learning
  - Imbalanced dataset
  - Labelling error
- Deployment: GPU?

# Techniques

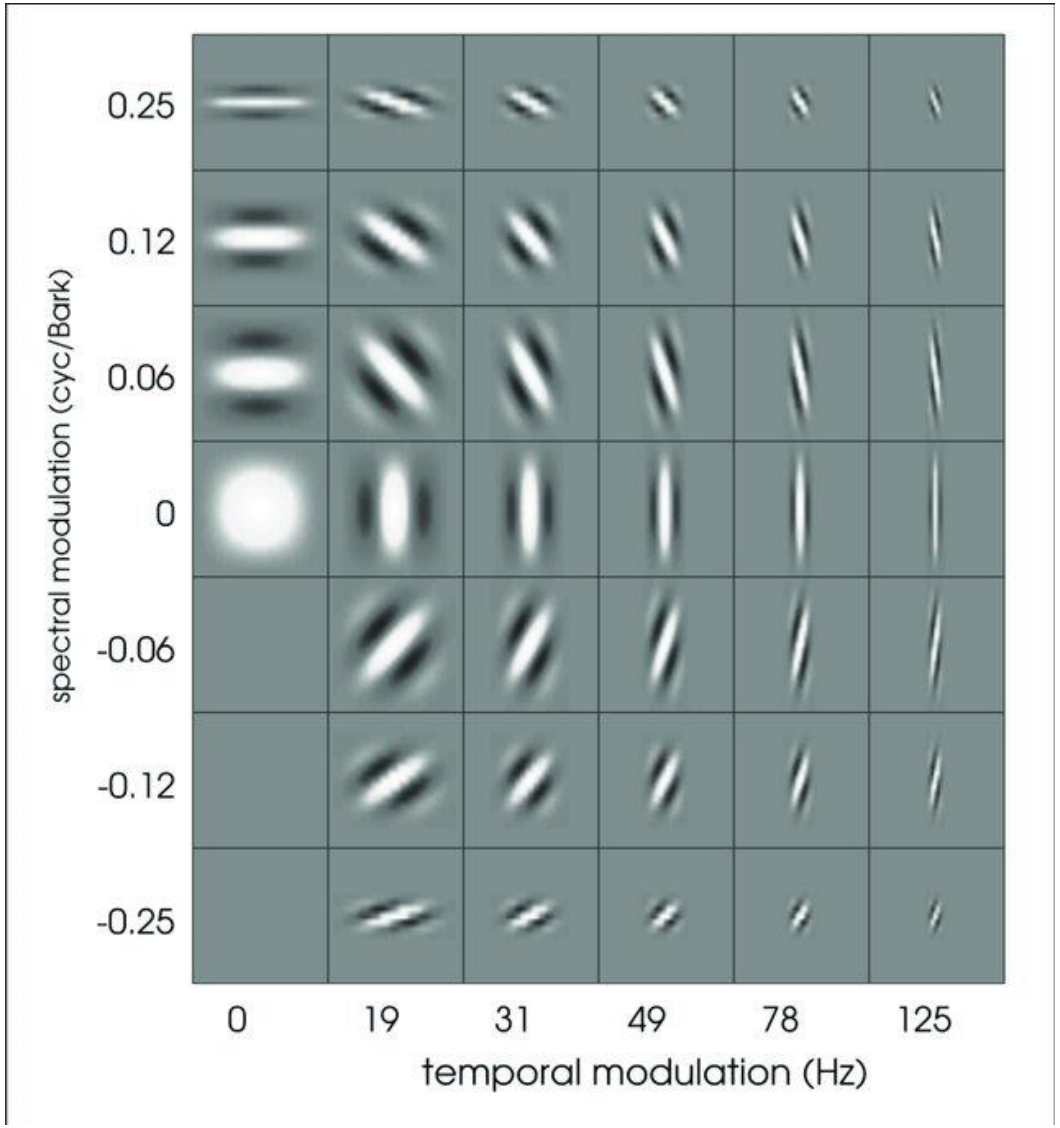
# Classical approaches

- Filtering (spatial domain)
- Frequency domain



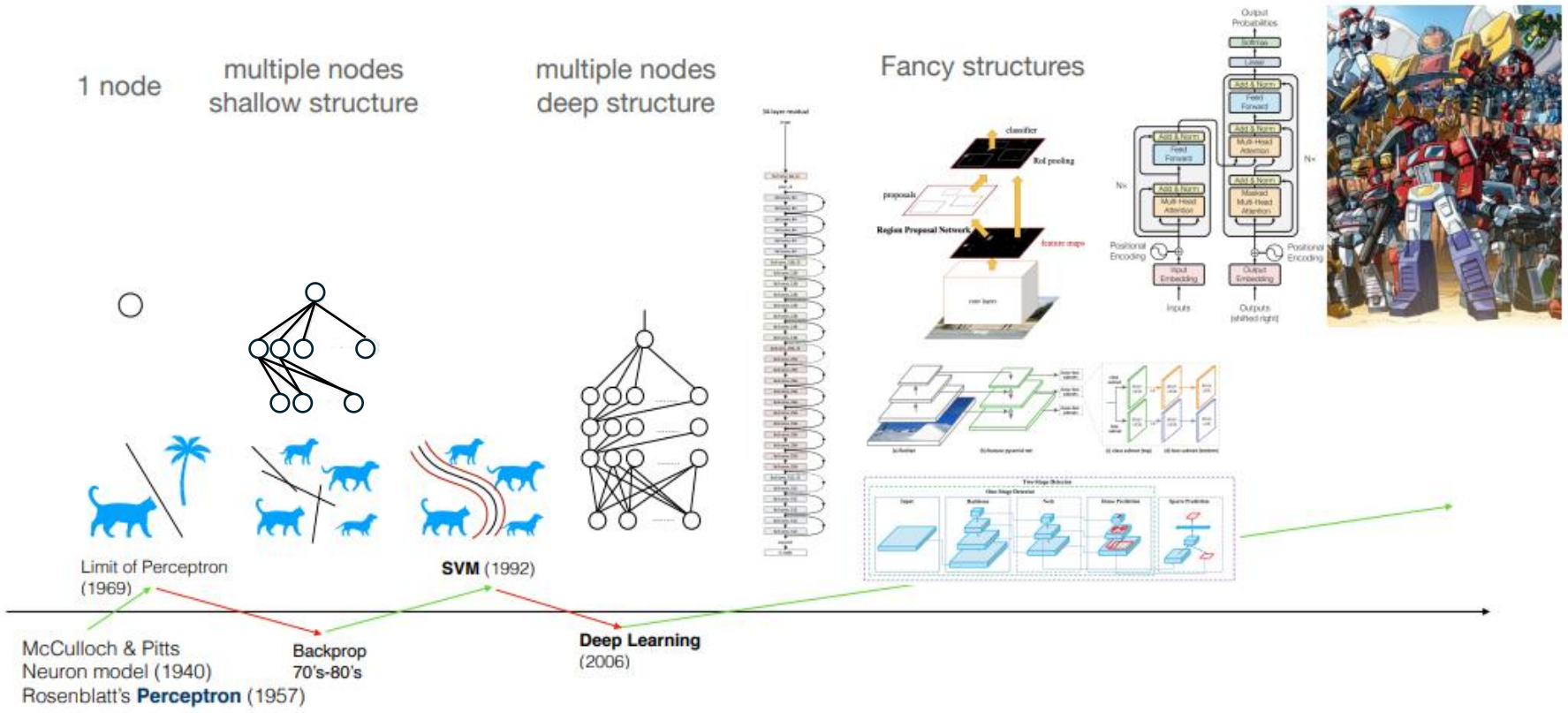
Gabor filter

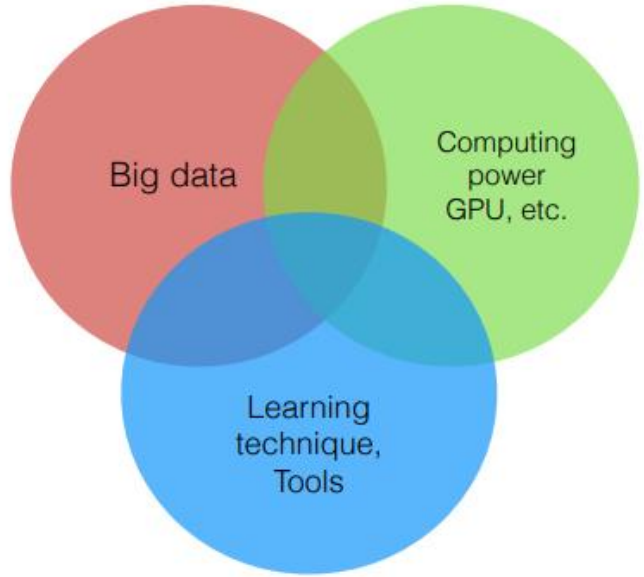
https://www.researchgate.net/publication/228411540_Directional_Adaptive_WSSG_Filter/figures?lo=1

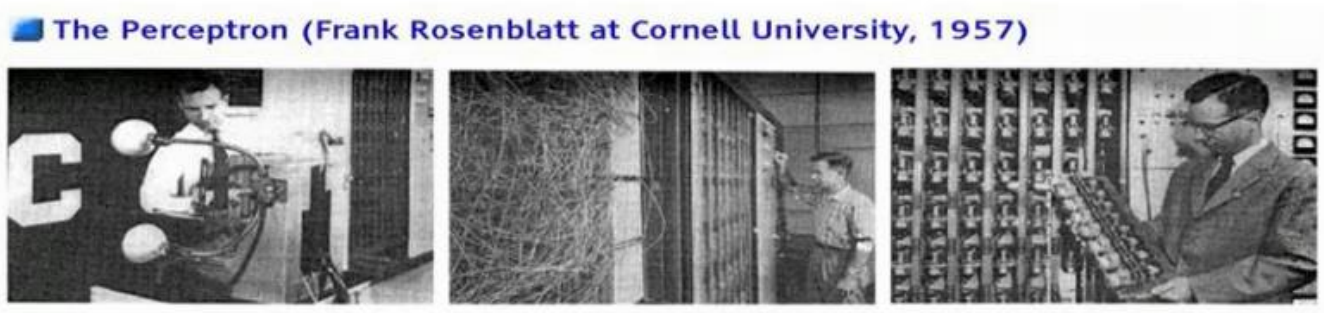


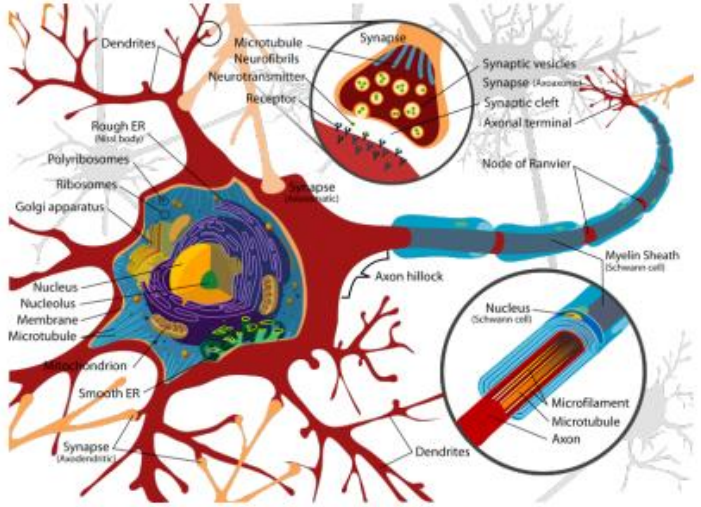

https://www.researchgate.net/publication/313489901_Matching_Pursuit_Analysis_of_Auditory_Receptive_Fields%27_Spectro-Temporal_Properties/figures?lo=1

# Neural Network timeline



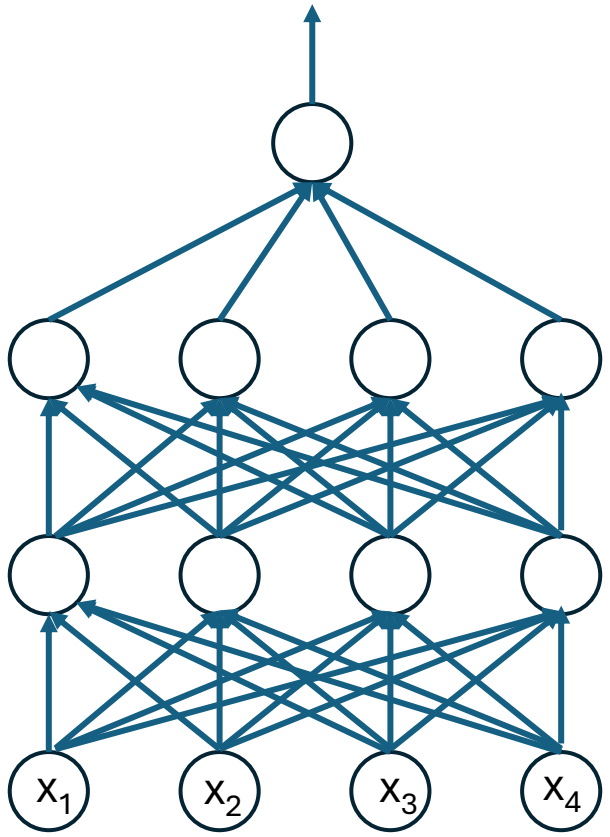




http://en.wikipedia.org/wiki/Neuron

# Basic NN

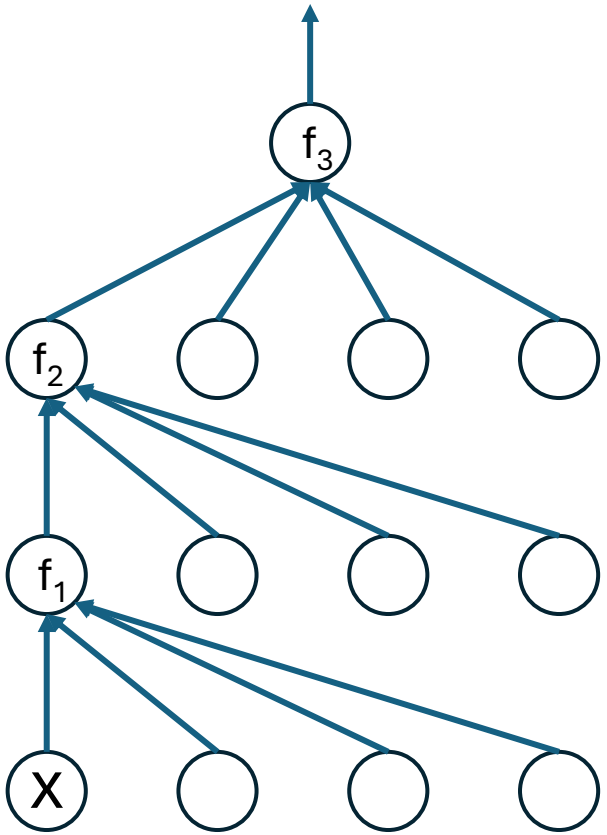
- Gradient of a function
  - Vector of partial dervatives of that function
  - Direction of changes that increases the value of the function
- Loss function
- Gradient descent
- Learning rate



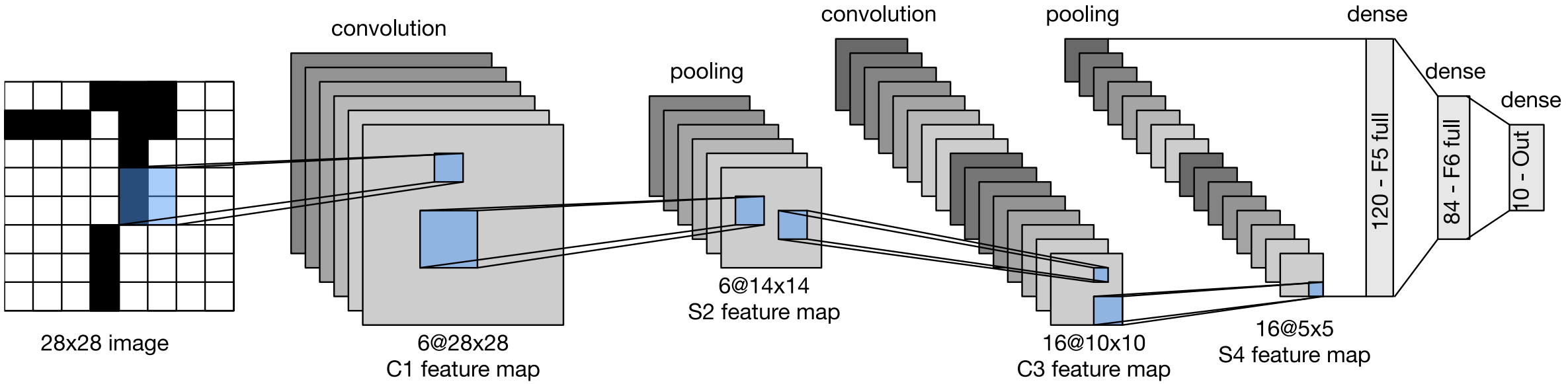
# Gradient Vanishing problem

$f_3(f_2(f_1(x)))=g(x)$

$g'(x)=f'_3(f_2(f_1(x)))\ f'_2(f_1(x))\ f'_1(x)$

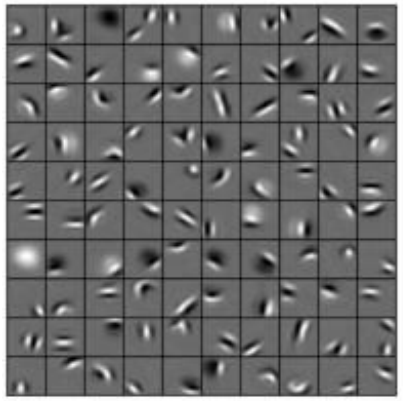


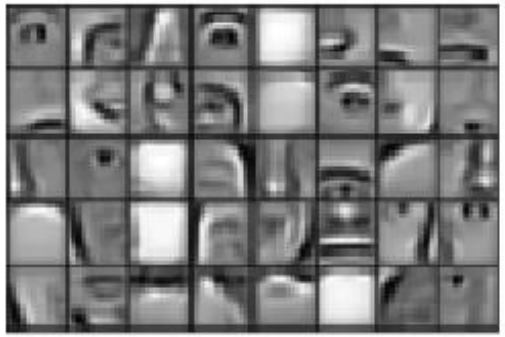
# Gradient vanishing solution 1: CNN



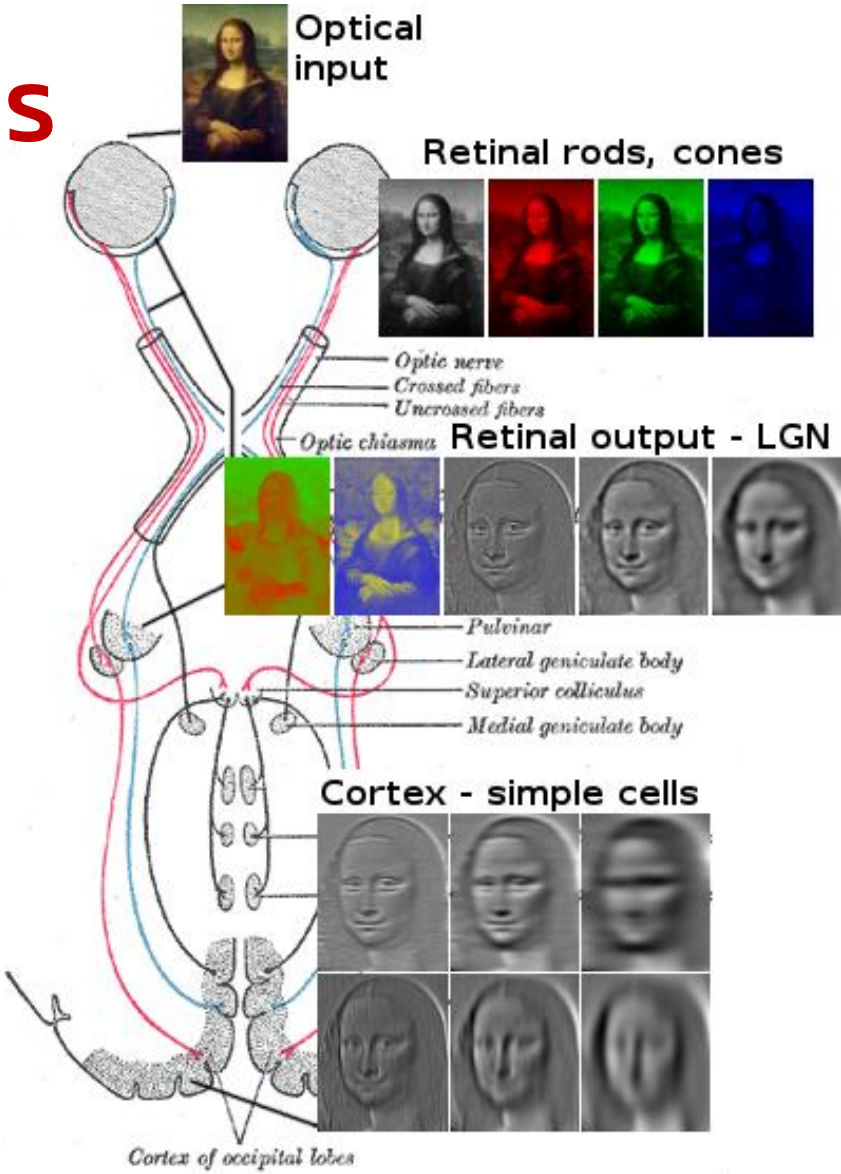
# Convolutional Neural Networks

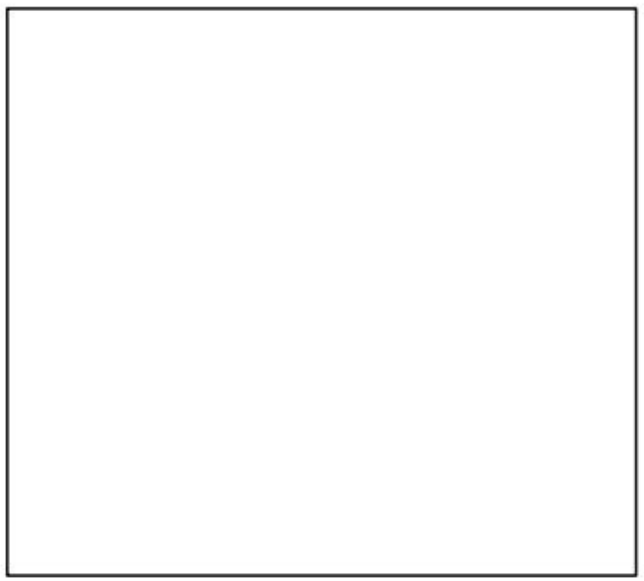
- Classical machine learning approach is composed of
  - Feature extractor
  - Classifier
- Deep learning extracts features
  - Less heuristic
  - Hierarchical representation



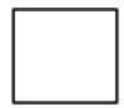


https://developer.nvidia.com/blog/deep-learning-nutshell-core-concepts/



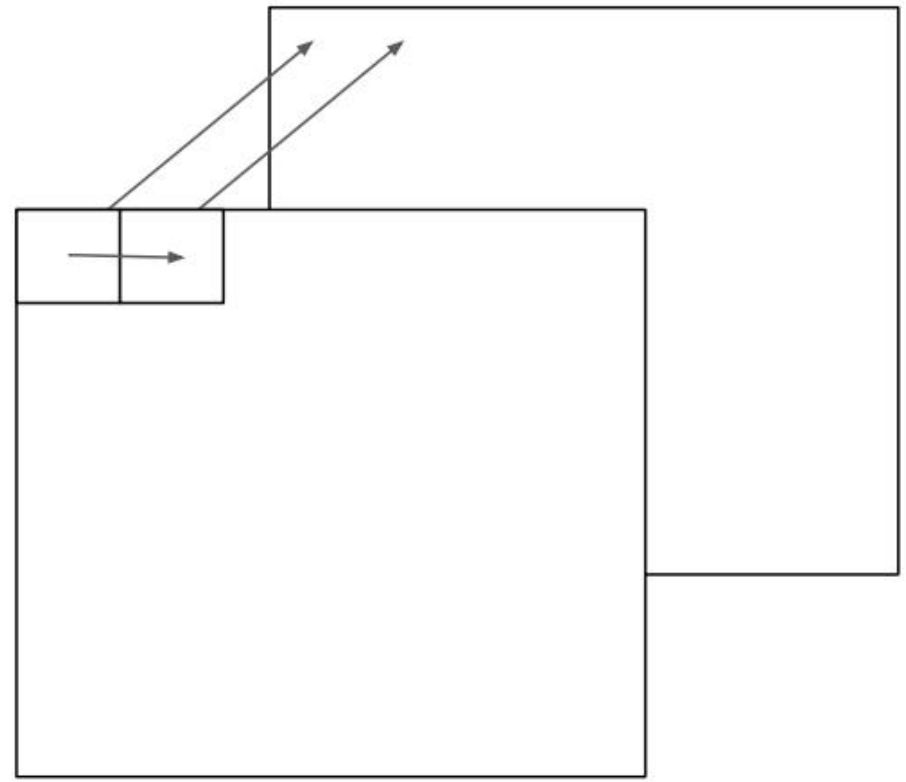


Input image



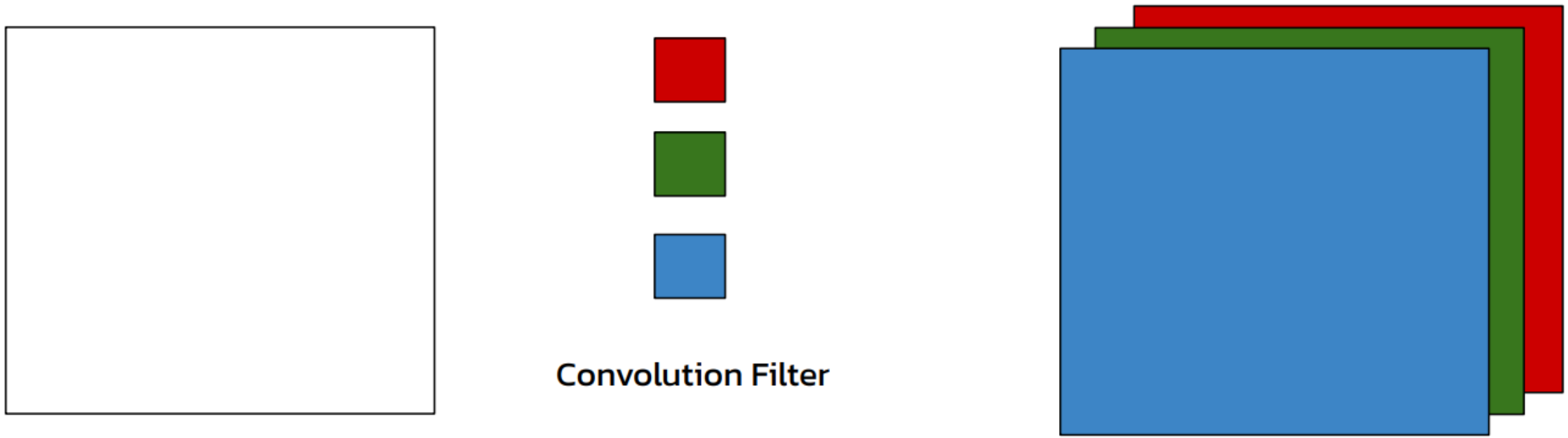
Convolution Filter

Feature map



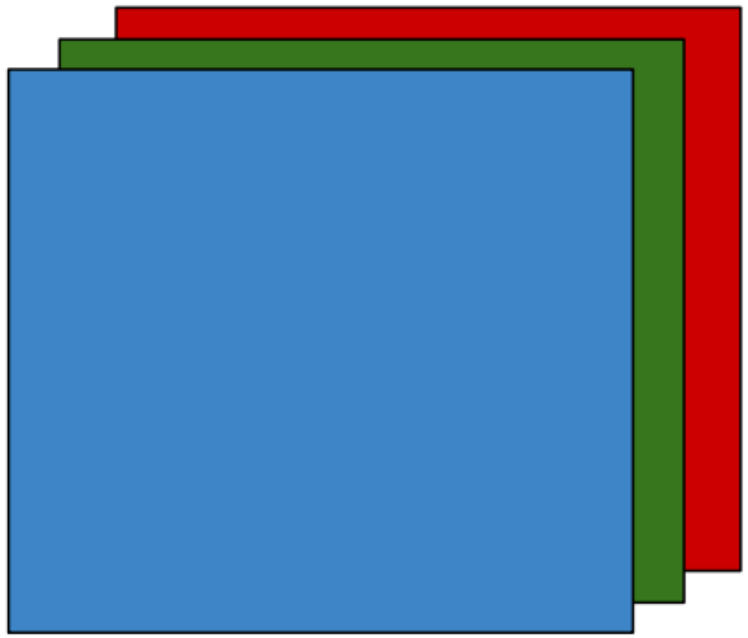
Input image

Feature map



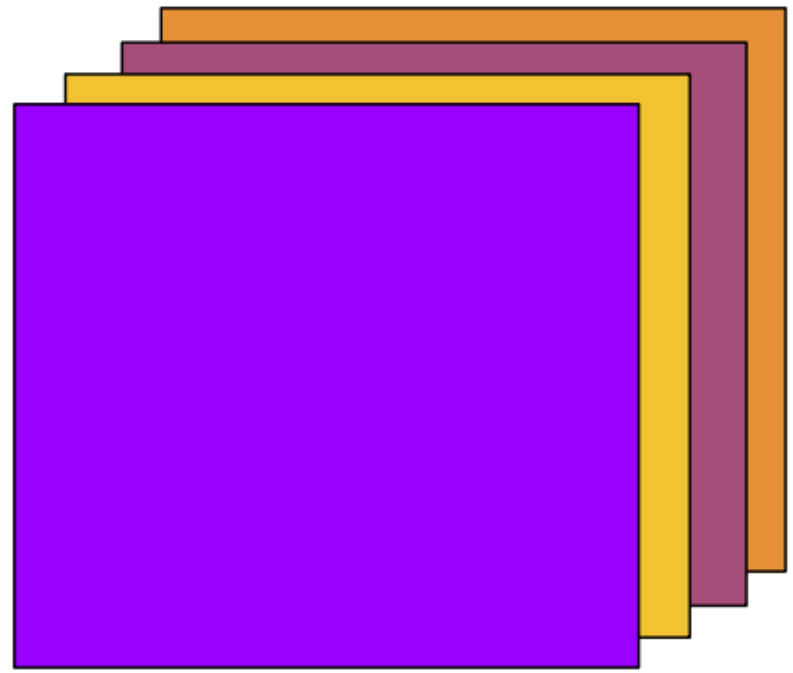
Input image

Convolution Filter



Input feature maps

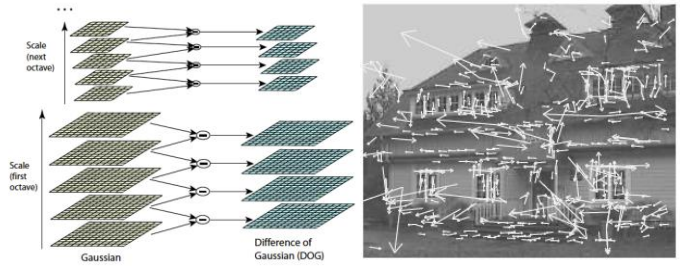
Convolution Filter

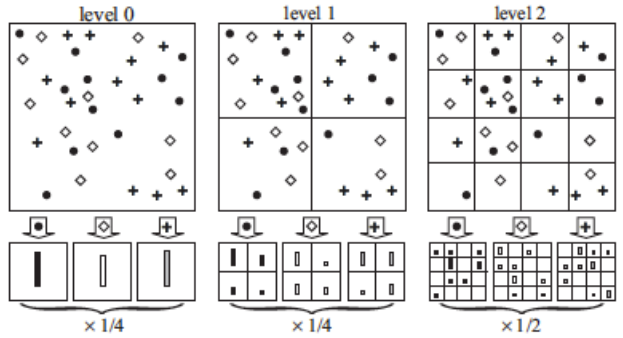


Output feature maps

- In the past, convolution filters needed to be prepared first
- Now, convolution filters = Weight of a neural network
- Weight sharing strategy
  - Need to sum all gradient before update
    - Why this helps gradient vanishing?

# Deep Features




- GIST (global feature) + SVM (RBF): 85.57%
- **SIFT** (local feature) + BoF + SVM (Histogram intersection): 89.69%
- SIFT + **SPM** (spatial pyramid matching) + **LLC** (locality-constrained linear coding) + SVM (linear): 91.48%
- **CNN (AlexNet trained on other dataset)** + SVM (linear): 93.58%

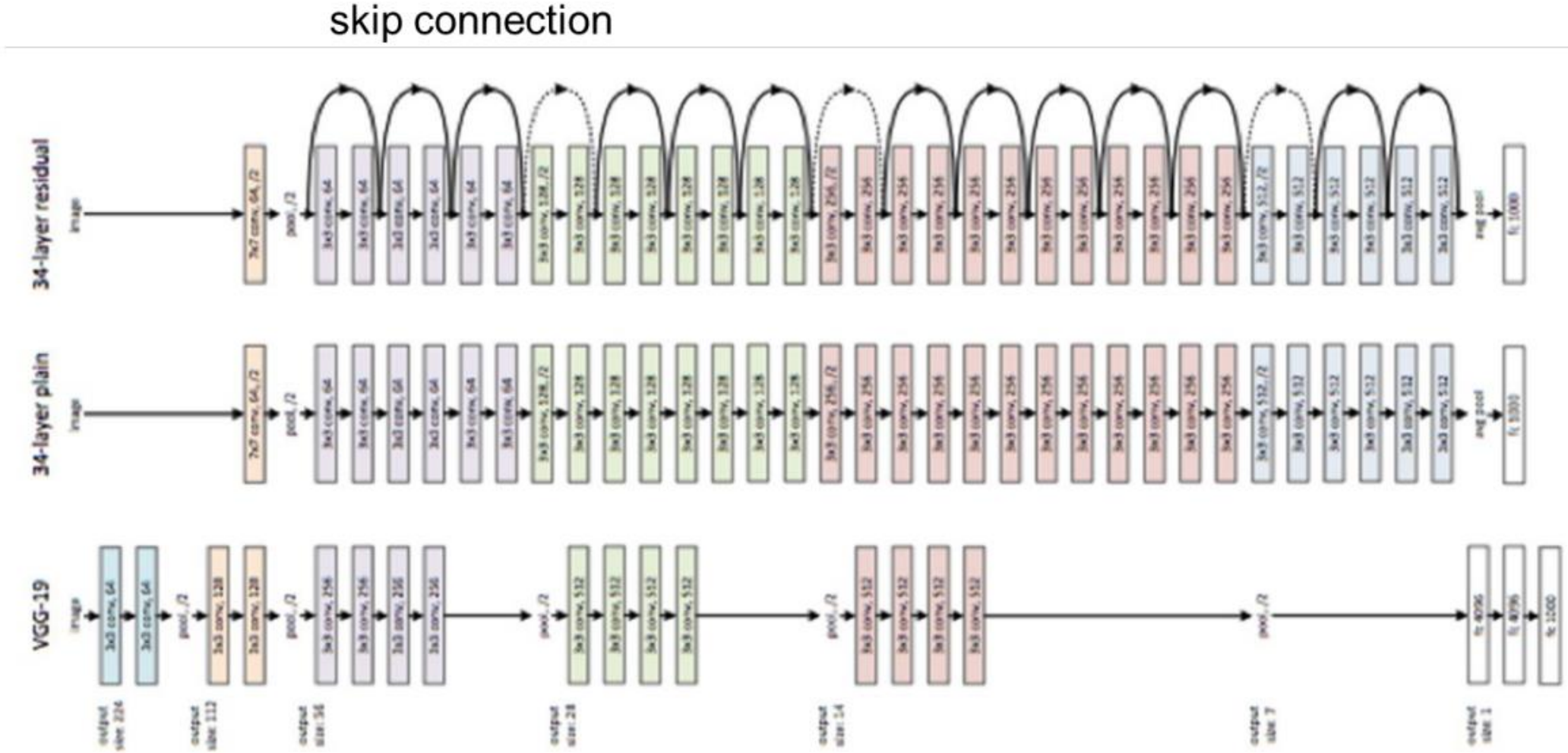
$$\min_C \sum_{i=1}^{M} \|x_i - Bc_i\|^2 + \lambda \|d_i \odot c_i\|^2$$

$$\text{subject to } 1^T c_i = 1$$

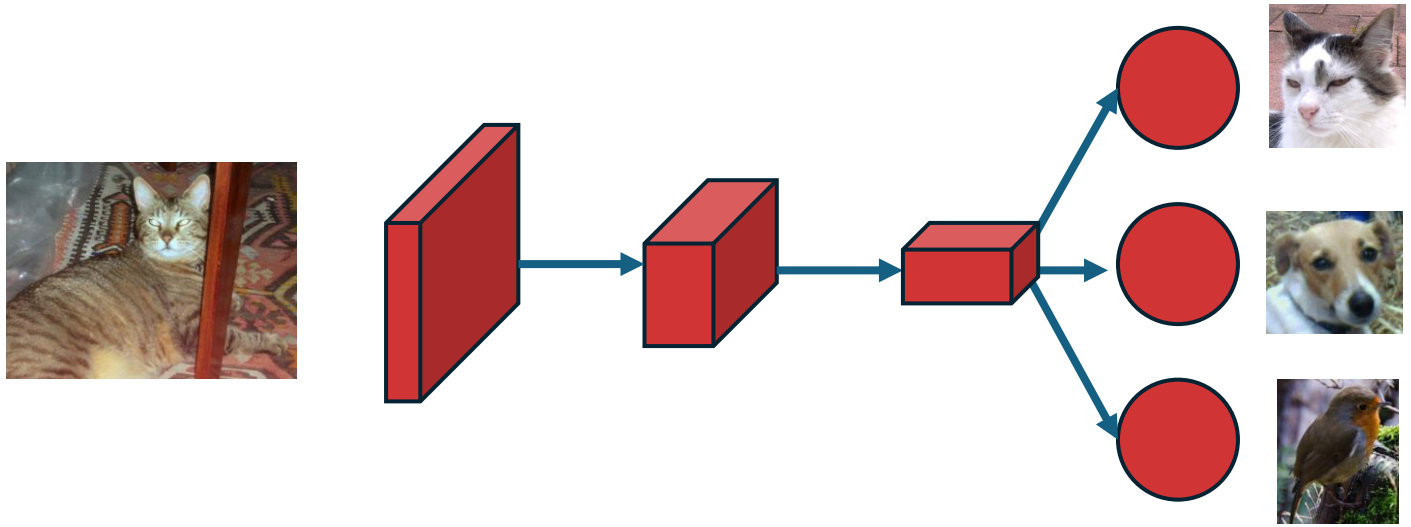
$$d_{im} = \exp\left(\frac{1}{Z\sigma}\|x_i - b_m\|^2\right)$$

$$Z = \max_k \|x_i - b_k\|^2,$$

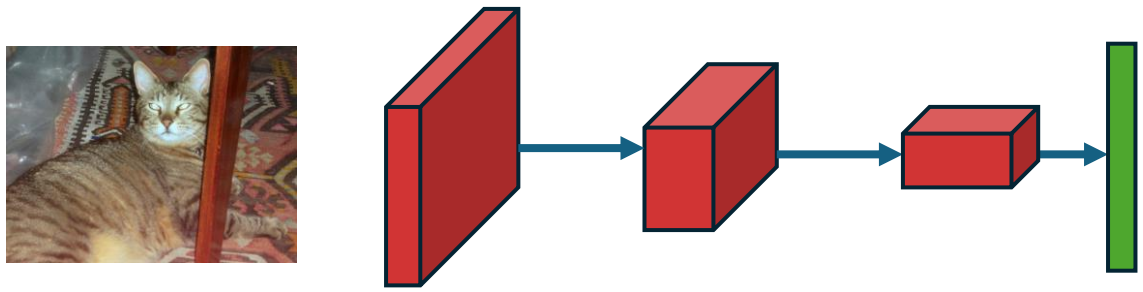
# Gradient Vanishing solution 2: ResNet / skip connection



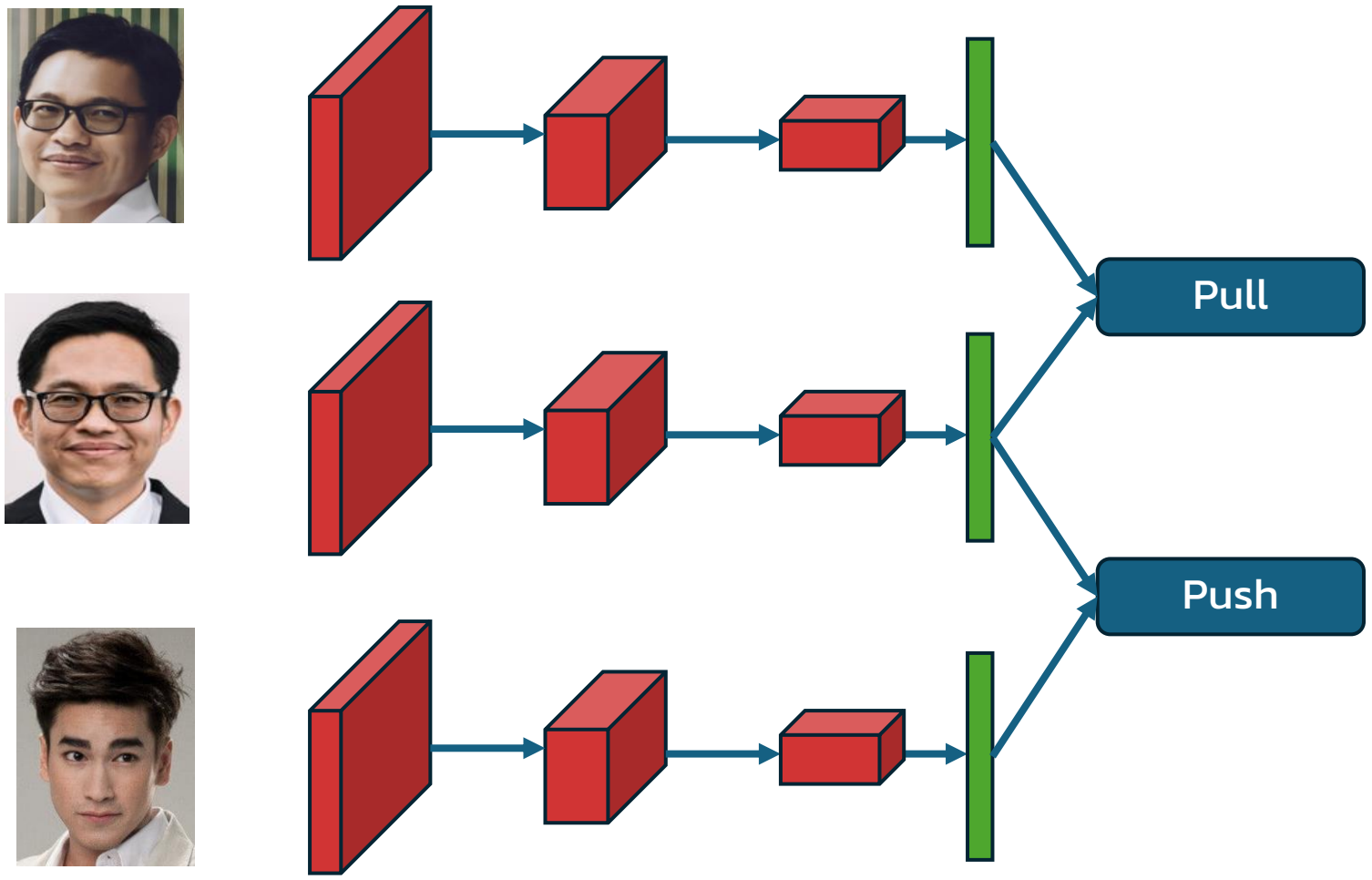
# Some common architectures & tasks



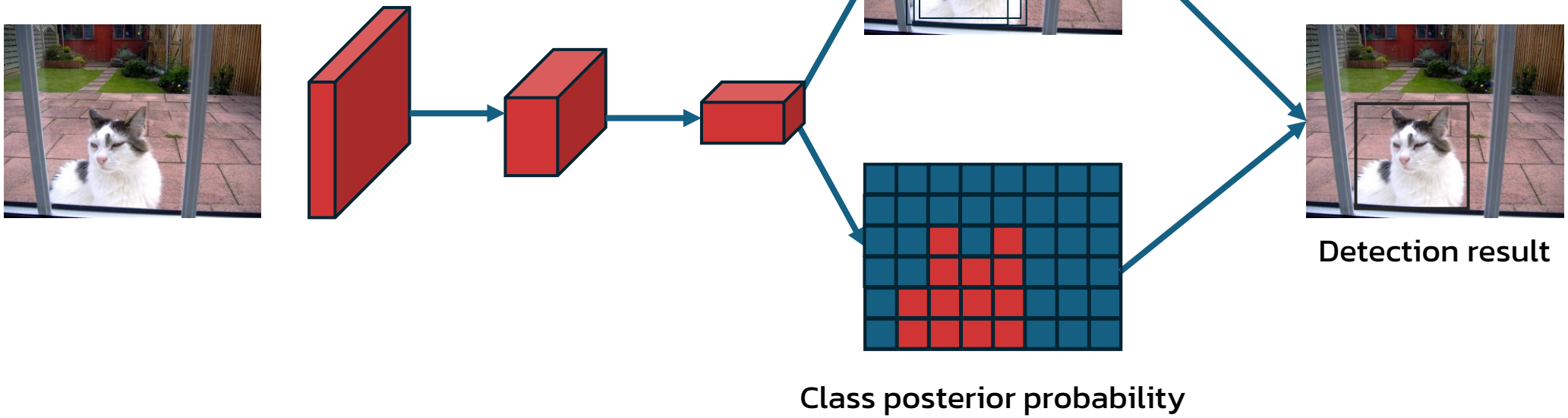
# Classical CNN for Image Classification
# Chest X-ray, Thalassemia



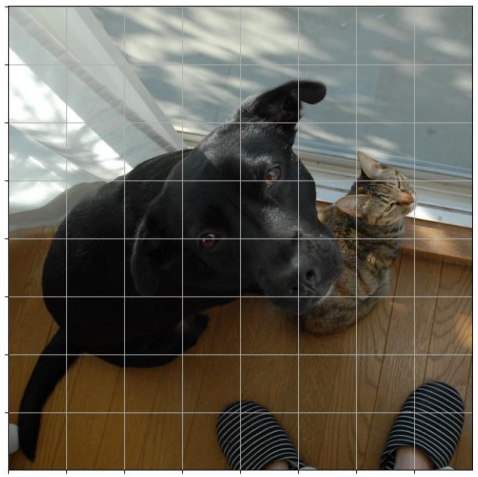
# Classical CNN for Feature Extraction

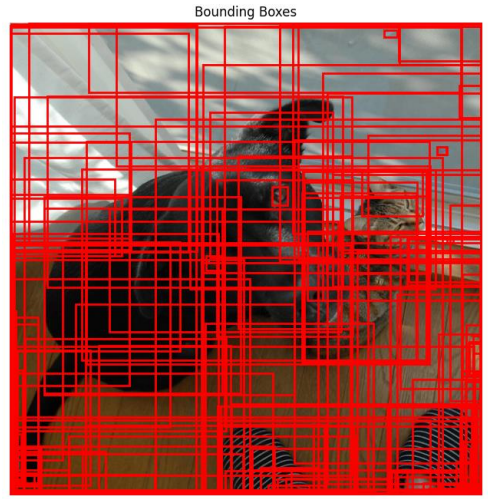


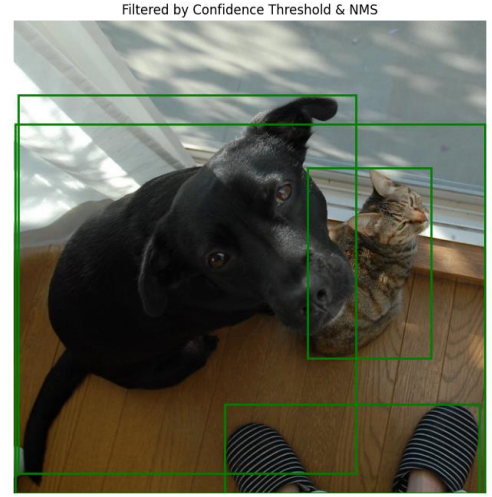

Feature for Face Verification

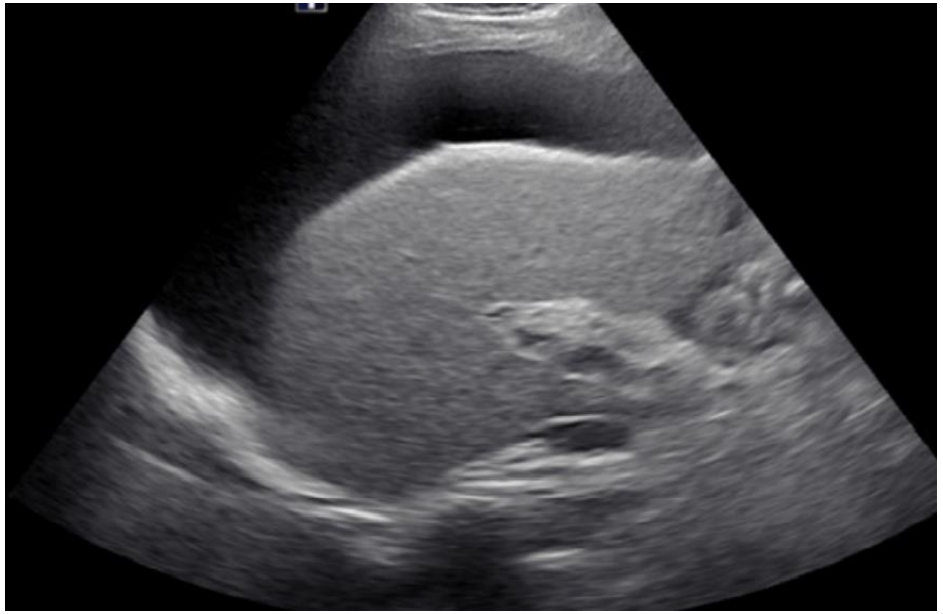


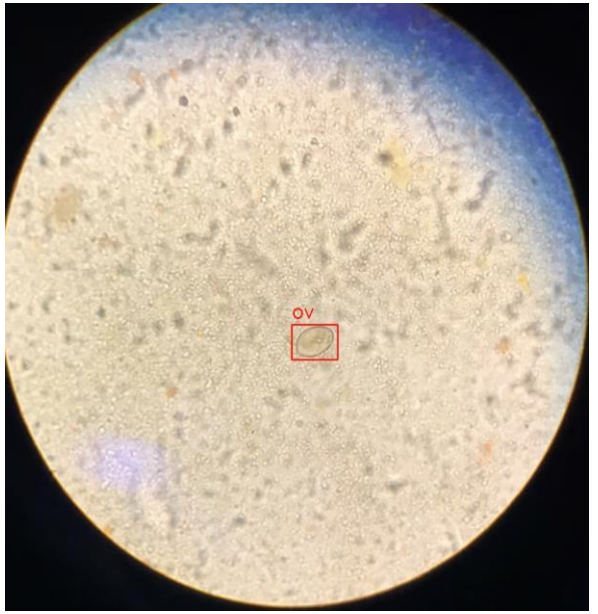

# YOLO for Object Detection



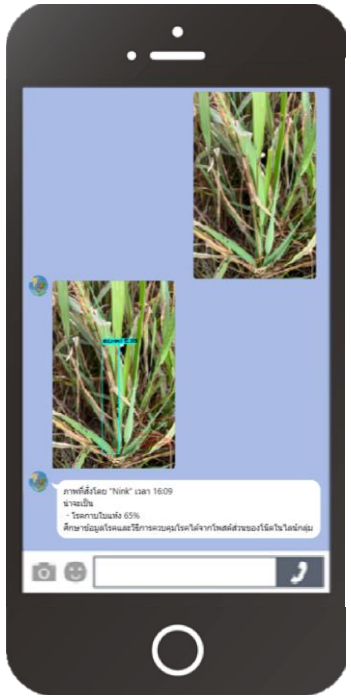

Bounding Boxes


Filtered by Confidence Threshold & NMS

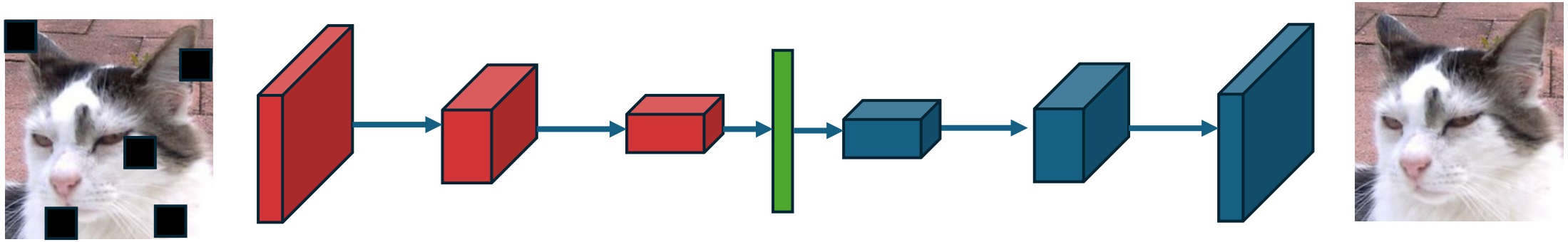



ov

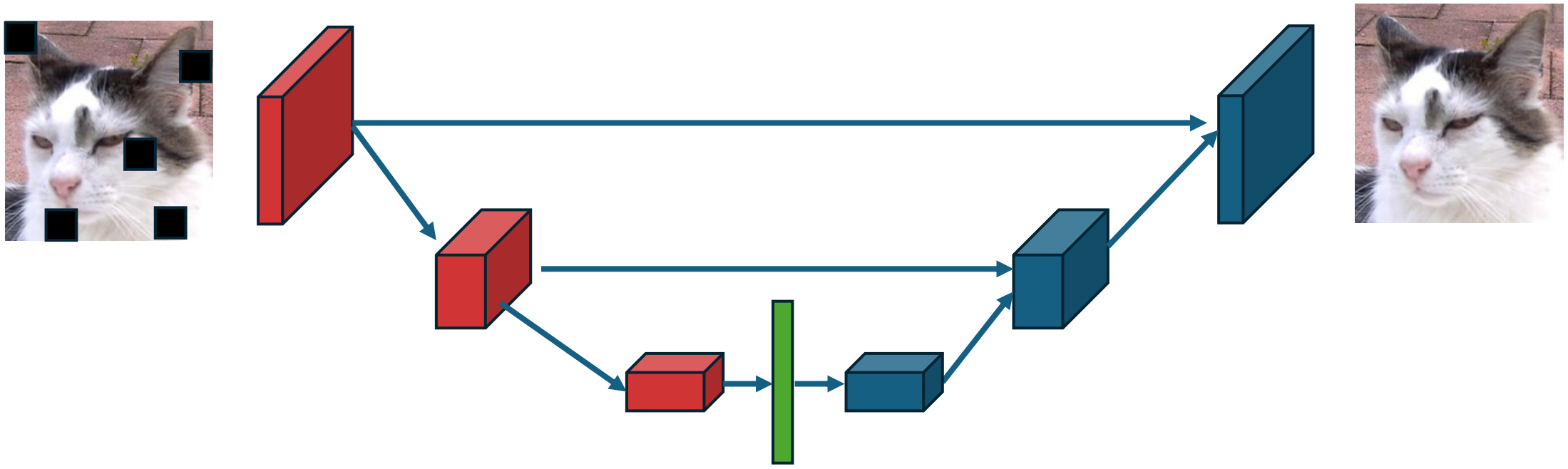

500 car 0.82
823 car 0.65
811 car 0.65
2 car 0.90
TAXI




26 cars,7 motorcycles
59 car 0.85
0:01 / 0:14
CC

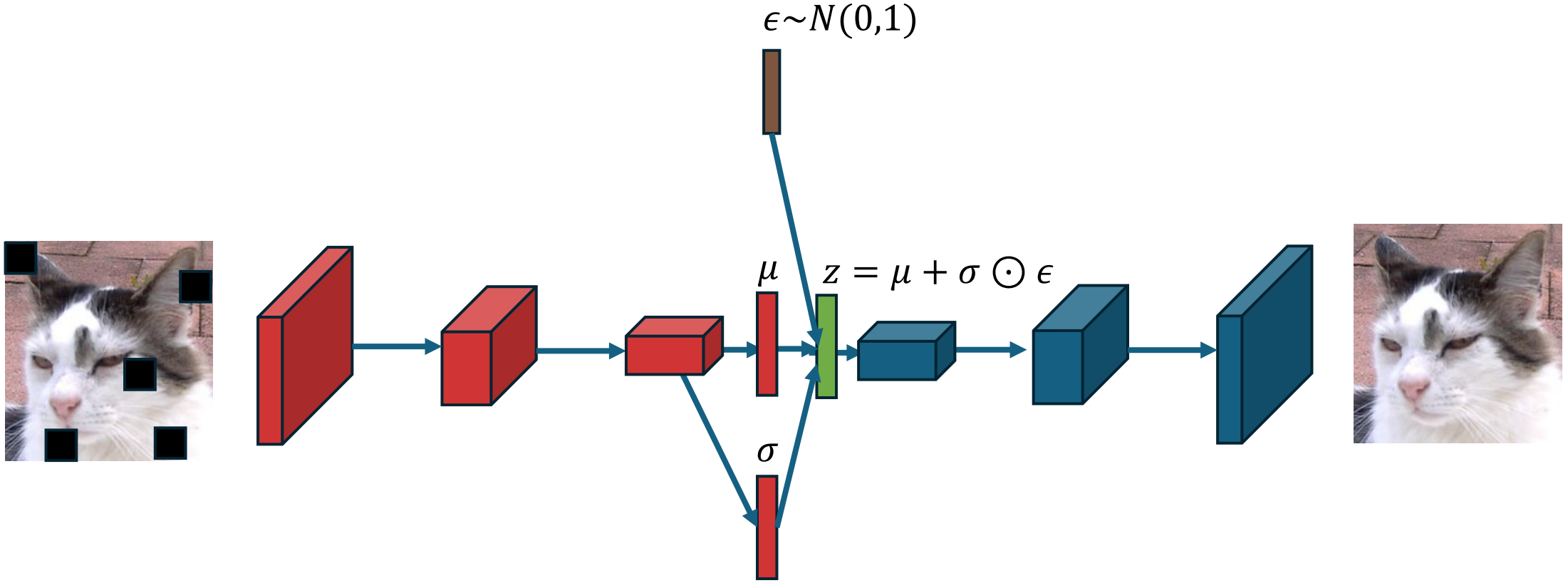


Encoder-Decoder for
Image Denoising, Image Segmentation

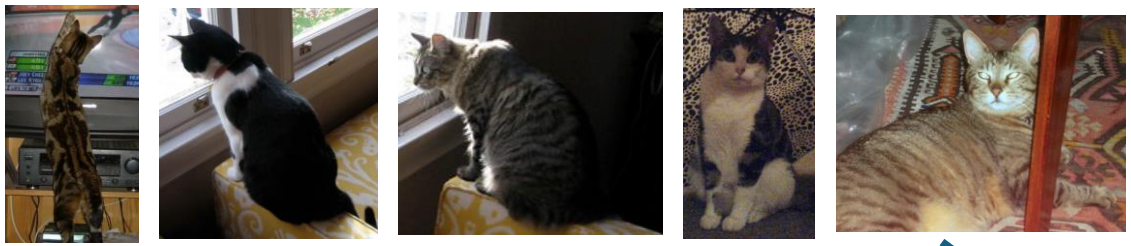


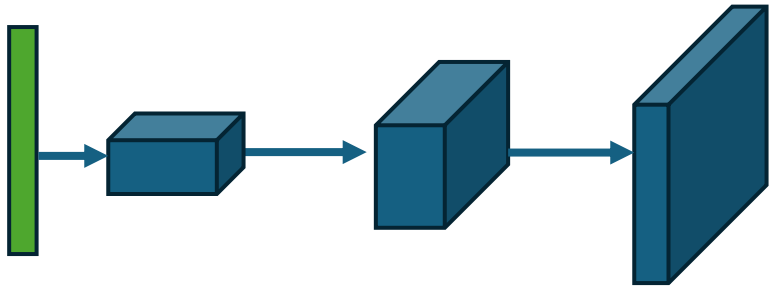
U-Net for

Image Denoising, Image Segmentation, Diffusion Model




# Variational Autoencoder (VAE) for Image Generation





Real

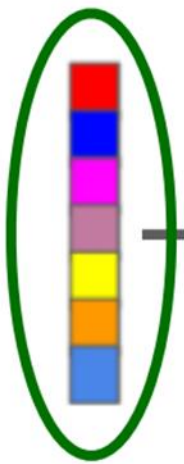
Fake

Discriminator

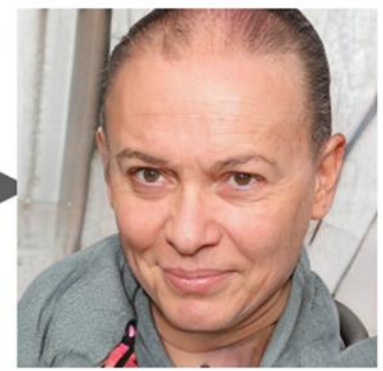
Generator

# Generative Adversarial Networks (GAN) for Image Generation, Super Resolution

# GAN



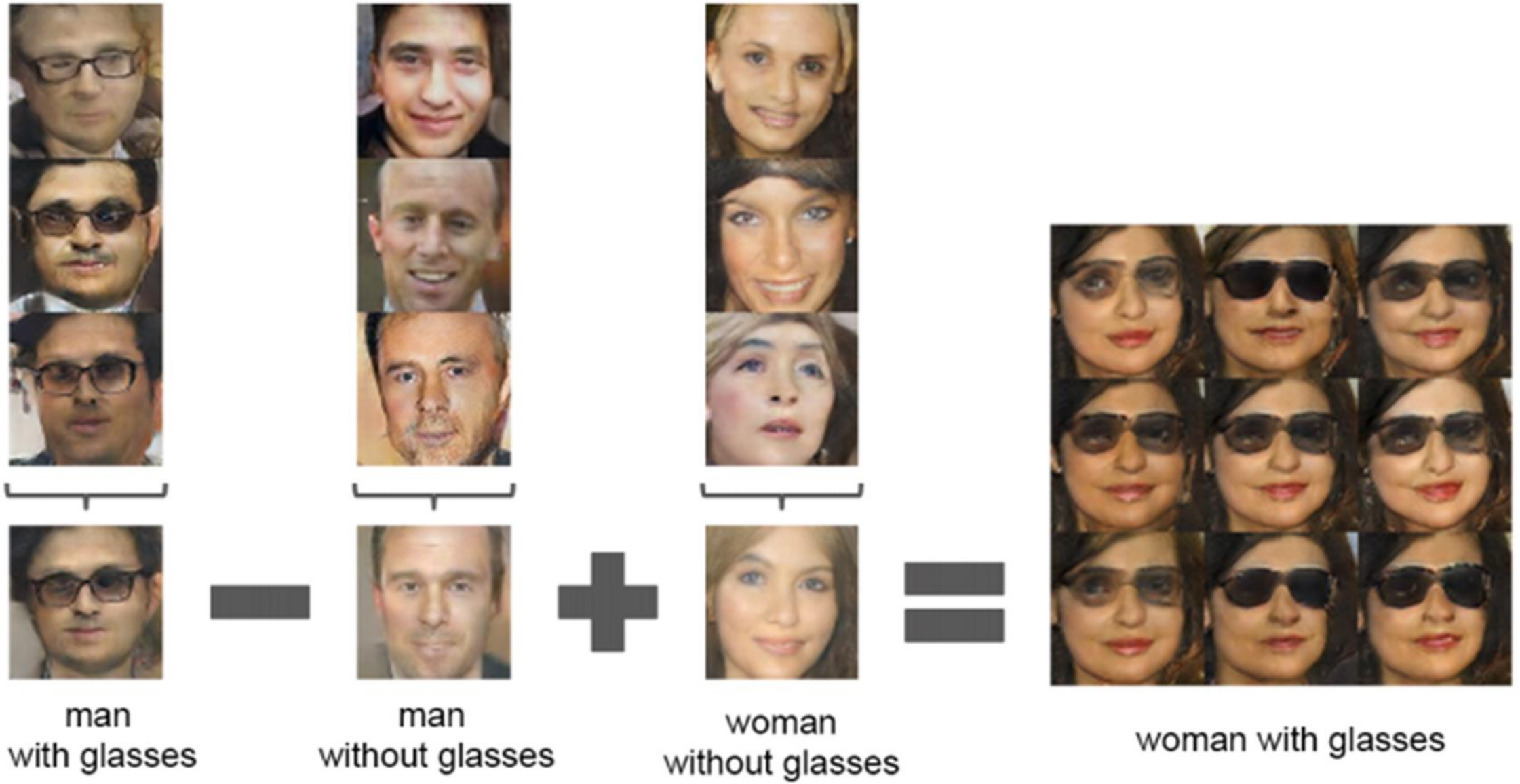
Generator



Discriminator

Real / Fake

Adjust face appearance

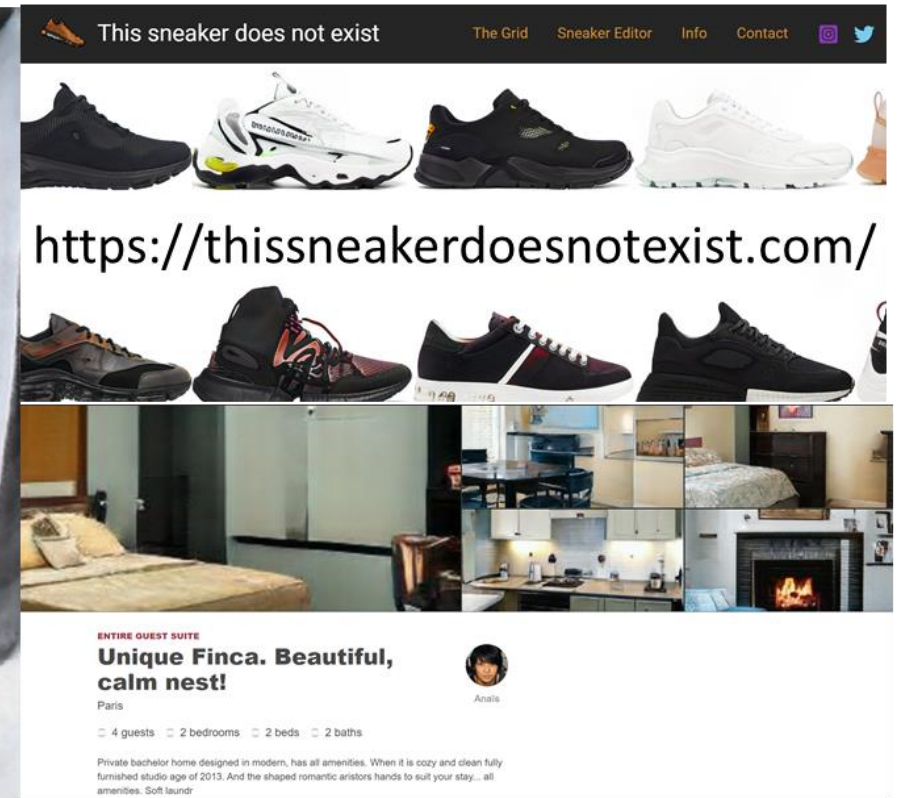



A. Radford et al. “Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks”, 2016

# X does not exist

https://thispersondoesnotexist.com/

https://thiscatdoesnotexist.com/




https://thisrentaldoesnotexist.com/

# Another gradient vanishing (for sequence)

# Classical Language Model

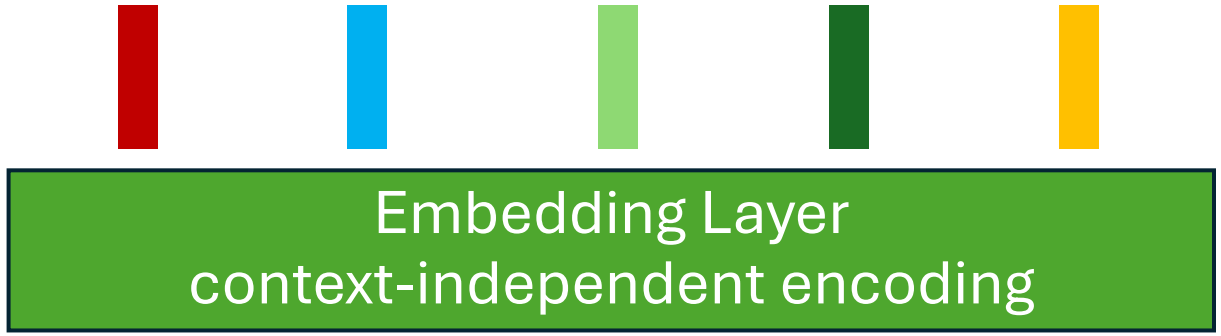
- "The wheel on the bus goes …"

- **N-gram model** $p(x_t|x_{<t}) = p(x_t|x_{t-(n-1)}, \dots, x_{t-1})$
  - Each word depends on n-1 previous words
  - Unigram, Bigram, Trigram
  - Probability table, discrete observations
- Improved models: Hidden Markov Model (HMM), Conditional Rendom Field (CRF)
  - Still rely on discrete observations

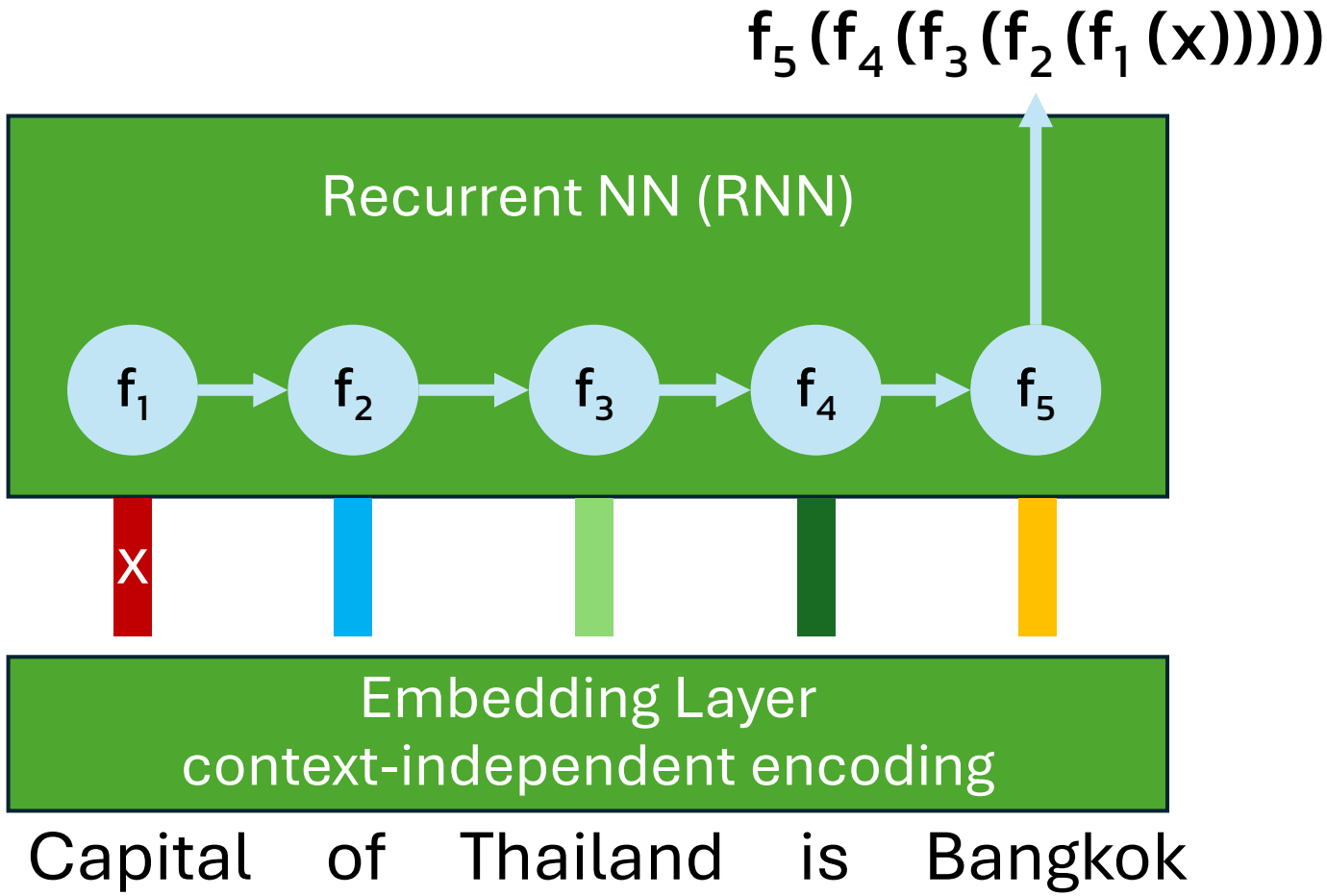
# Continuous representation

- **One-hot vector**
  - Dimension = number of words in vocabulary
  - 1 on position of this word in the vocab, 0 on every other dimensions
- **Bag-of-words** = present/absent of words in the input document
- **Latent Semantic Analysis (LSA)**= **Principal Component Analysis (PCA)** on bag-of-words vectors
  - Obtain a dense vector for each word
- **Word2vec**
  - Autoencoder for one-hot vector, with 1 encoding layer & 1 decoding layer
  - Train either to predict surrounding words or to use surrounding words to predict this word
  - After training, the latent from encoding layer = vector represent the word
  - vec(King) – vec(Man) + vec(Woman) = vec(Queen)

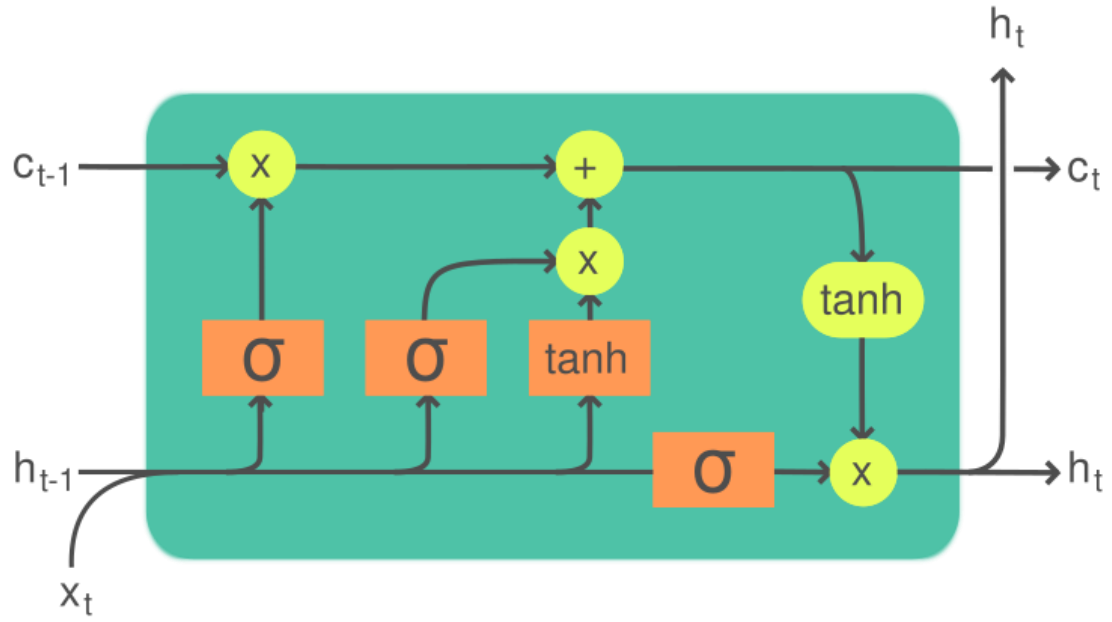
# Context-independent representation

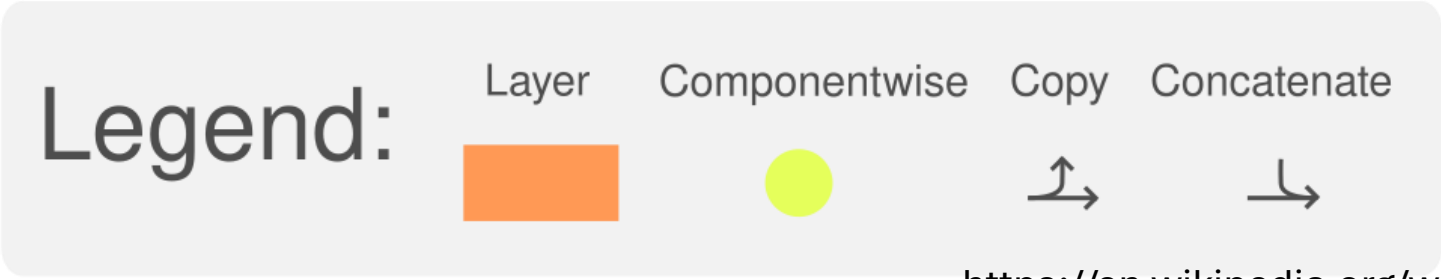


# Context-dependent representation



# Long Short-Term Memory (LSTM)




https://en.wikipedia.org/wiki/Long_short-term_memory

# LSTM

Forget gate → $f_t = \sigma_g(W_f x_t + U_f h_{t-1} + b_f)$

Input gate → $i_t = \sigma_g(W_i x_t + U_i h_{t-1} + b_i)$

Output gate → $o_t = \sigma_g(W_o x_t + U_o h_{t-1} + b_o)$

$\tilde{c}_t = \sigma_c(W_c x_t + U_c h_{t-1} + b_c)$

$c_t = f_t \odot c_{t-1} + i_t \odot \tilde{c}_t$

$h_t = o_t \odot \sigma_h(c_t)$

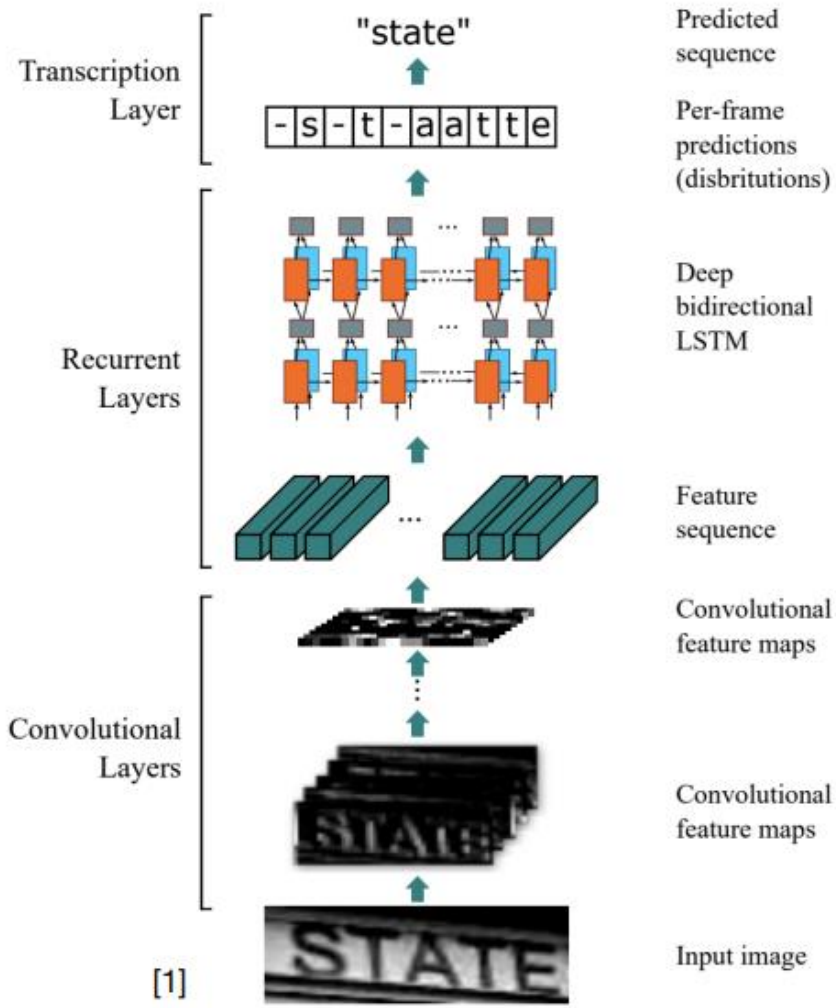
Sigmoid fct

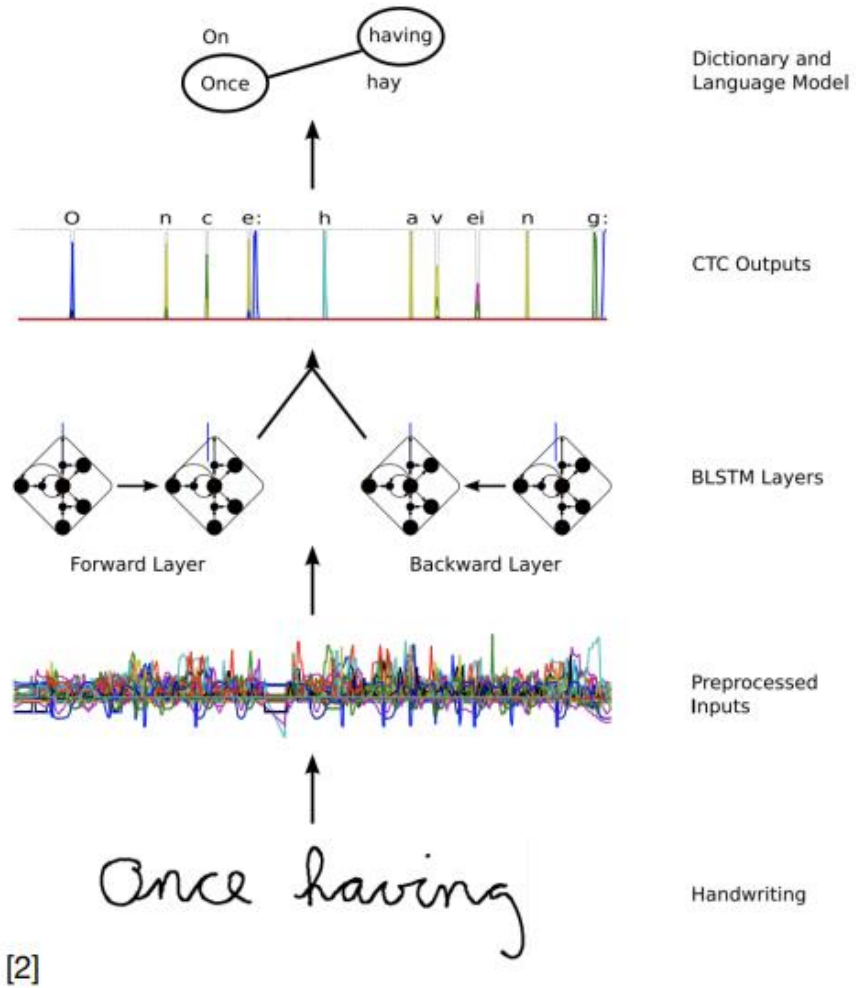
Hadamard product

$$c_t = \sigma_h(c_{t-1} + \sigma_c(W_c x_t + U_c h_{t-1} + b_c))$$

~Skip connection

W. Byeon et al. “Scene Labeling with LSTM Recurrent Neural Networks”, CVPR 2015

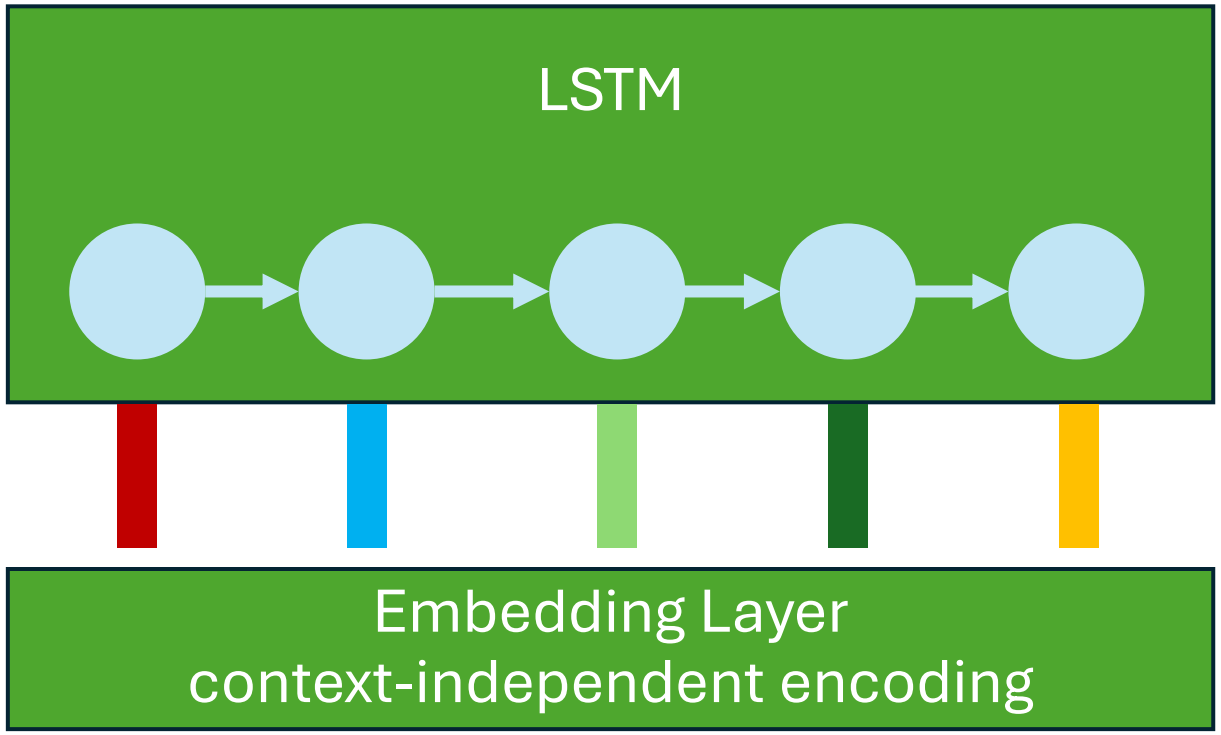


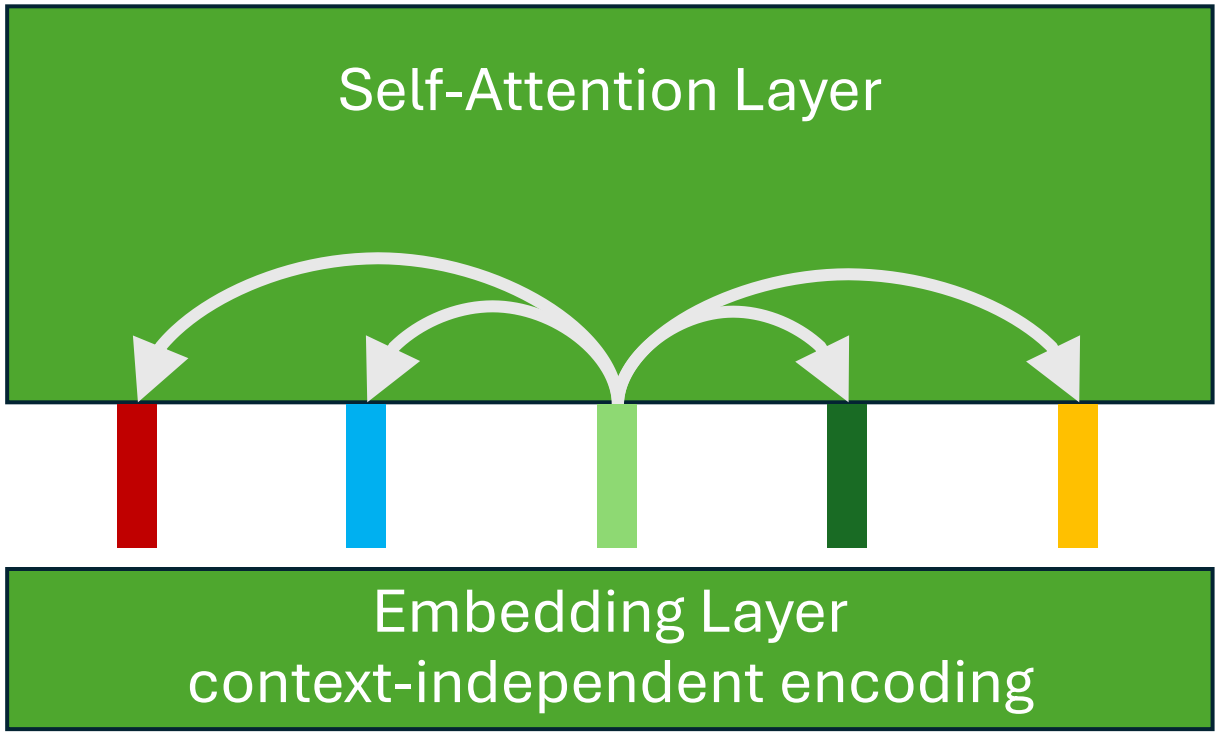




[1] B. Shi, X. Bai, and C. Yao, "An End-to-End Trainable Neural Network for Image-based Sequence Recognition and Its Application to Scene Text Recognition", https://arxiv.org/pdf/1507.05717.pdf

[2] A. Graves, M. Liwicki, S. Fernández, R. Bertolami, H. Bunke, and J. Schmidhuber, "A Novel Connectionist System for Unconstrained Handwriting Recognition", http://www.cs.toronto.edu/~graves/tpami_2009.pdf

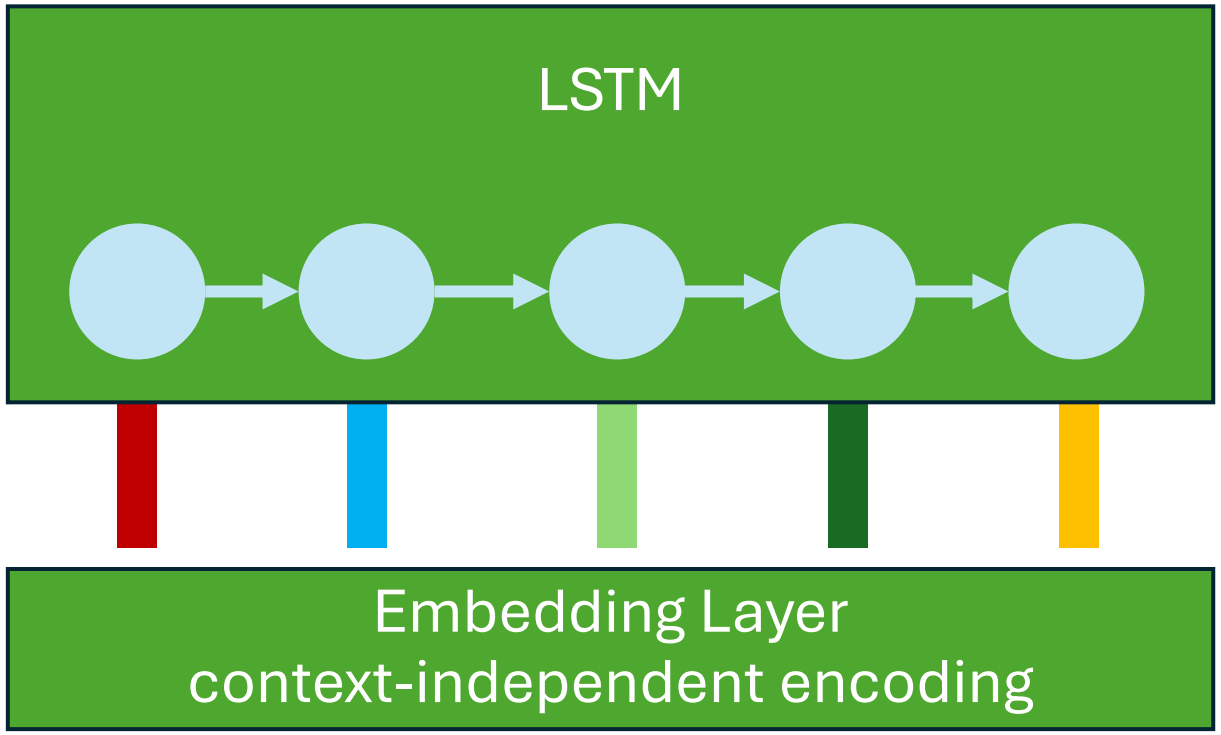
# Context-dependent representation

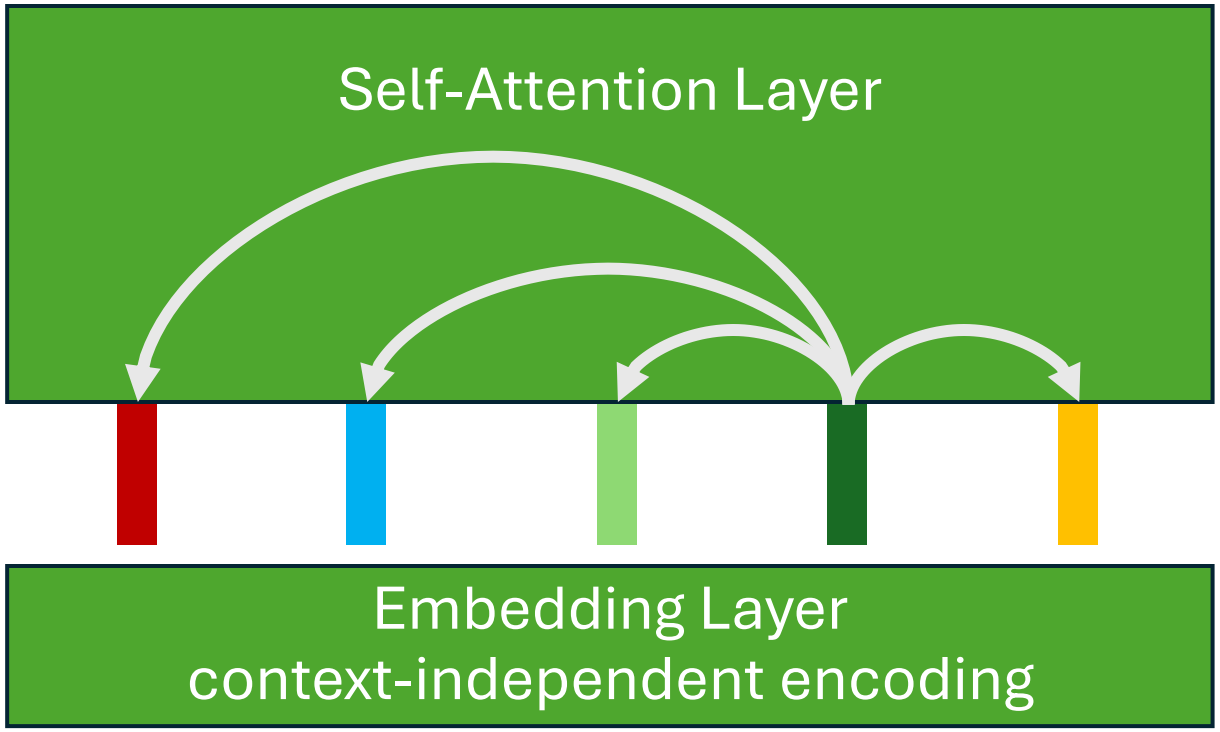


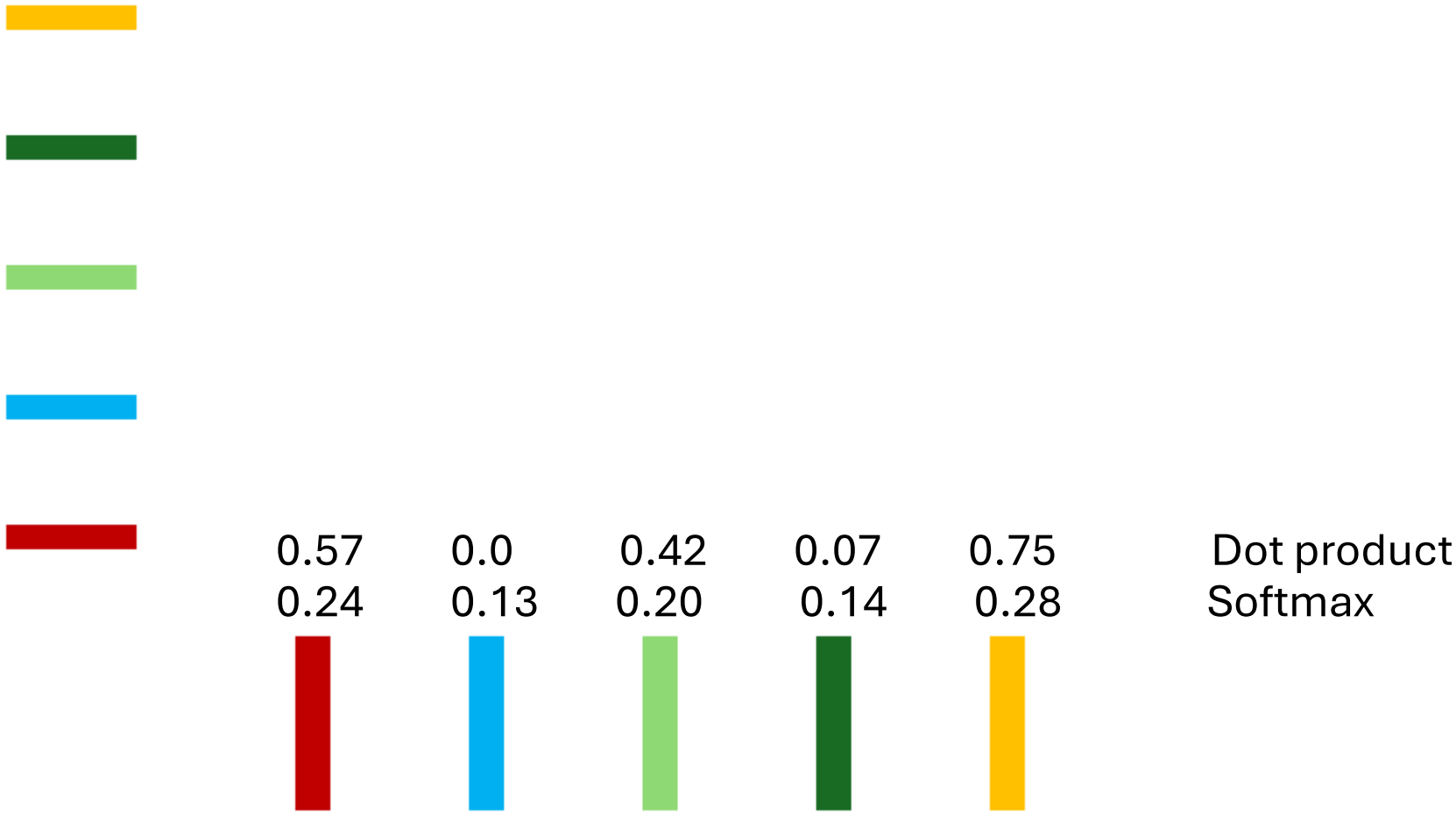
# Gradient vanishing problem

- Self-attention mechanism solves gradient vanishing with increasing computational cost

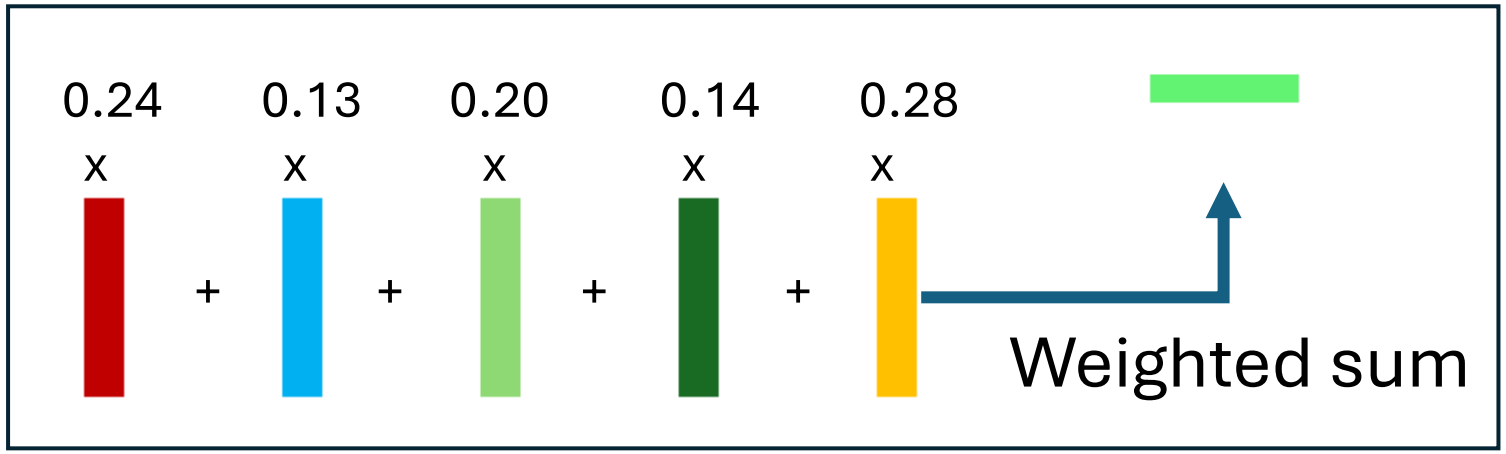


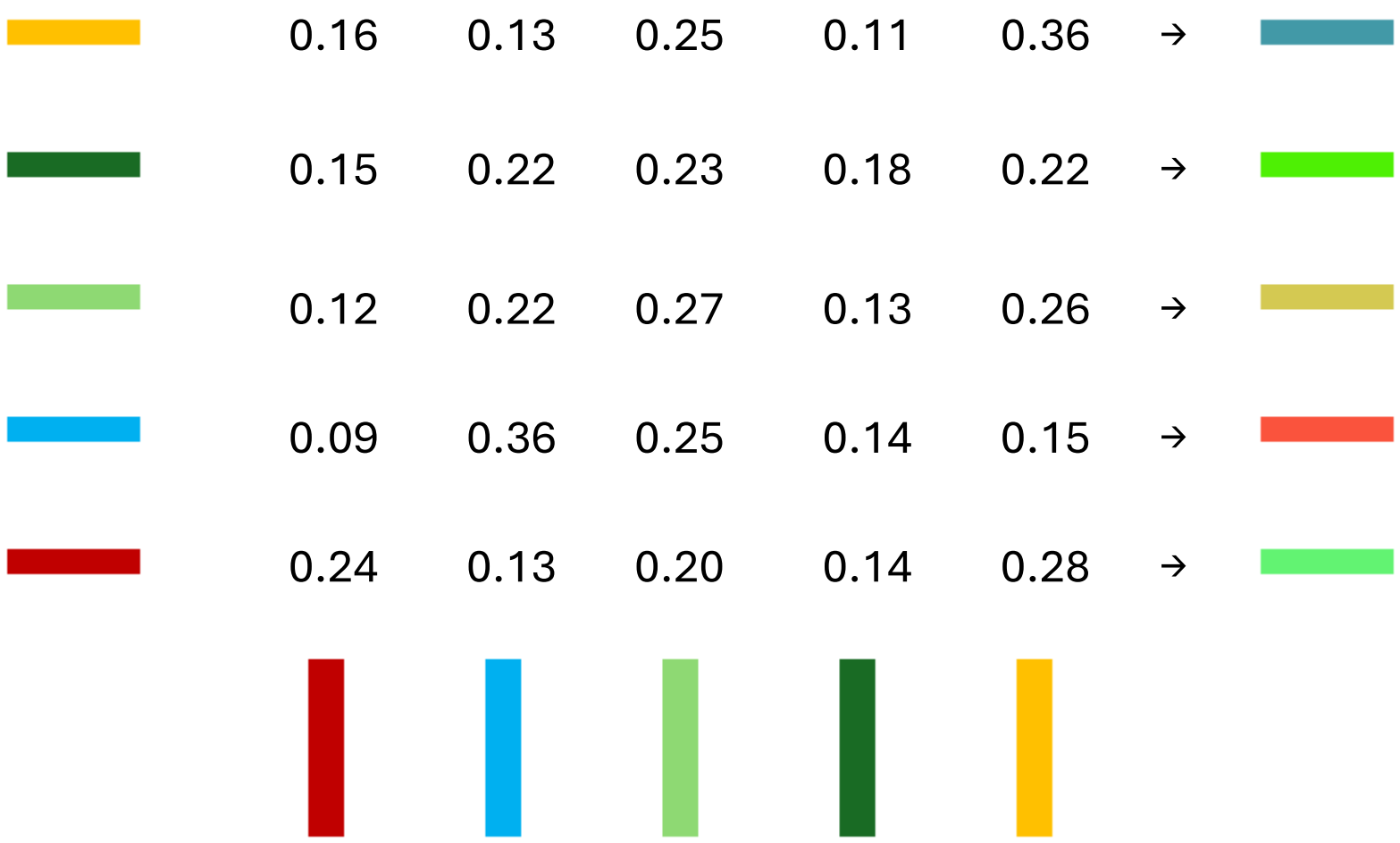
# Self-Attention Layer



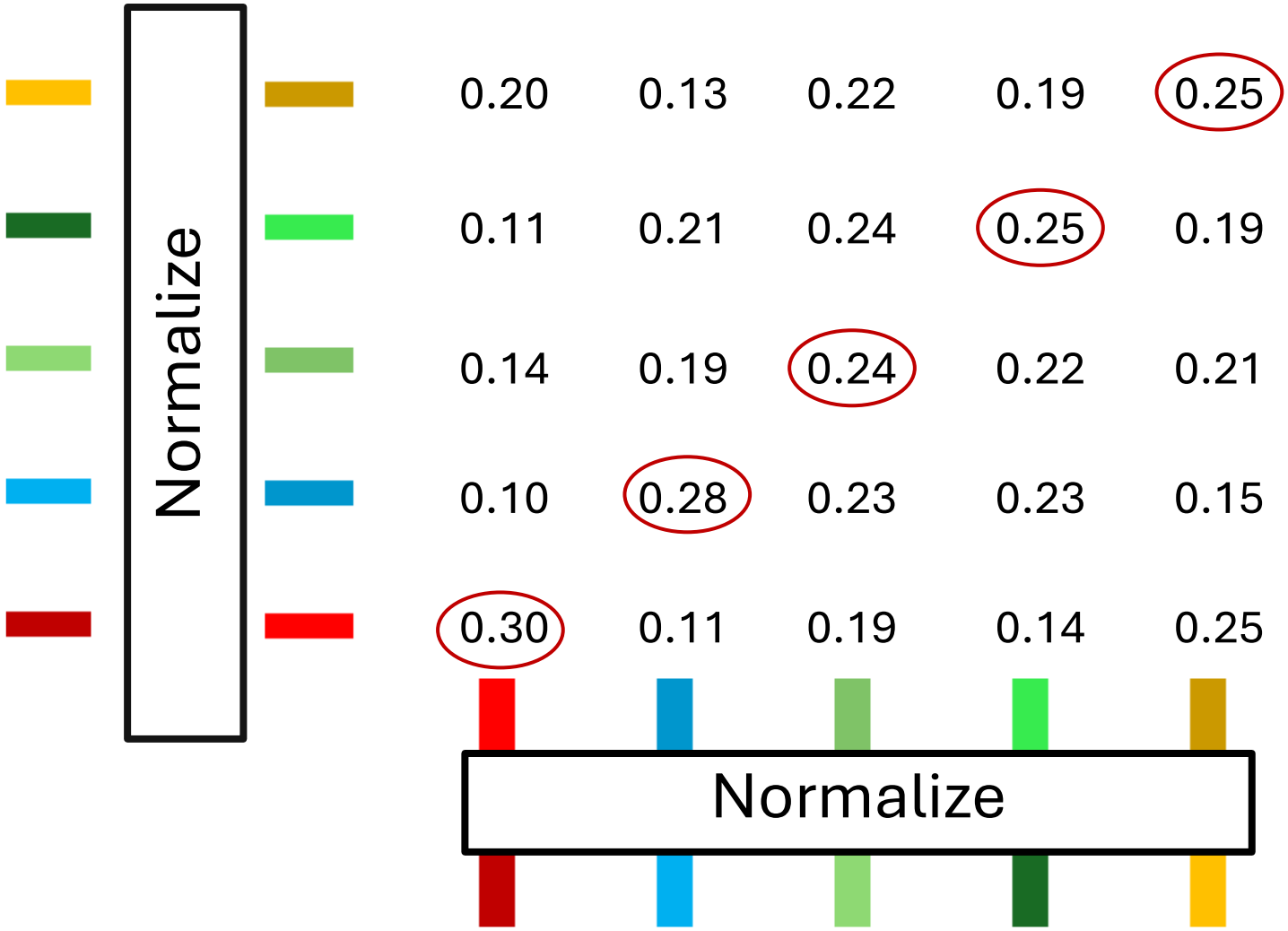
# Self-Attention Layer



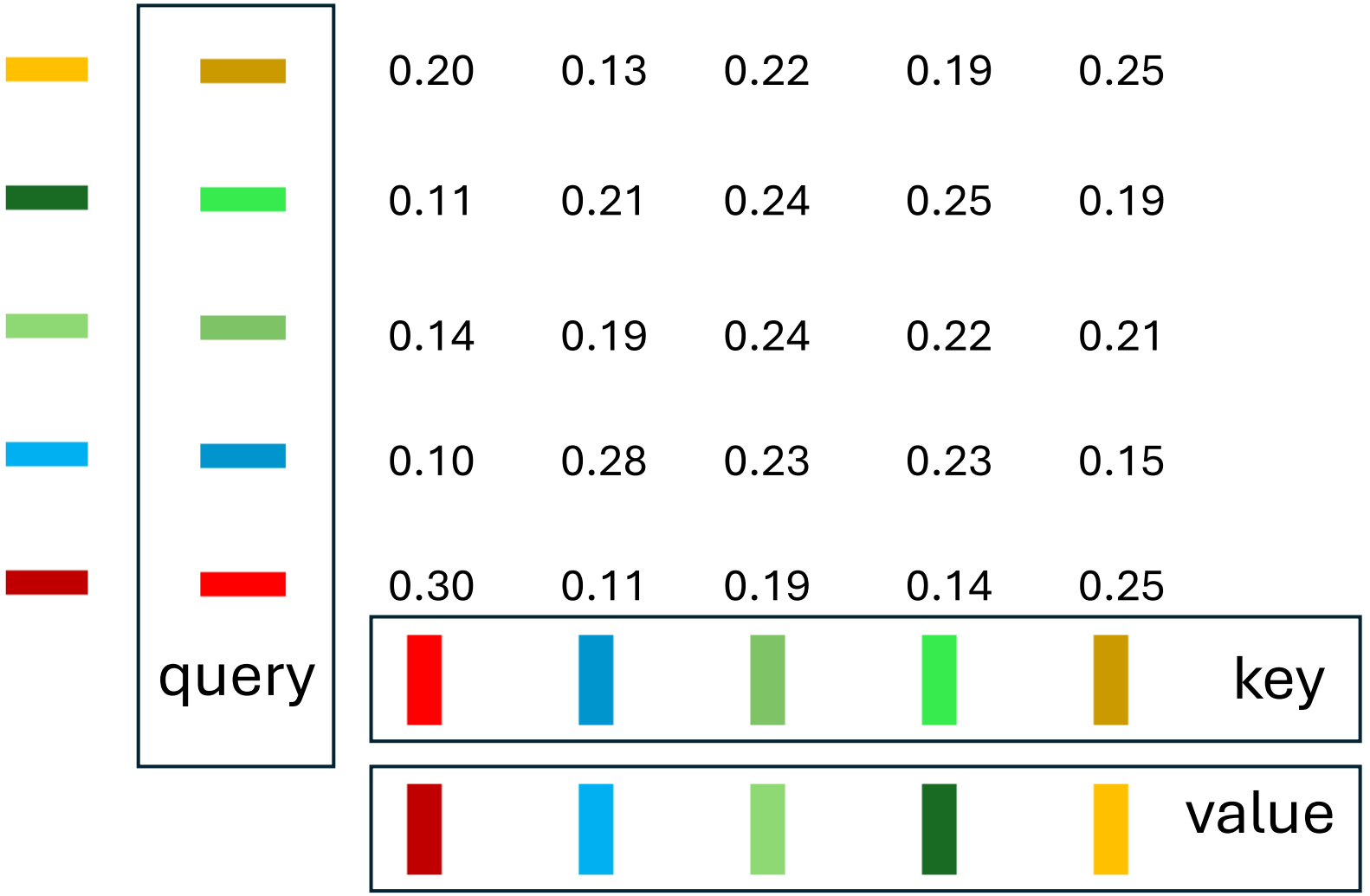
# Self-Attention Layer



# Self-Attention Layer

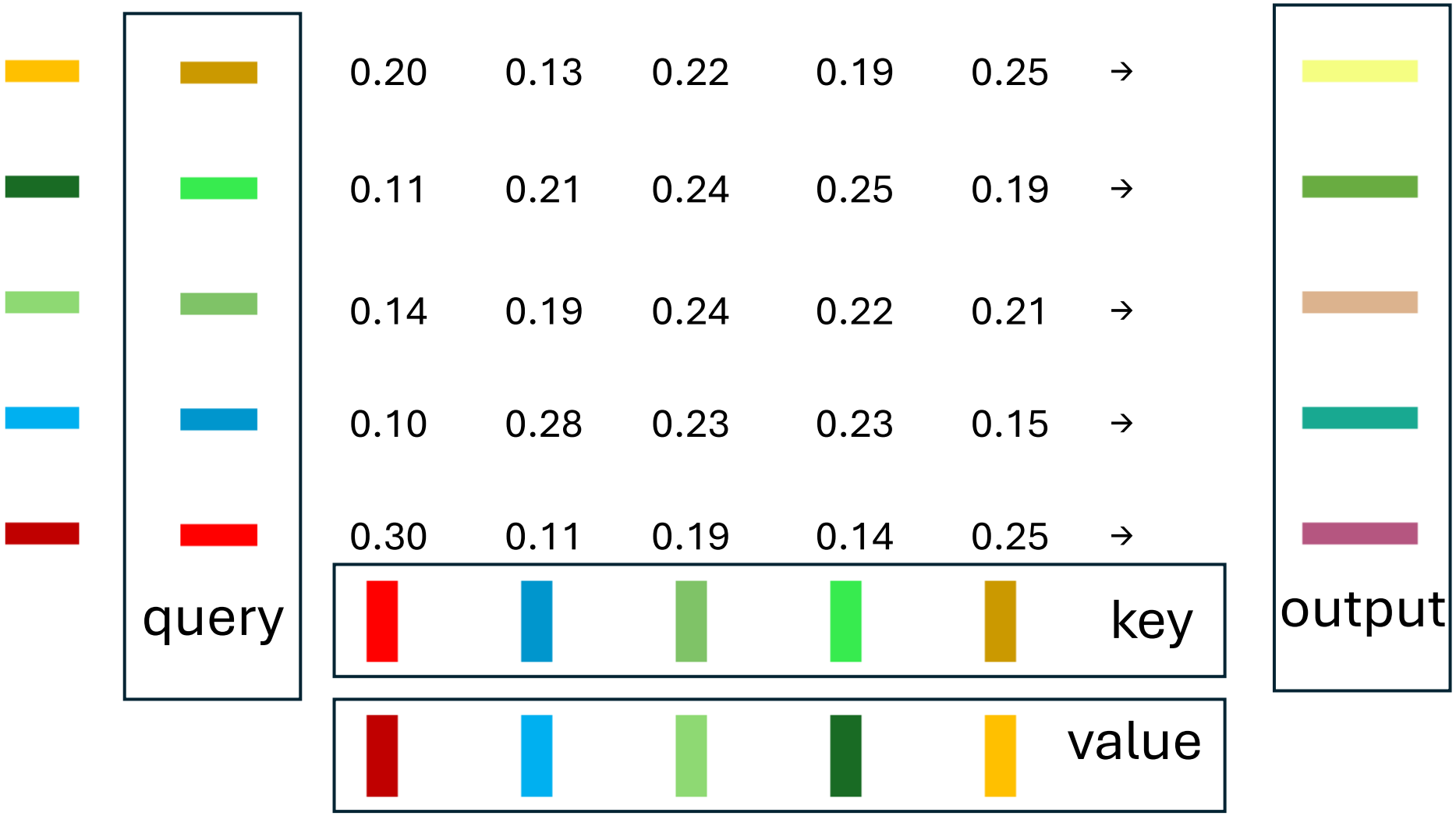


# Self-Attention Layer



# Self-Attention Layer

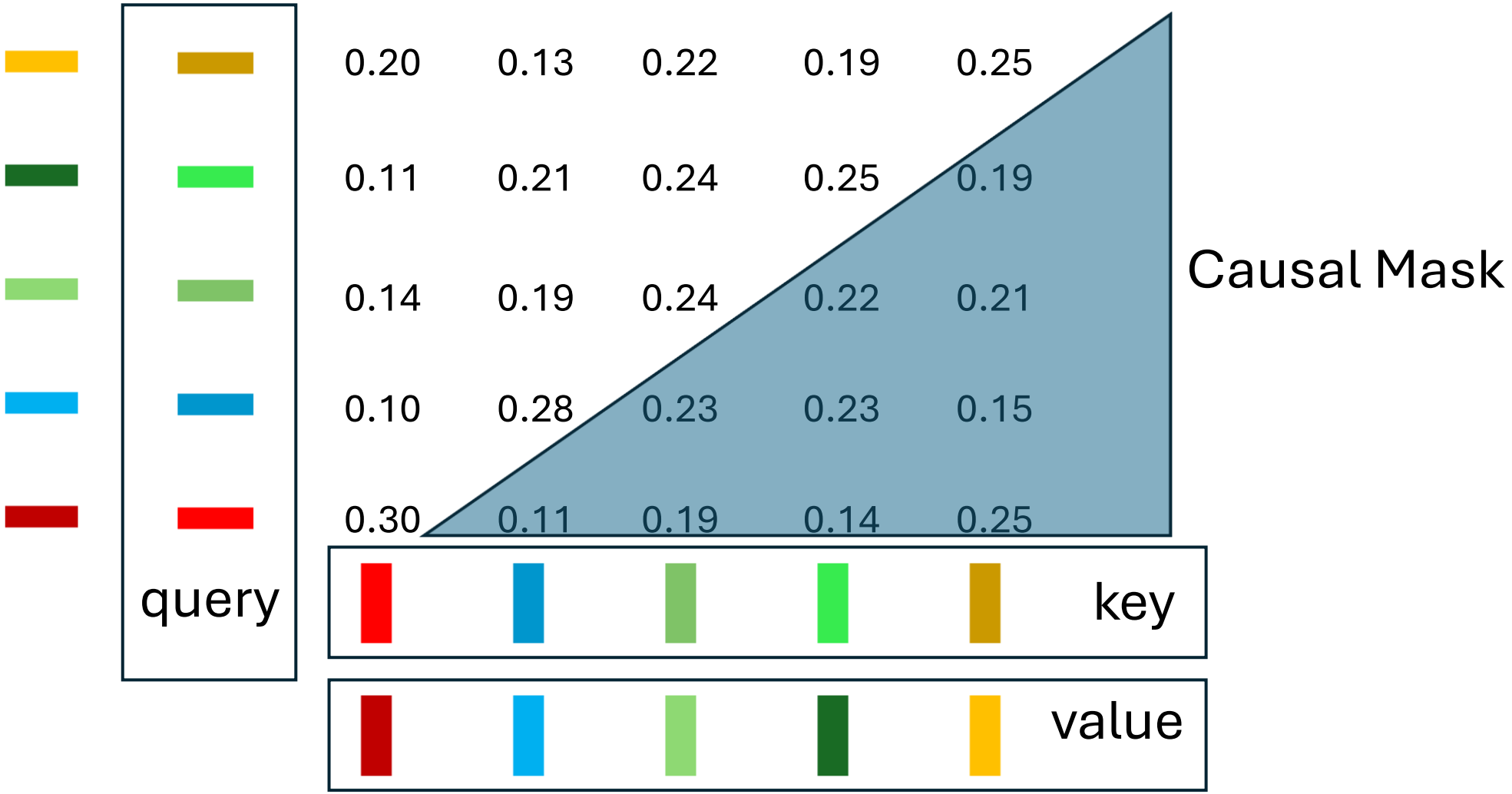
# Self-Attention Layer



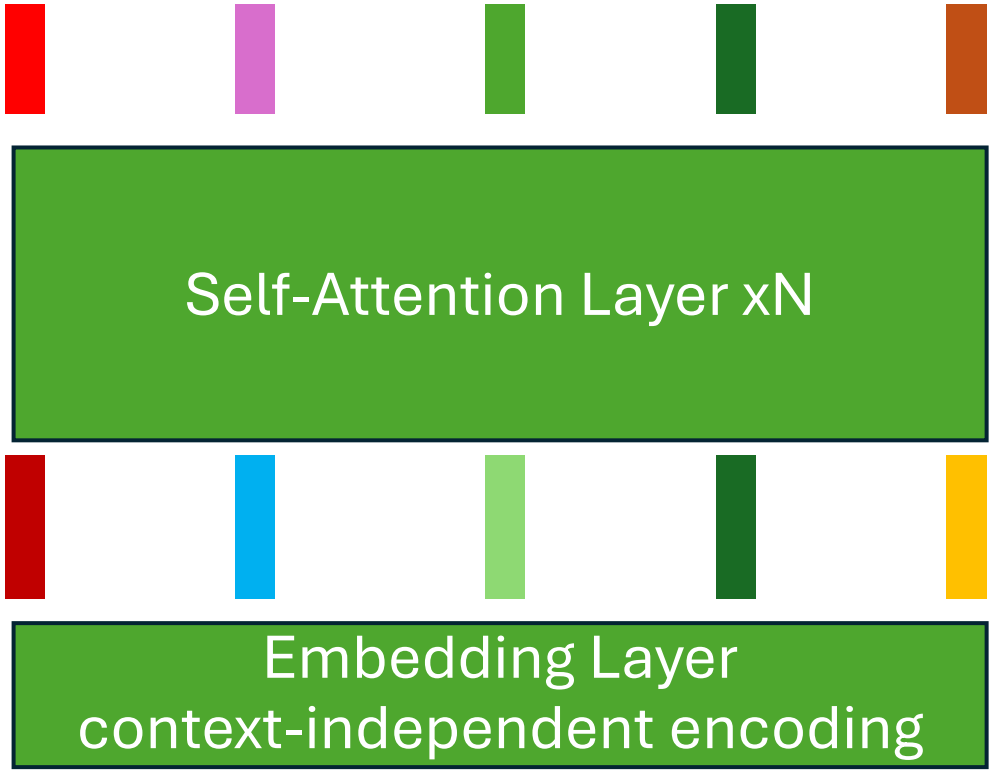
# Variations

- Attention with **Causal Mask**
  - Force left->right generation (also called autoregressive)
  - Useful for text generation)
- **Cross-attention** (useful to combine information from different modalities)
- **Multi-Head Attention (MHA)**
  - Use sets of Attention in parallel
  - Each Attention head focus on a particular features similar to Convolutional filter
  - **This is interesting. LSTM has units, Transformer has attention heads**
- Attention block = transform a sequence of vectors into another sequence of vectors of the same length
- **Positional Encoding** = Embedding layer for position of each token in input sequence

# Self-Attention Layer

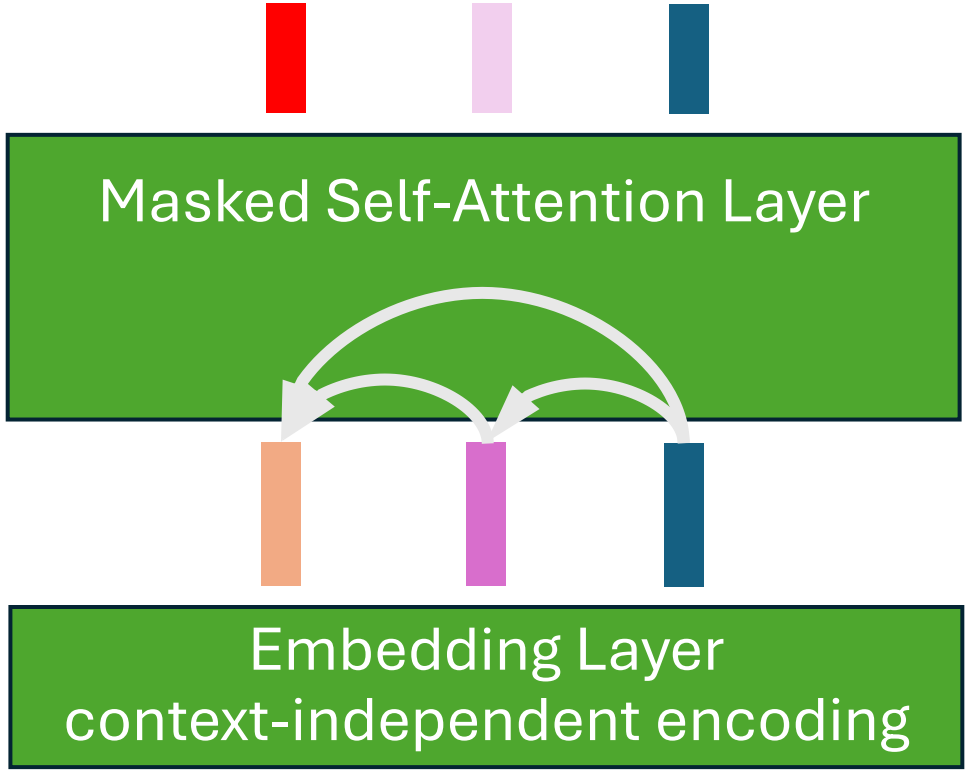


## Transformer Encoder
## BERT

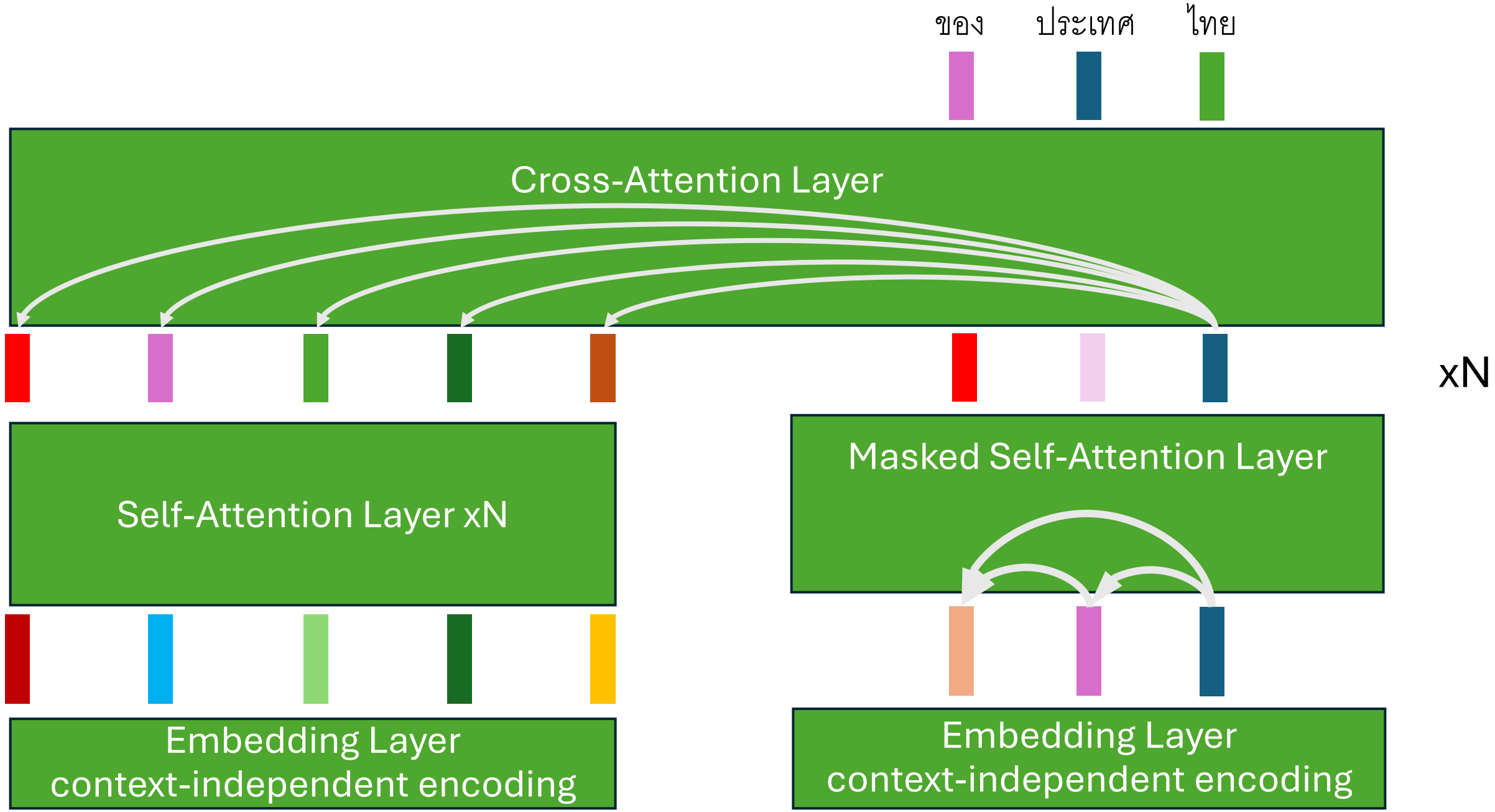


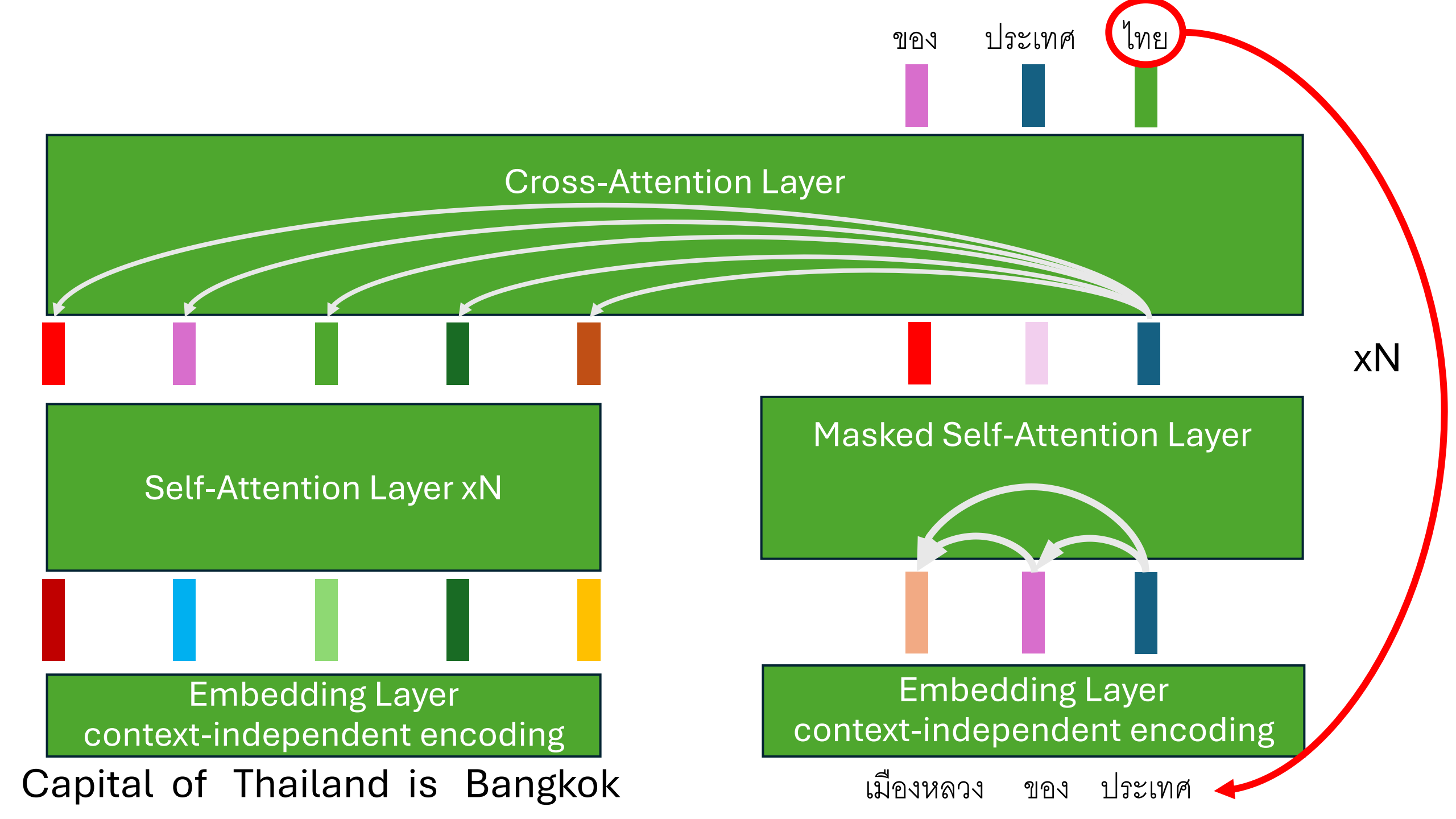

Capital of Thailand is Bangkok

## Transformer Decoder

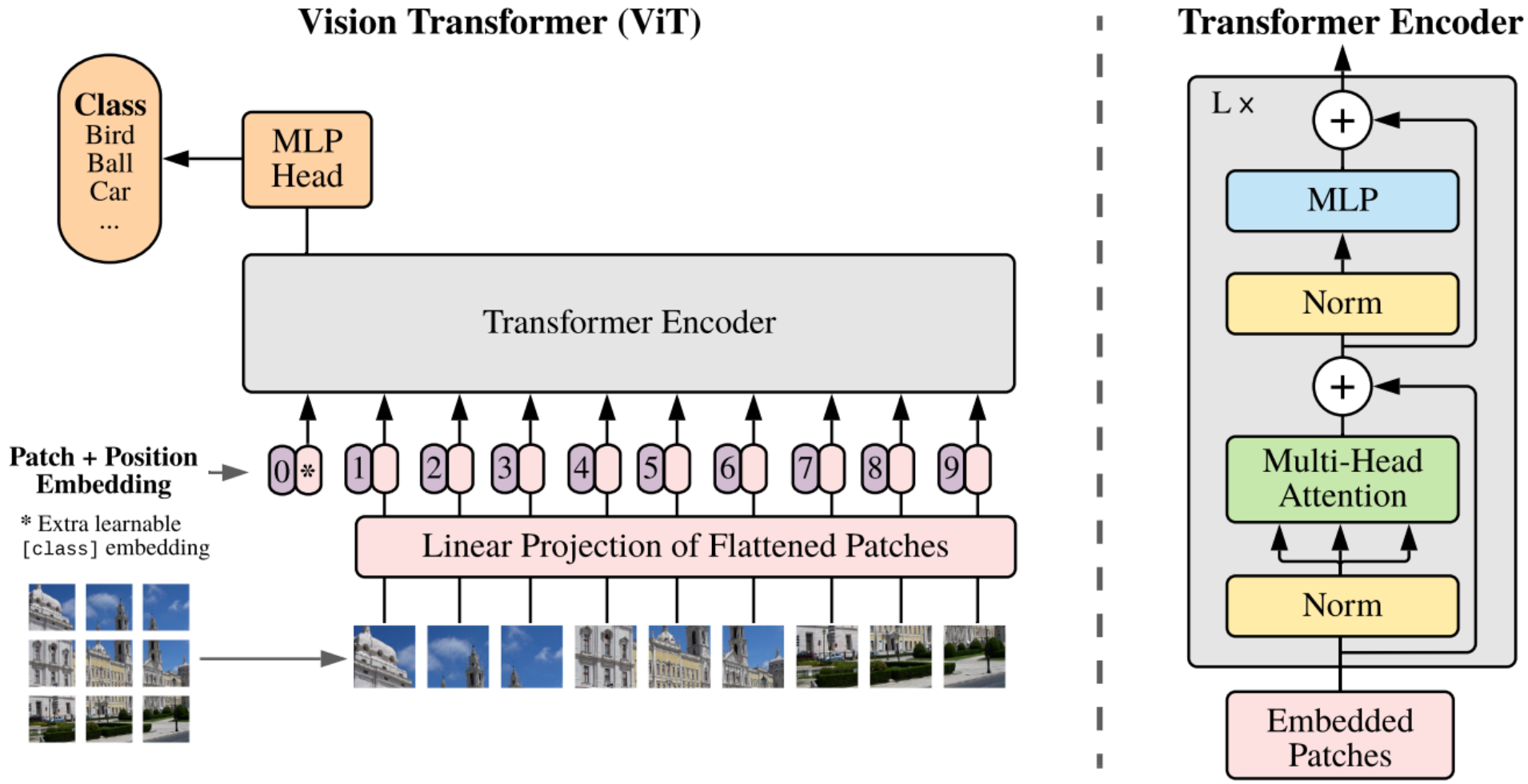



เมืองหลวง ของ ประเทศ


ของ
ประเทศ
ไทย
Cross-Attention Layer
xN
Self-Attention Layer xN
Masked Self-Attention Layer
Embedding Layer
context-independent encoding
Embedding Layer
context-independent encoding
Capital of Thailand is Bangkok
เมืองหลวง ของ ประเทศ


ของ
ประเทศ
ไทย
Cross-Attention Layer
xN
Self-Attention Layer xN
Masked Self-Attention Layer
Embedding Layer
context-independent encoding
Embedding Layer
context-independent encoding
Capital of Thailand is Bangkok
เมืองหลวง ของ ประเทศ

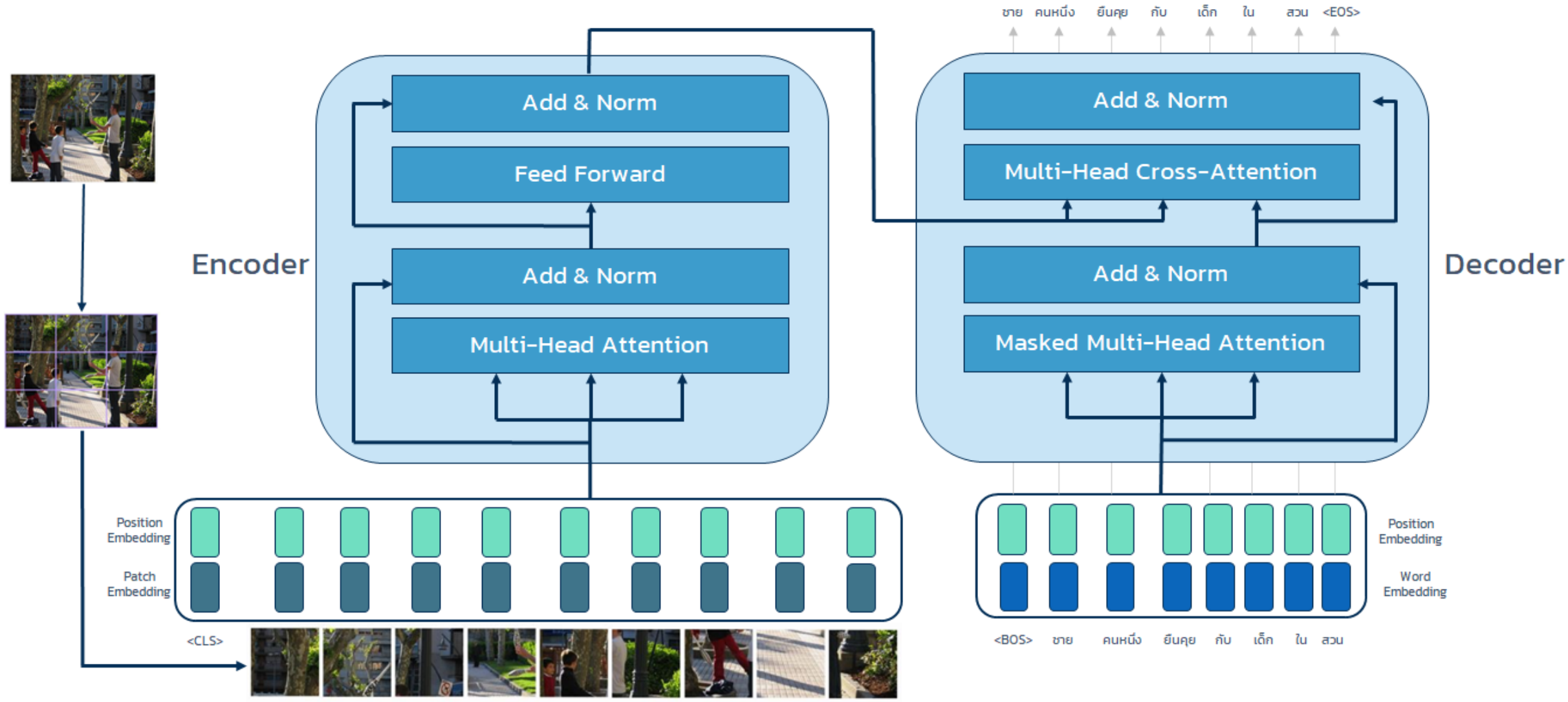
# Vision Transformer (ViT)

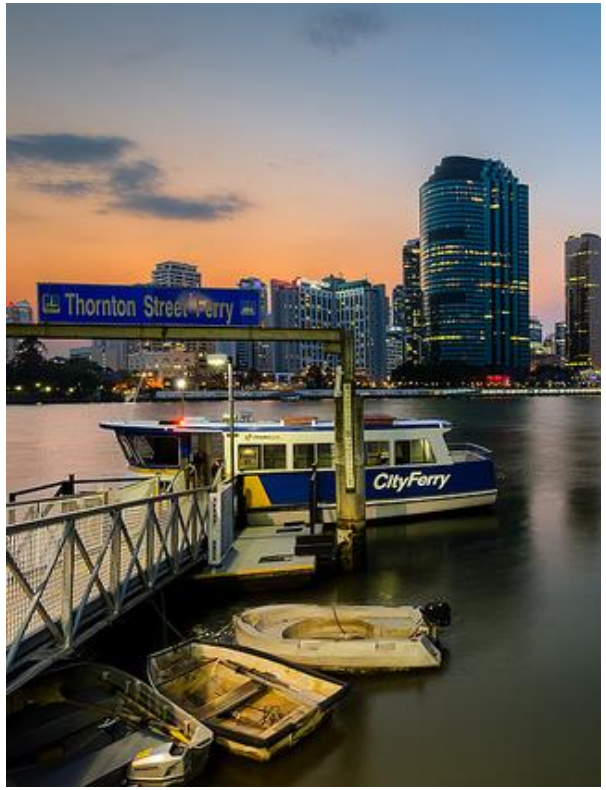


# Alternative approaches

- Depth-wise Separable Convolution
- ConvNeXt
- CoAtNet
- MLP-Mixer
- Graph NN

# Image Captioning



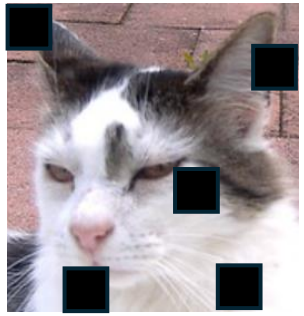


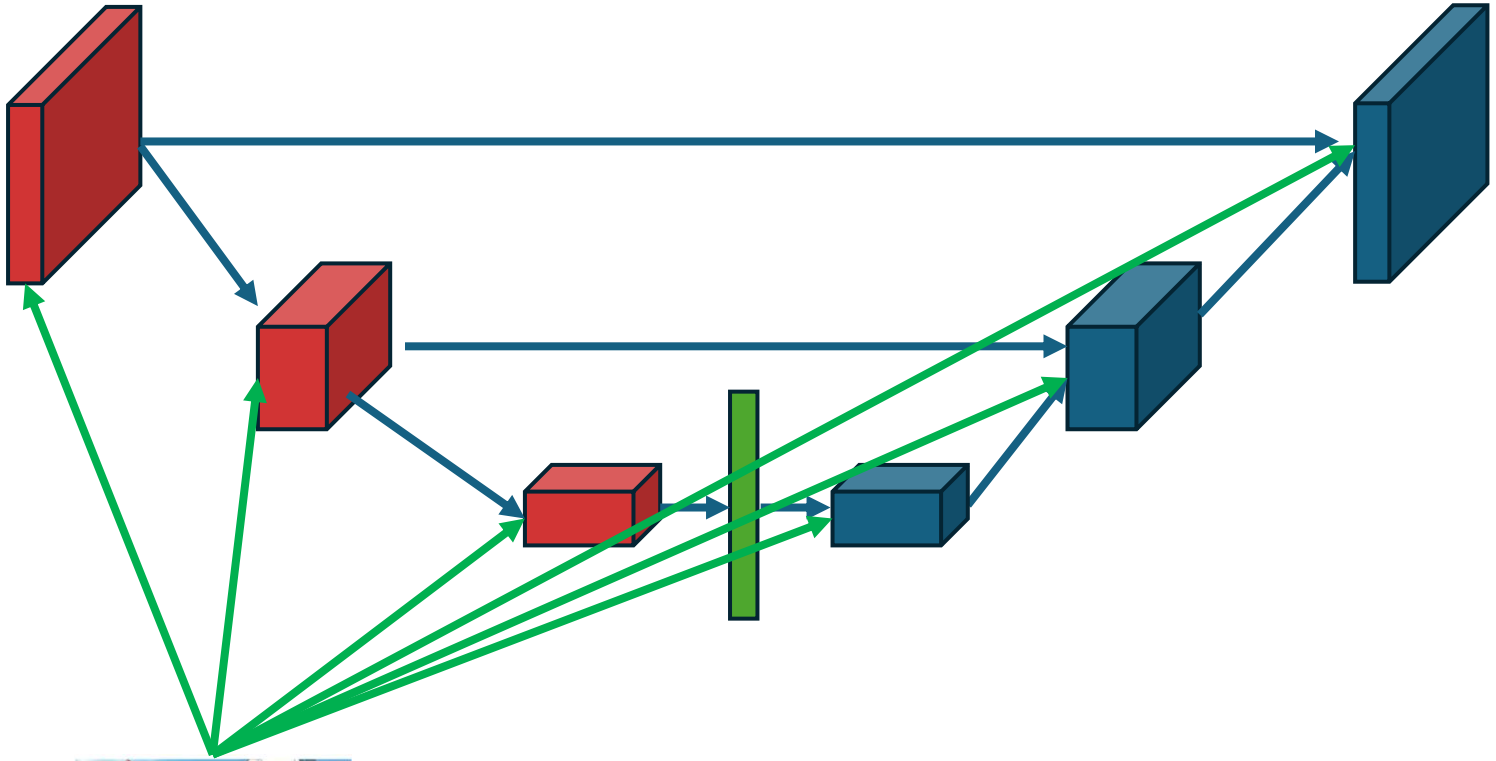

แม่น้ำแห่งหนึ่งมีเรือจำนวนหลายลำจอดอยู่บนแม่น้ำด้านข้างสะพาน

สนามหญ้ามีรถบรรทุกคันสีดำคันหนึ่งจอดอยู่ข้างบ้านหลังหนึ่ง

คนกำลังเล่นว่าวอยู่ใกล้กับตึกสูงสีขาวและท้องฟ้ามีเมฆมาก

ผู้ชายผมสีดำมีหนวดสวมแว่นกำลังถือลูกบอลสีเหลือง



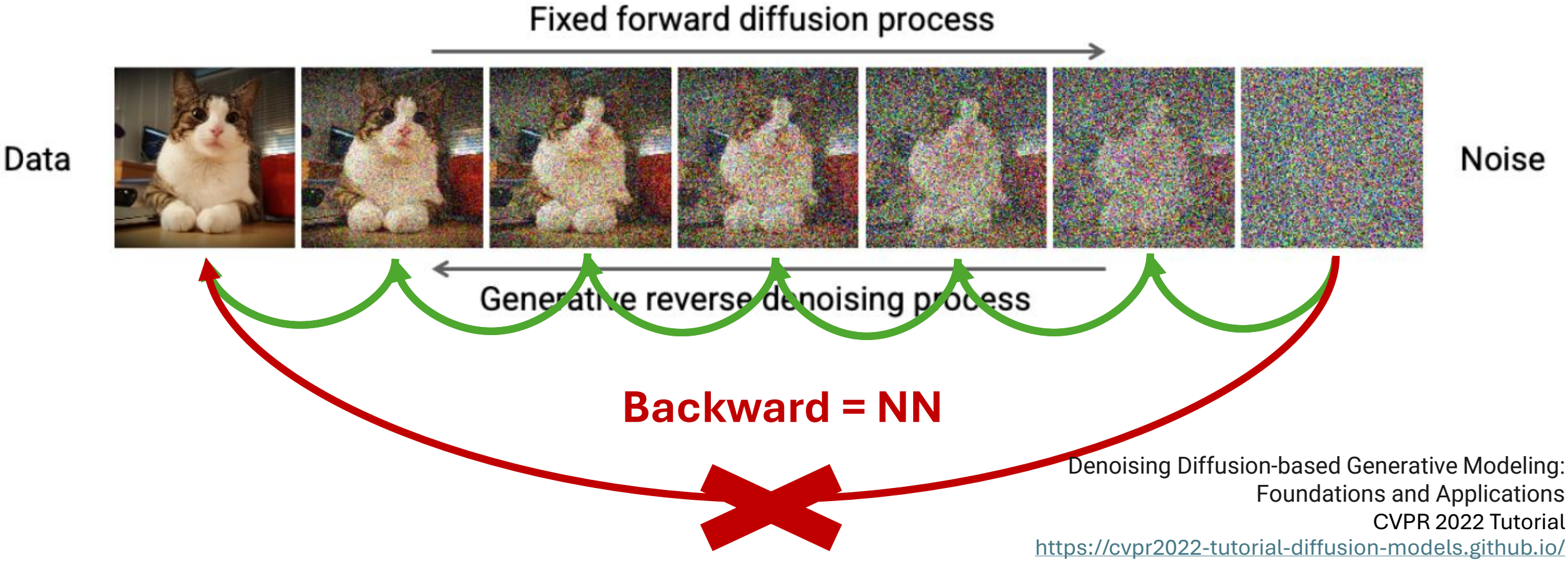


An image of a cat

Stable Diffusion

# Diffusion Model

## Forward = Add Noise



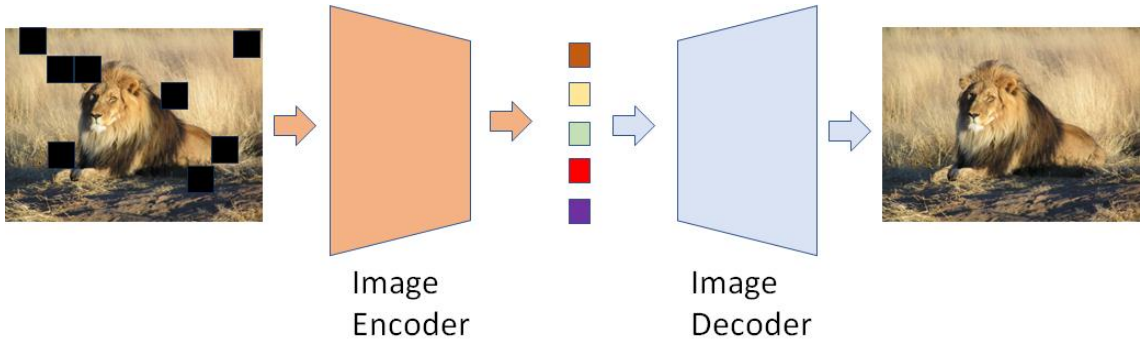

Denoising Diffusion-based Generative Modeling: Foundations and Applications
CVPR 2022 Tutorial
https://cvpr2022-tutorial-diffusion-models.github.io/

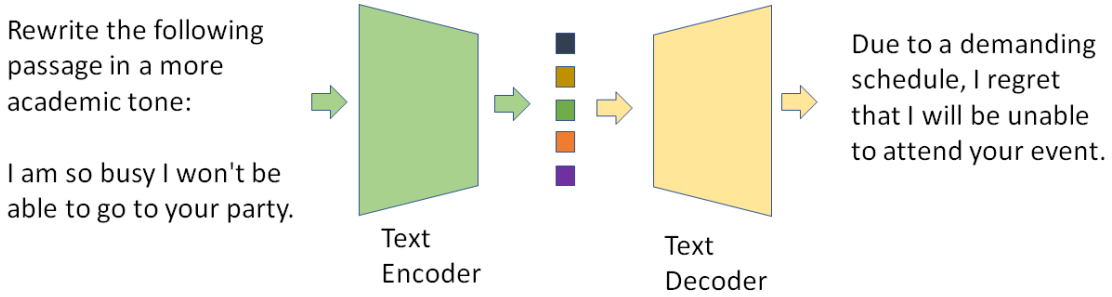


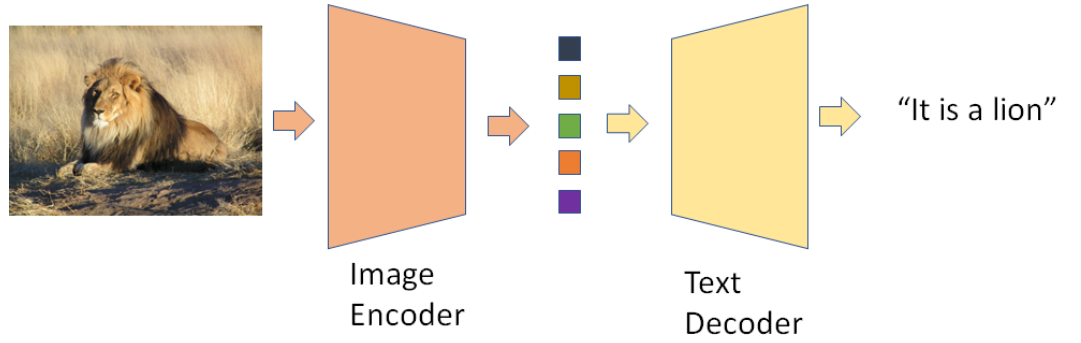


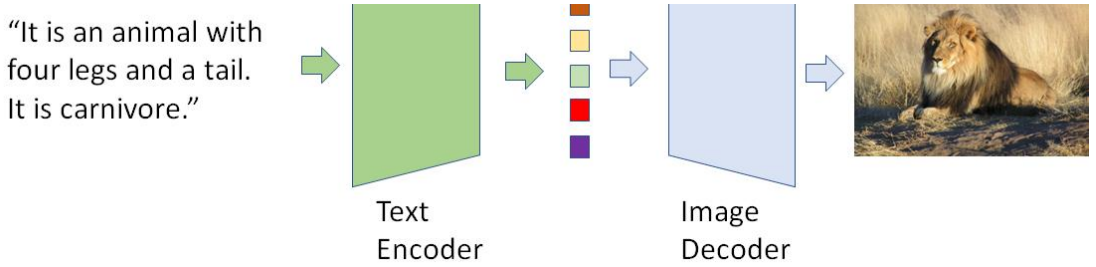

# Future of Computer Vision

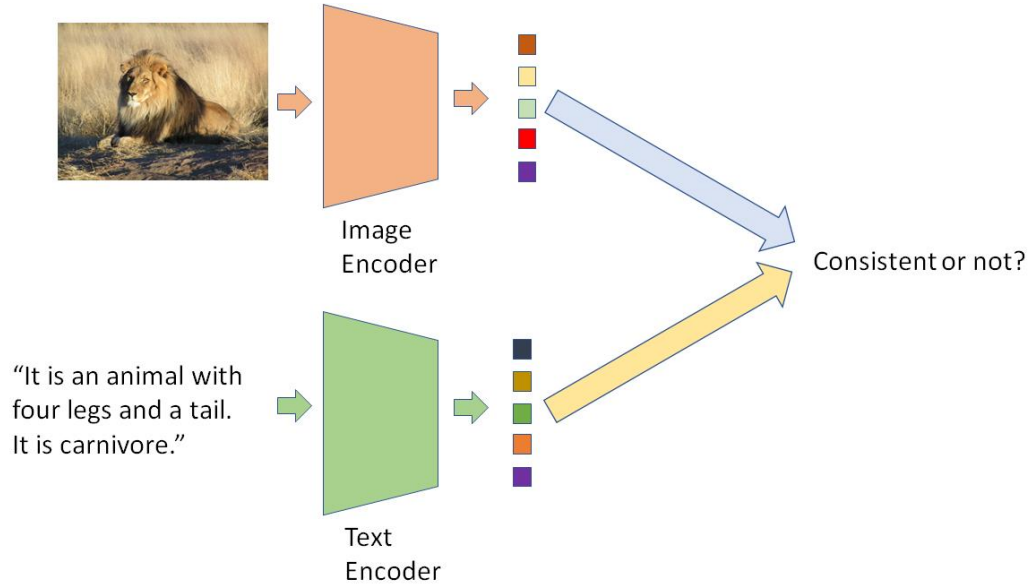
# Multi-Modal AI



# Image Search

- Use surrounding text as index of an image in a web page work with text query
- Use image feature vector as index of an image, work with image query (Content-Based Image Retrieval, CBIR)
- Use object detectors, e.g. YOLO, to index images with known keywords work with text query
- Use CLIP visual feature vector as index of an image
  - Compare against CLIP text feature vector from text query

- Synthetic Data
- Federated Learning
- Out-of-distribution
- Frameworks: Tensorflow Lite, TinyML
- Model:
  - Quantization, Prunning
  - Neural Architecture Search (NAS)
- Spiking NN
- Explanable AI (XAI)
- Counterfactual Analysis

# Thank you

Q & A