imc FH KREMS
UNIVERSITY OF APPLIED SCIENCES

# DEEP LEARNING #2

## Advanced Architectures

# Agenda

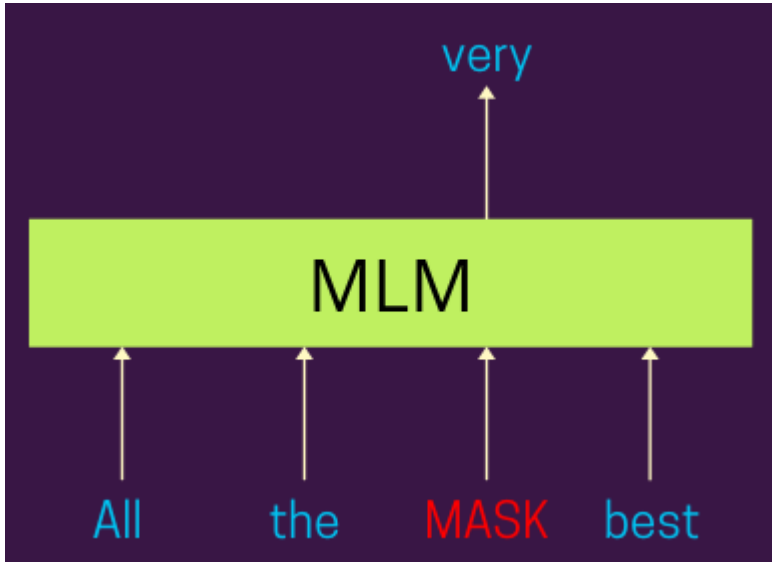
- Classification with Vision Transformers
- Object Detection
- Image Generation

# VISION TRANSFORMERS (VIT)

# Self-Supervised Learning

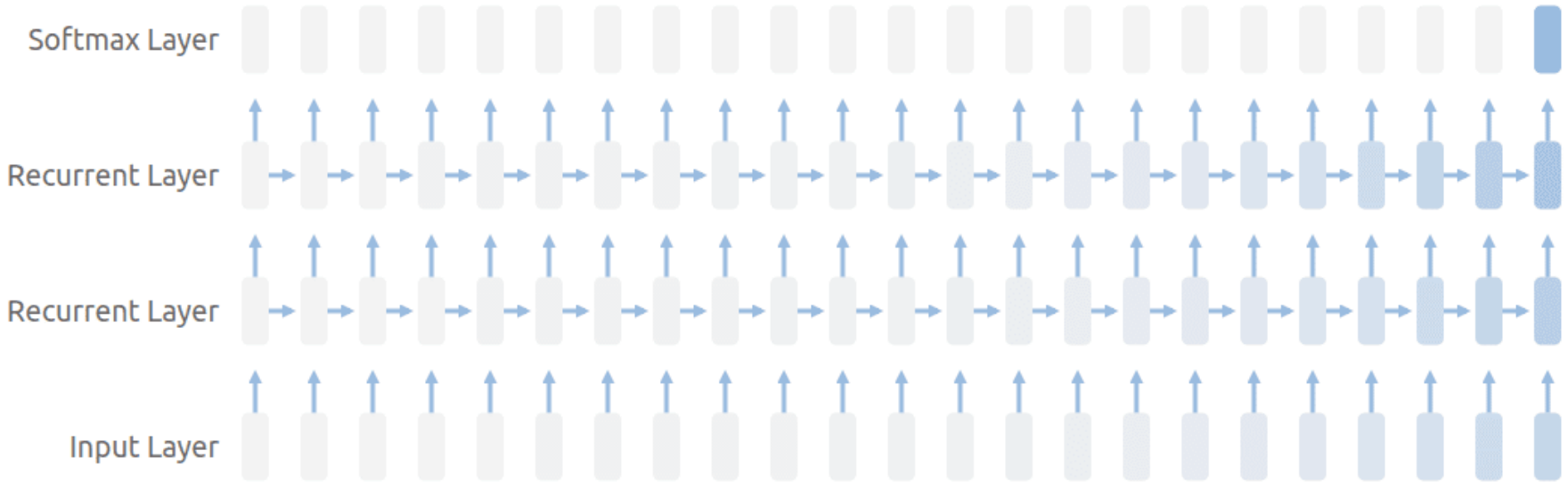
Learning without labels

- “You shall know a word by the company it keeps”.
- NLP Trick: Training a model to predict a masked word based on the neighboring words.
- Train on a large corpus
- Fine-tune on a down-stream task:
  fewer labeled samples are needed!



# RNN

## And the vanishing gradient problem




**Vanishing Gradient:** where the contribution from the earlier steps becomes insignificant in the gradient for the vanilla RNN unit.

# Attention

Classic attention:

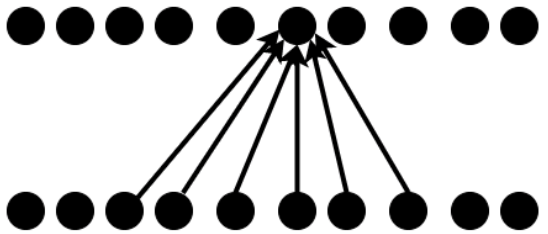
Unlike Convolution, values are constantly changing – depending on the input sequence.

Uses some specific function:

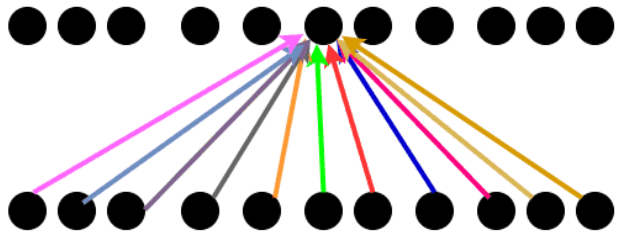
Cosine//softmax//tanh …

Also called “fast-weights”.

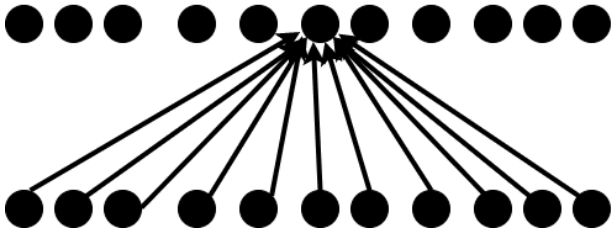
Convolution



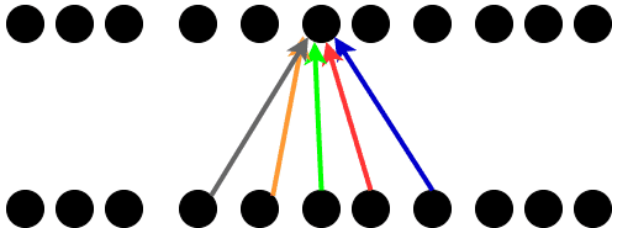
Global attention



Fully Connected layer



Local attention



https://theaisummer.com/attention/

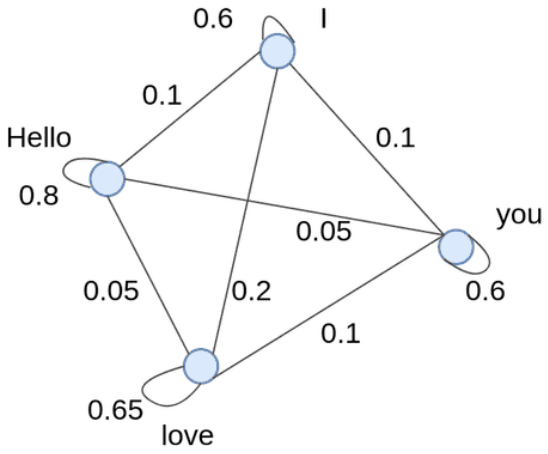
# Self-Attention

Self-attention
Probability score matrix

| | Hello | I | love | you |
|---|---|---|---|---|
| Hello | 0.8 | 0.1 | 0.05 | 0.05 |
| I | 0.1 | 0.6 | 0.2 | 0.1 |
| love | 0.05 | 0.2 | 0.65 | 0.1 |
| you | 0.2 | 0.1 | 0.1 | 0.6 |

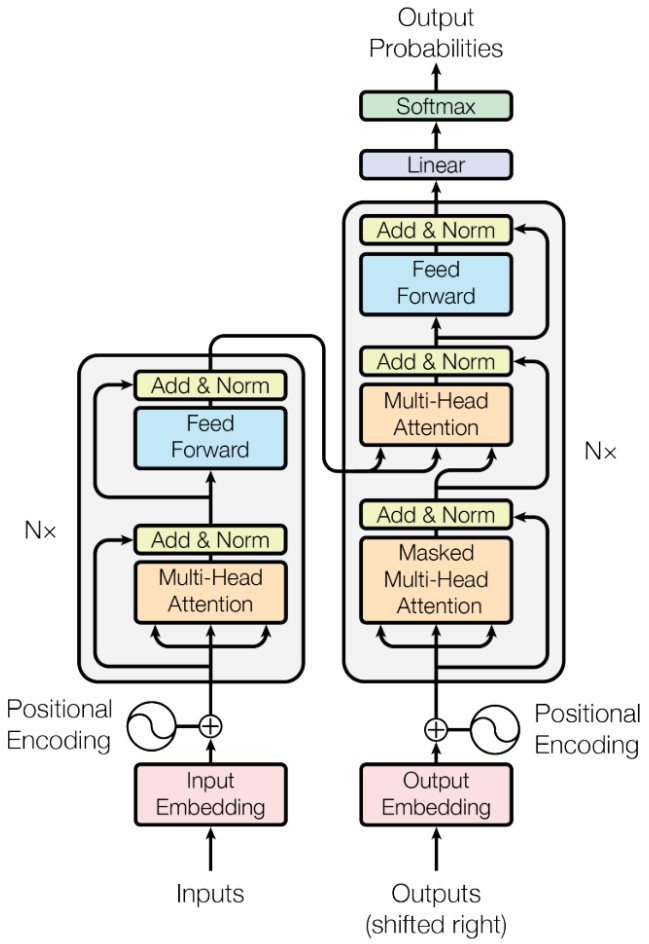
Softmax(Attention)
equation

Self-attention as a
undirected weighted graph



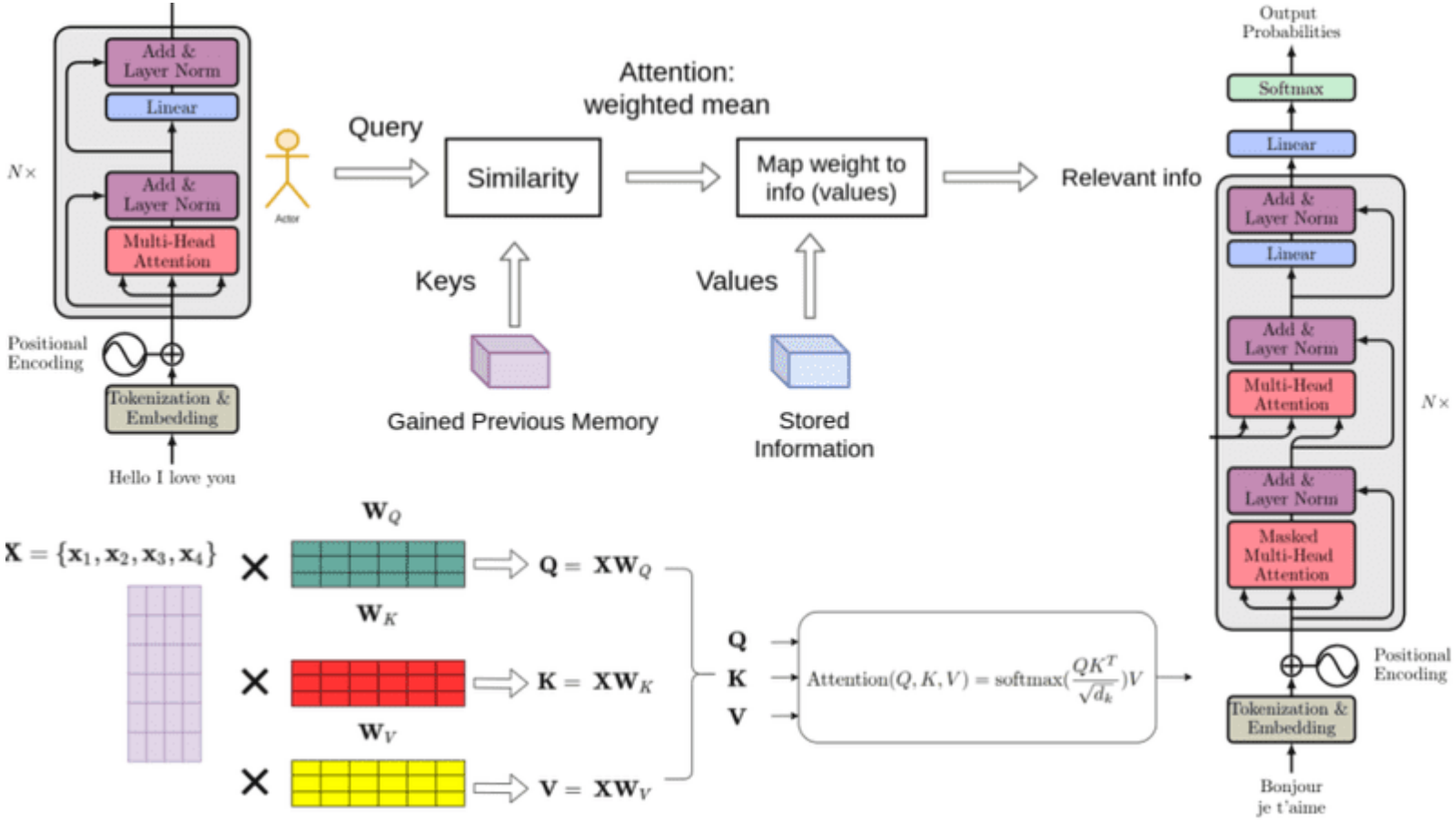
# Attention is all you need

2015



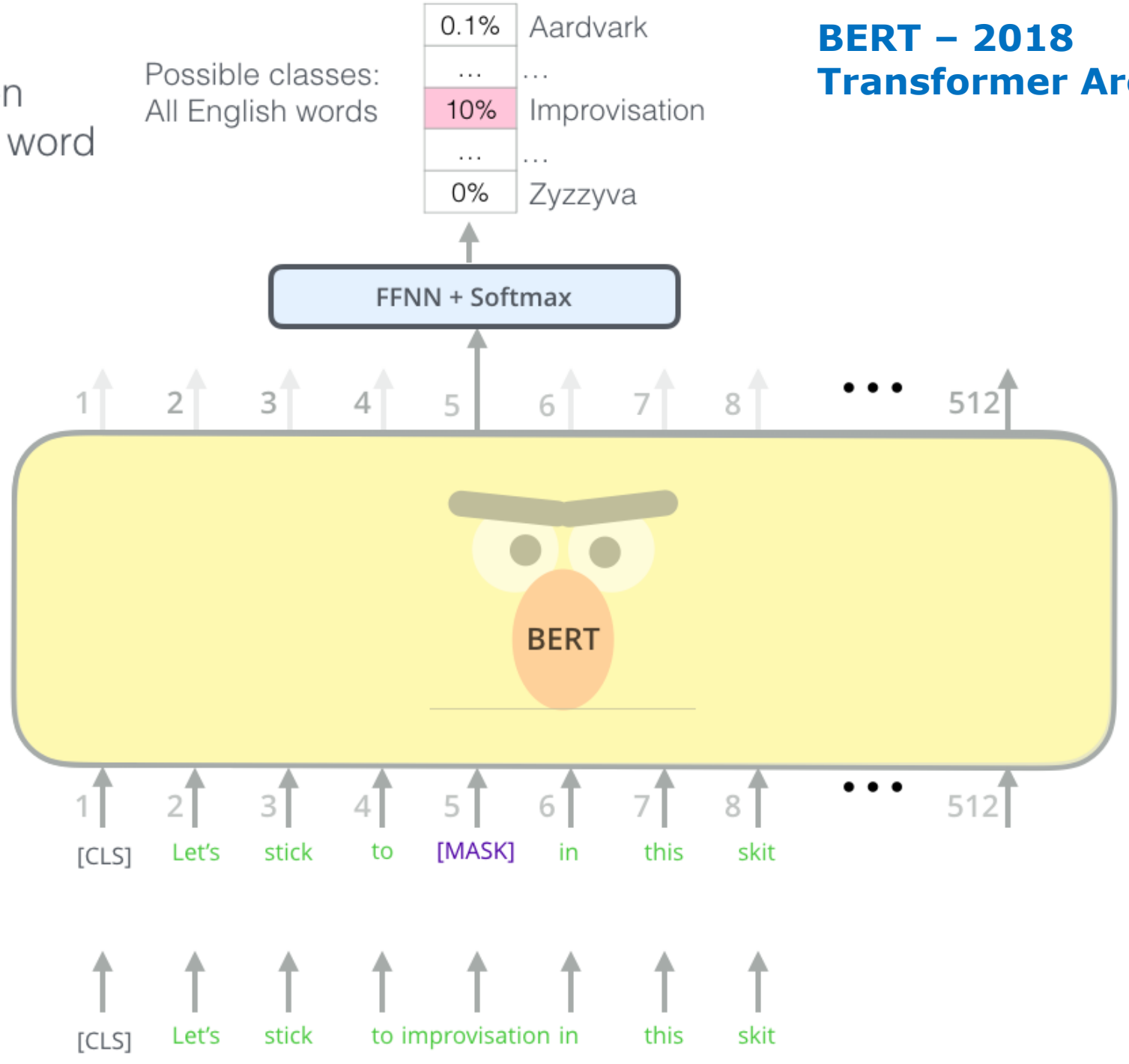
# Transformer Architecture

- How Transformers work in deep learning and NLP: an intuitive introduction | AI Summer (theaisummer.com)
- How Attention works in Deep Learning: understanding the attention mechanism in sequence models | AI Summer (theaisummer.com)
- 3D Medical image segmentation with transformers tutorial | AI Summer (theaisummer.com)



# BERT – 2018
# Transformer Architecture

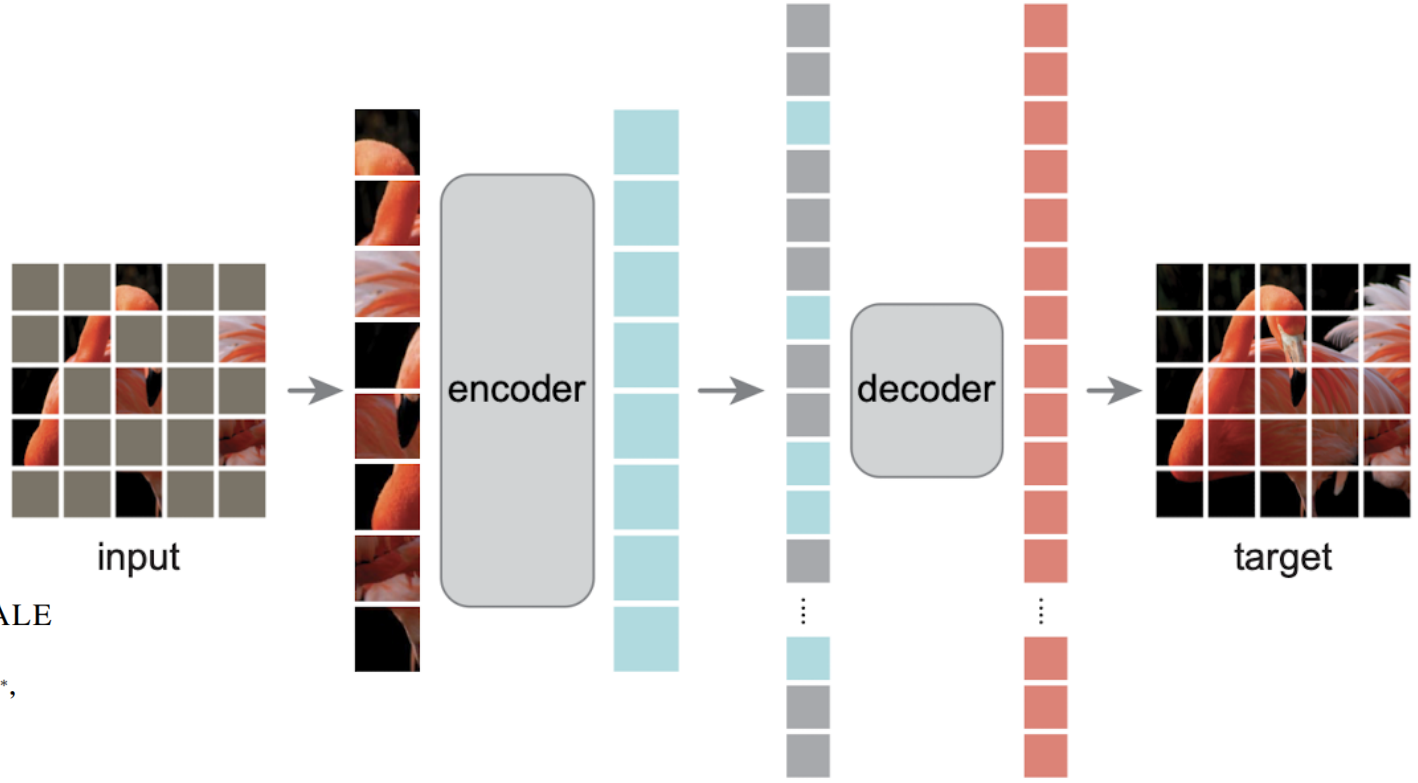
Use the output of the masked word's position to predict the masked word



# Self-Supervised Learning

## Learning without labels

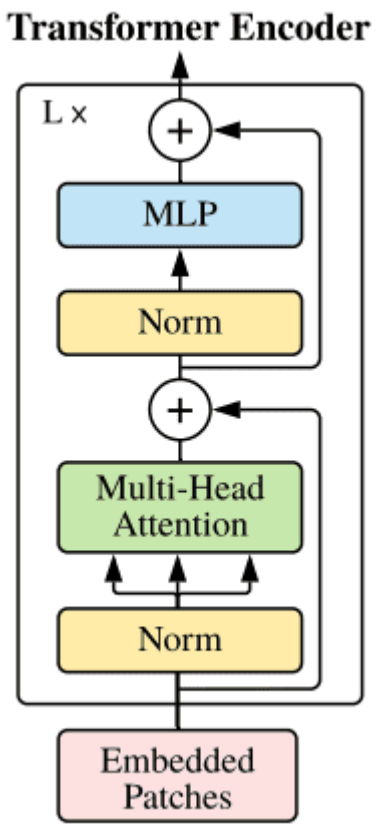
- How to translate it into CV:
  - Masking a pixel?
  - Masking a $n \times n$ patch?




2020:

AN IMAGE IS WORTH 16X16 WORDS:
TRANSFORMERS FOR IMAGE RECOGNITION AT SCALE

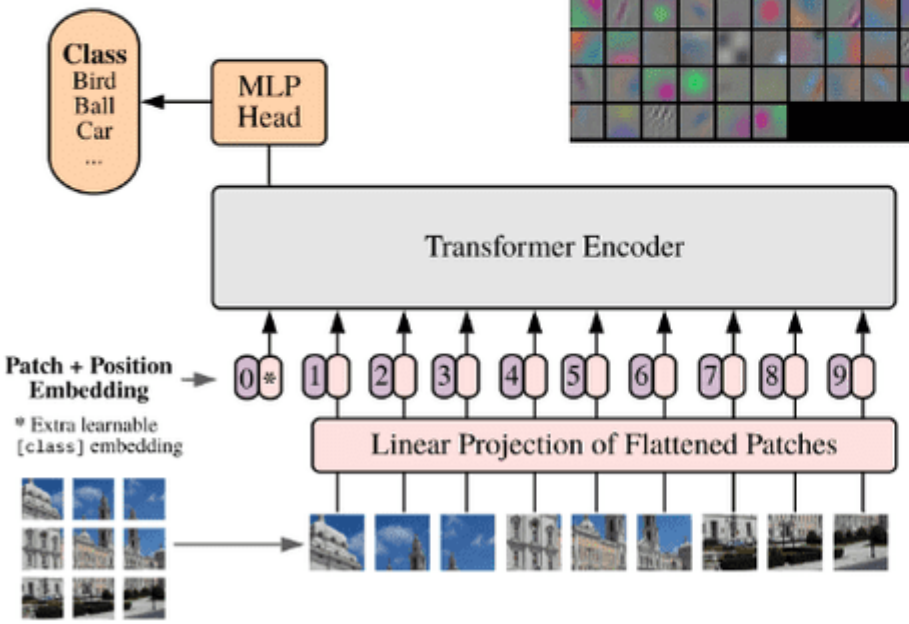
**Alexey Dosovitskiy**[*,†], **Lucas Beyer**[*], **Alexander Kolesnikov**[*], **Dirk Weissenborn**[*],
**Xiaohua Zhai**[*], **Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer,**
**Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, Neil Houlsby**[*,†]
[*]equal technical contribution, [†]equal advising
Google Research, Brain Team
{adosovitskiy, neilhoulsby}@google.com


Transformer Encoder
L x
MLP
Norm
Multi-Head Attention
Norm
Embedded Patches

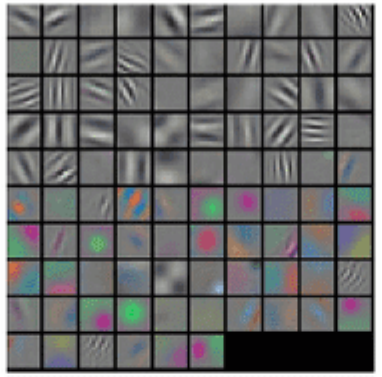
# Self-Supervised Learning

An image is worth 16x16 words

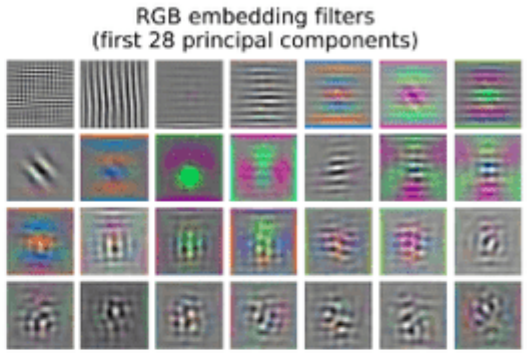
- Vision Transformer
- (18) An image is worth 16x16 words: ViT | Is this the extinction of CNNs? Long live the Transformer? - YouTube
- How the Vision Transformer (ViT) works in 10 minutes: an image is worth 16x16 words | AI Summer (theaisummer.com)




Alexnet 1st conv filters



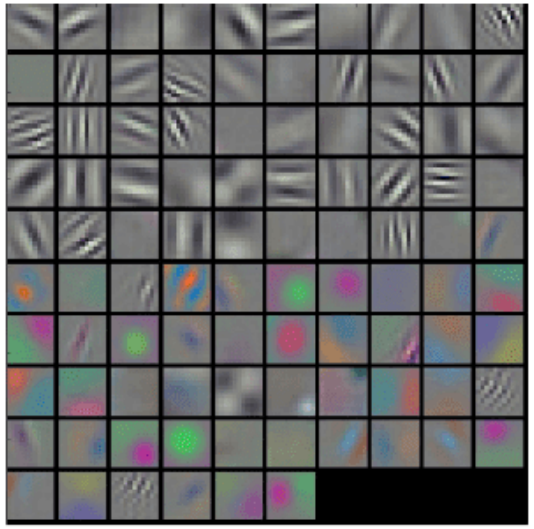
ViT 1st linear embedding filters



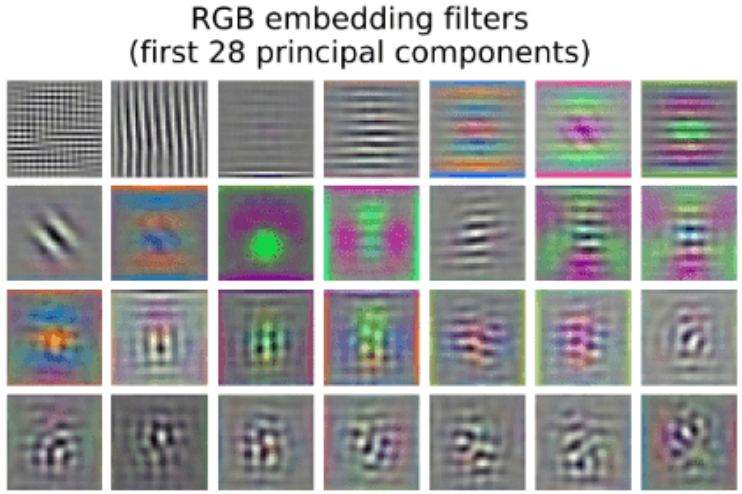
# 1st-Layer Conv Filters

## AlexNet filters vs ViT filters

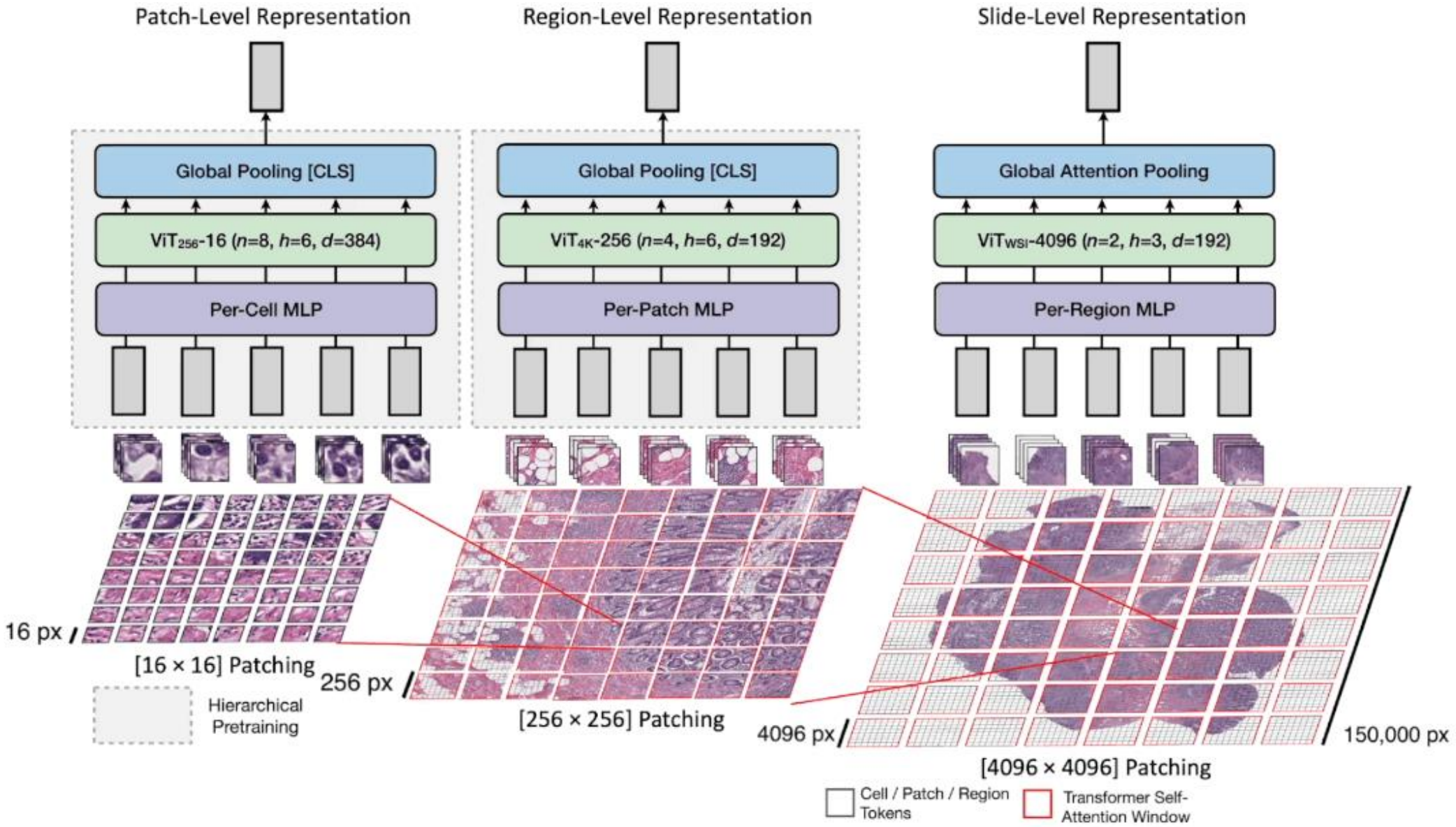
Alexnet 1st conv filters



ViT 1st linear embedding filters

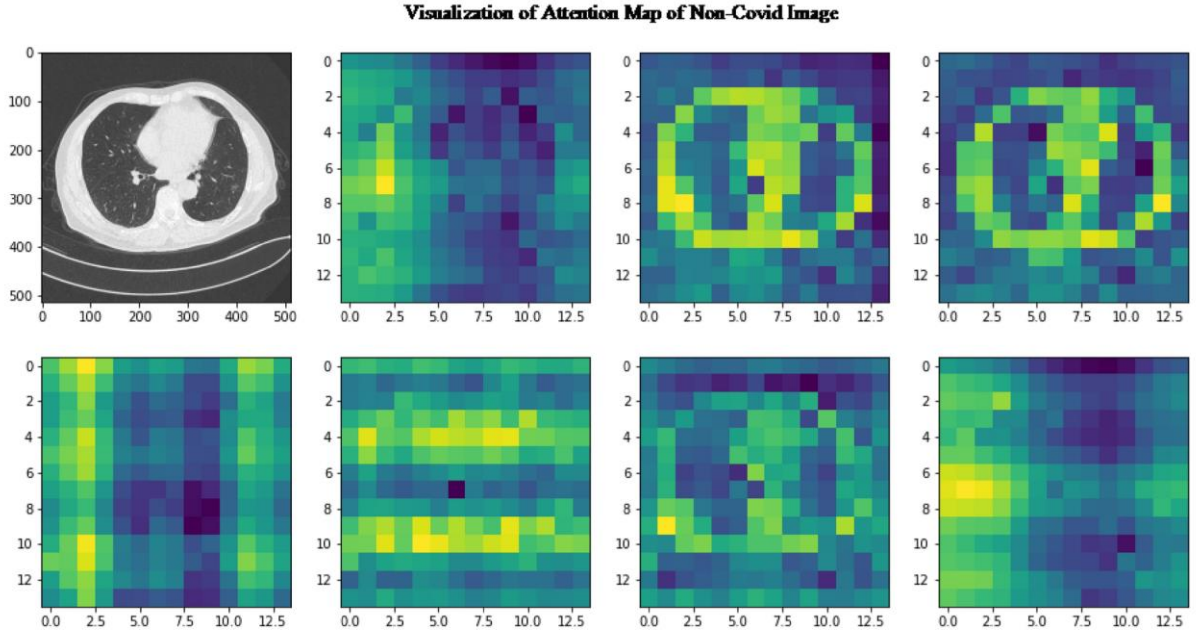


# Hierarchical Image Pyramid Transformer (HIPT)

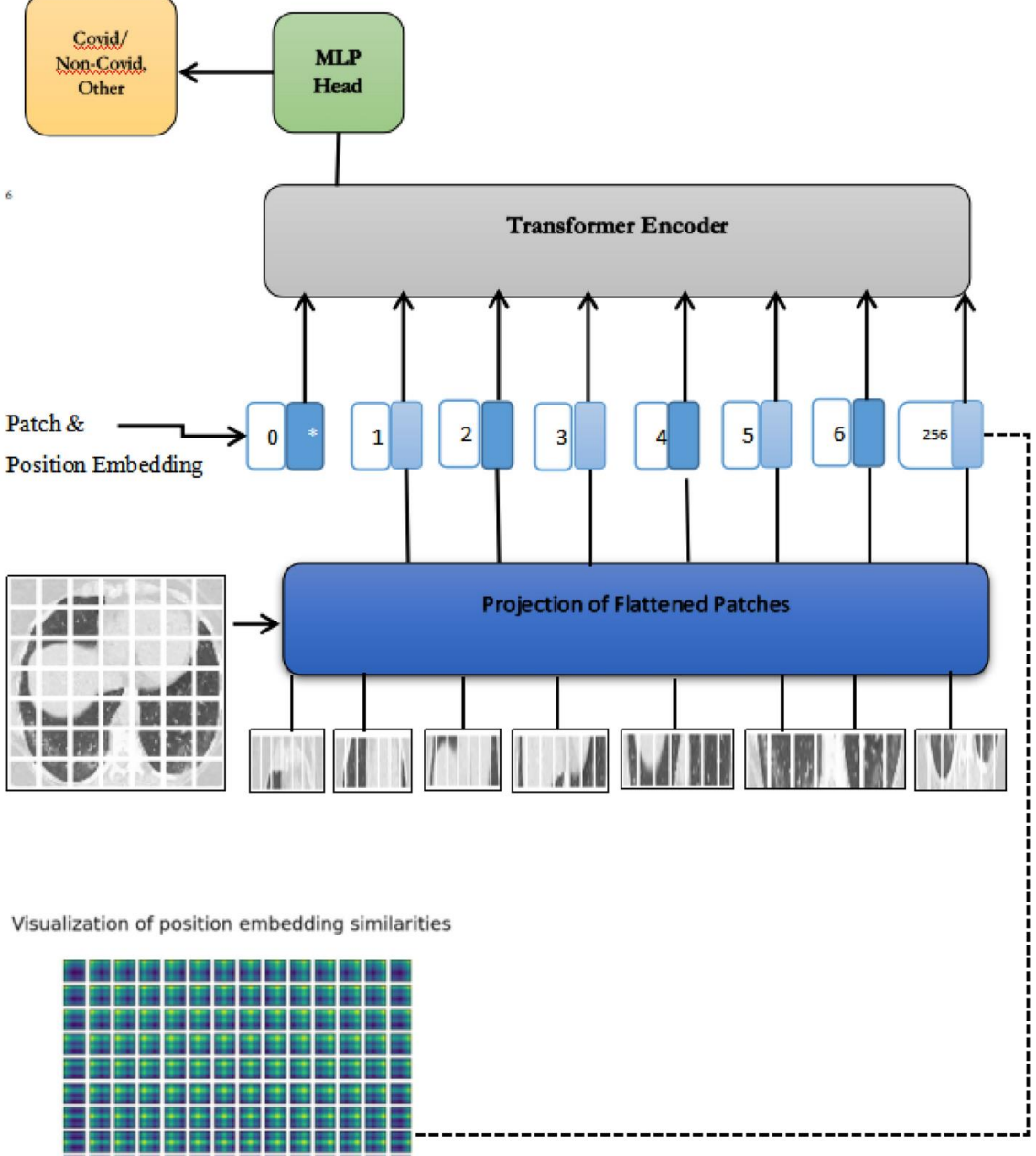


# CoV-19 Detection

2022




(b): Visualization of Attention Map of Non-Covid Image



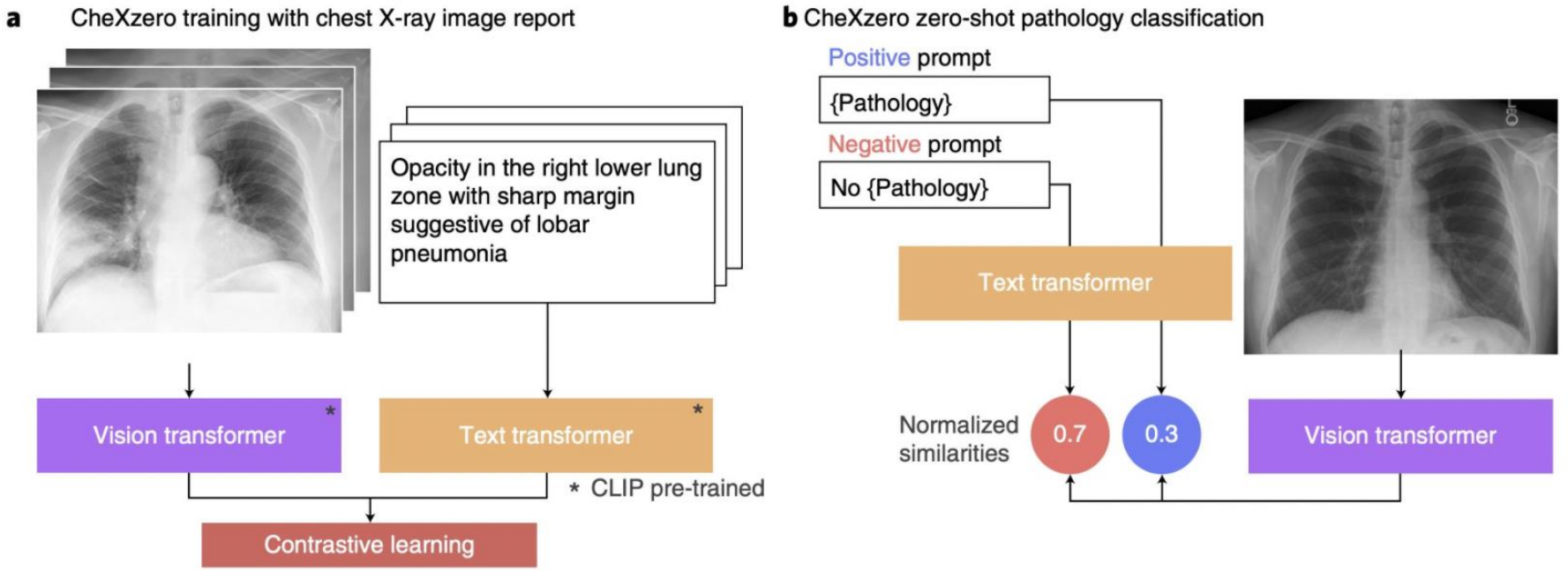

Visualization of position embedding similarities



[Towards robust diagnosis of COVID-19 using vision self-attention transformer | Scientific Reports (nature.com)]

# Self-Supervised Learning

[Expert-level detection of pathologies from unannotated chest X-ray images via self-supervised learning | Nature Biomedical Engineering]



imc FH KREMS
UNIVERSITY OF APPLIED SCIENCES

# R-CNN

How does DL used for Image Recognition?

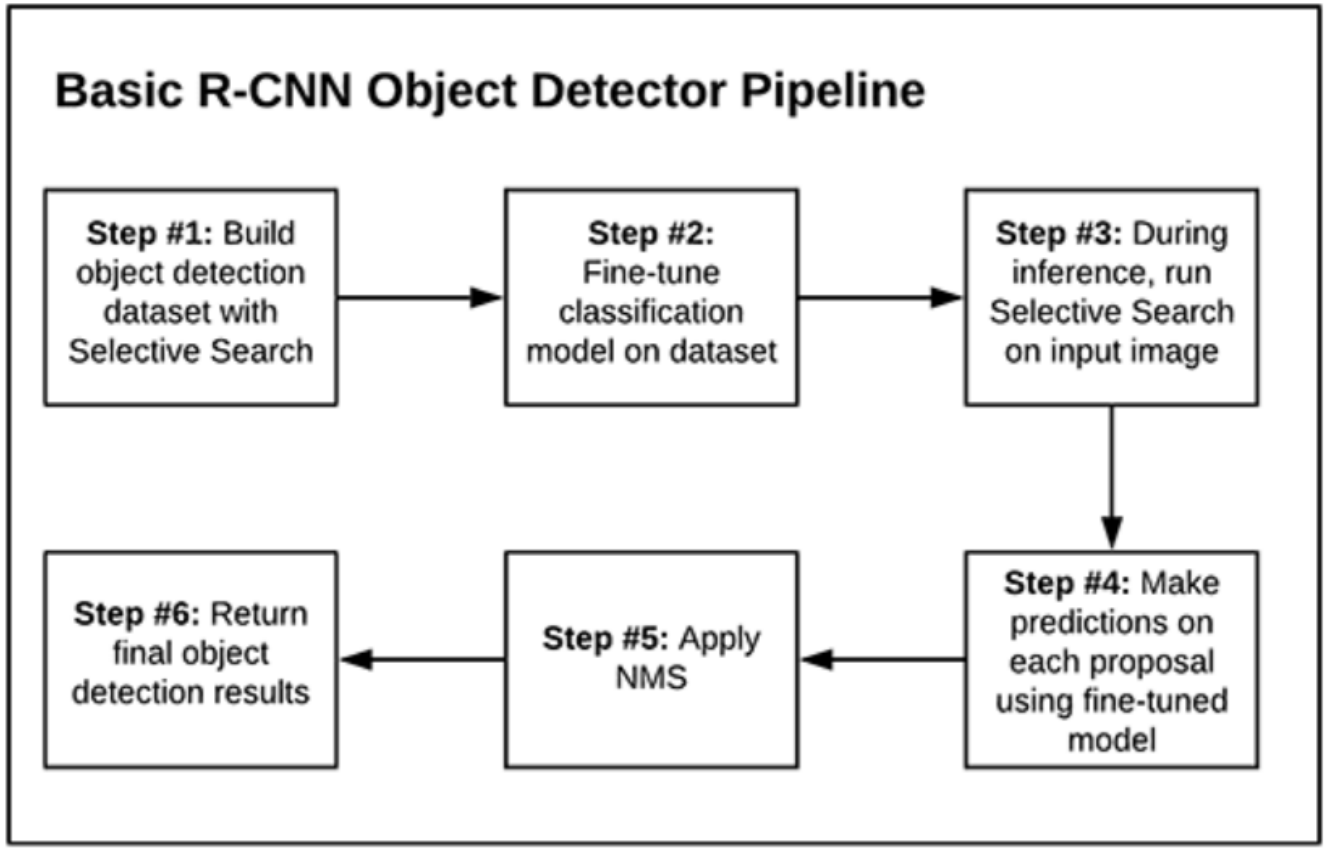
# Image Recognition

So far, we've discussed classification & Segmentation…

- How is Image Recognition done with DL?

# Image Recognition with DL

## R-CNN



# R-CNN

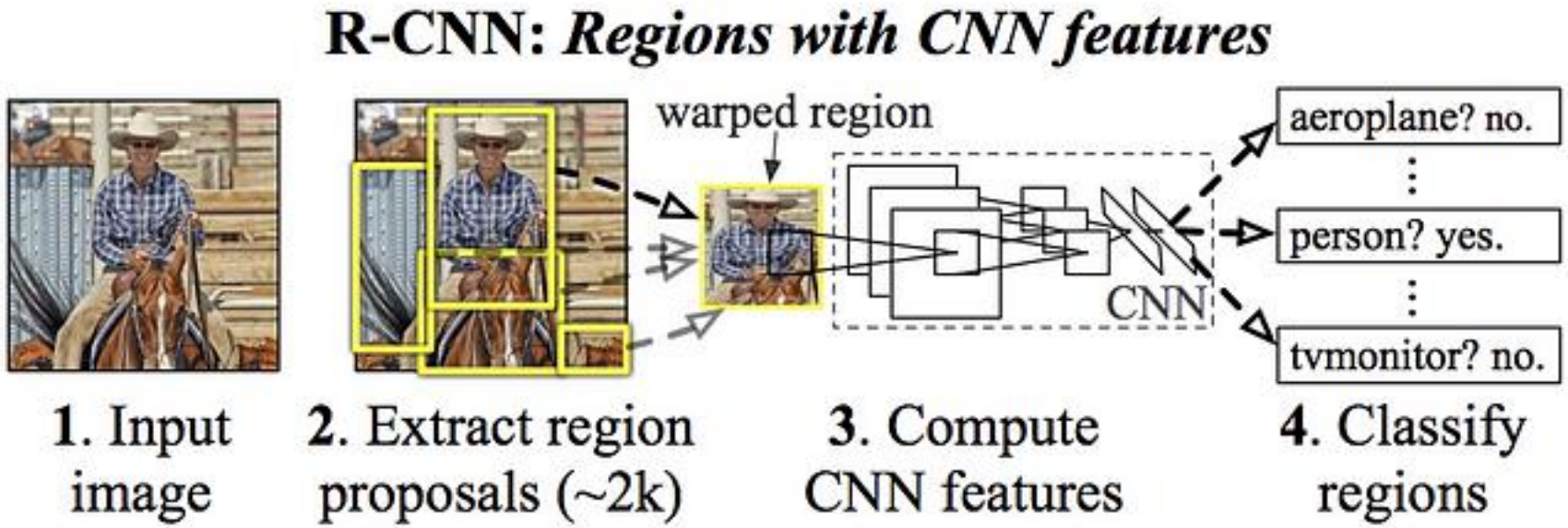
## Region proposal - getting square candidates


Figure 2: Two examples of our selective search showing the necessity of different scales. On the left we find many objects at different scales. On the right we necessarily find the objects at different scales as the girl is contained by the tv.

UijlingsIJCV2013.pdf (uva.nl)

# R-CNN

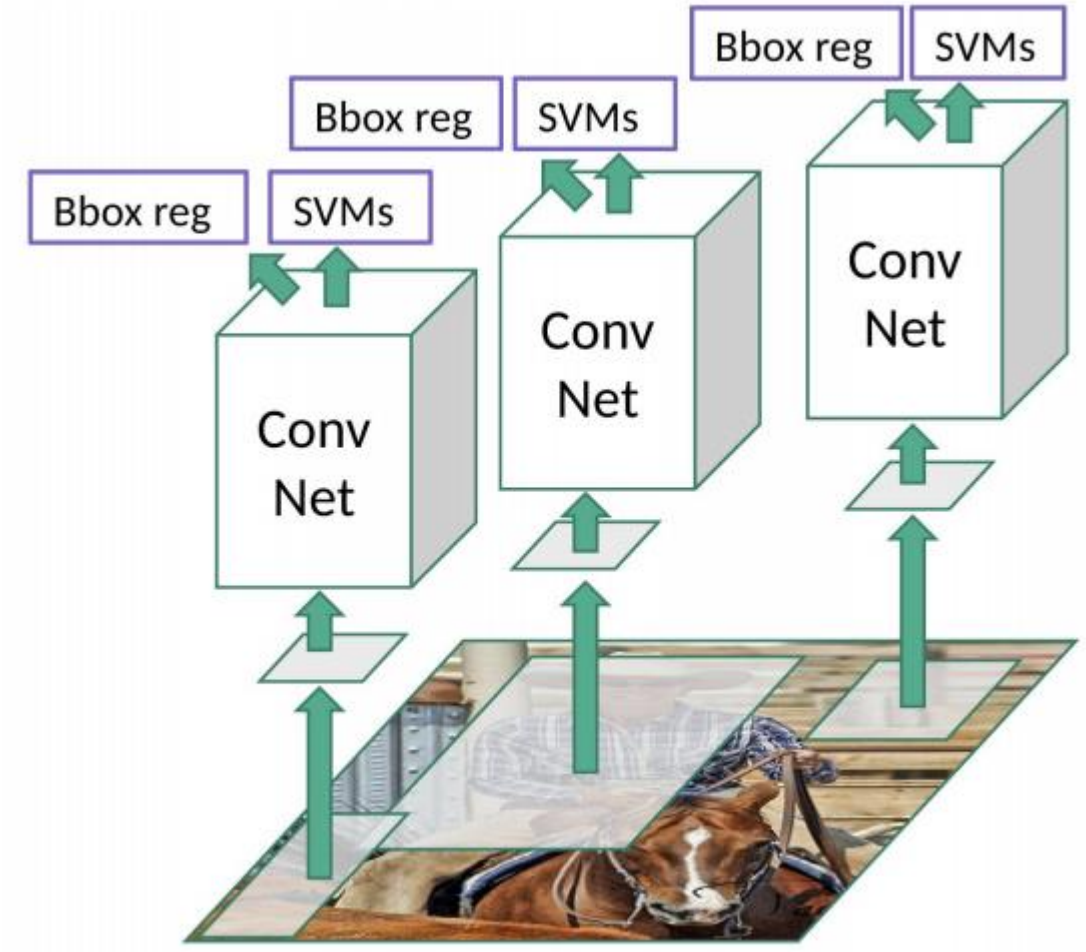
## The CNN part



# R-CNN

## SVM layer

- The region candidates goes through SVM
- Classify if the object is in this bounding-box square

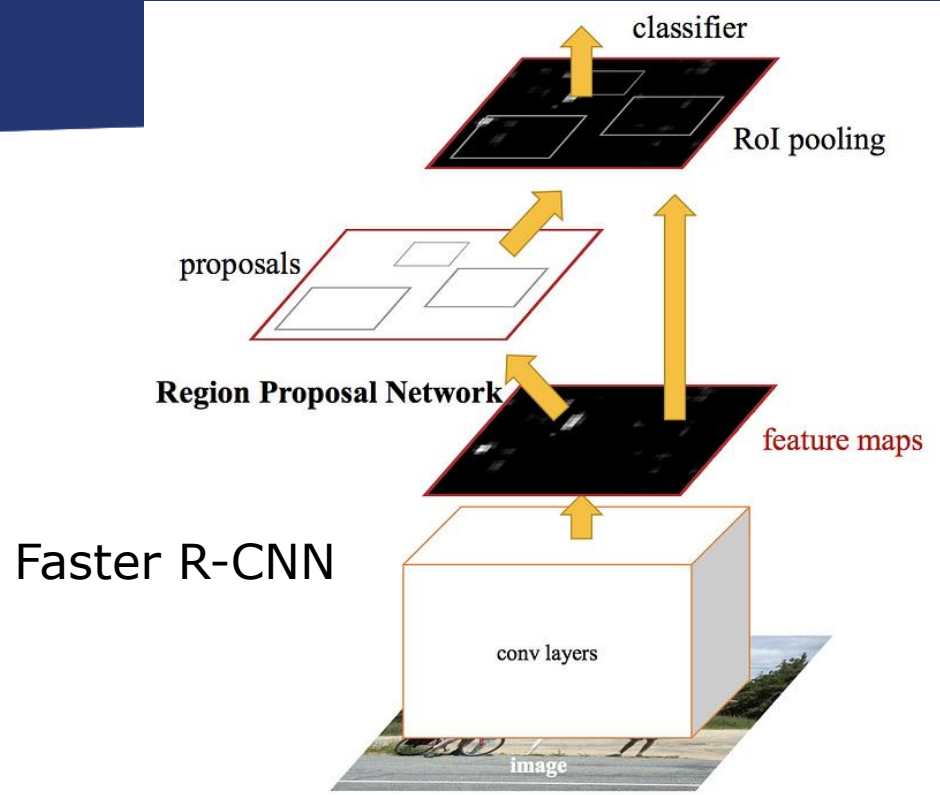


# Image Recognition

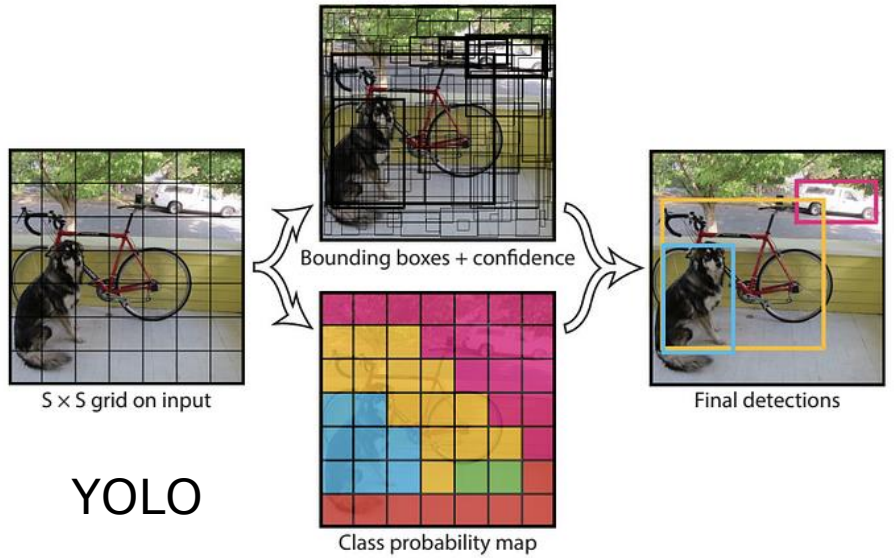
## R-CNN vs the rest

- Issues with R-CNN:
  - Slow: extracting and processing 2000 candidates takes time
  - Recognition for an image takes nearly a minute
  - The extraction is a fixed process – without learning

- Other (better) implementation exist:
  - Fast-R-CNN
  - Faster R-CNN
    - Is used for OCR
  - YOLO – You Only Look Once
    - Does not work well for overlapping objects



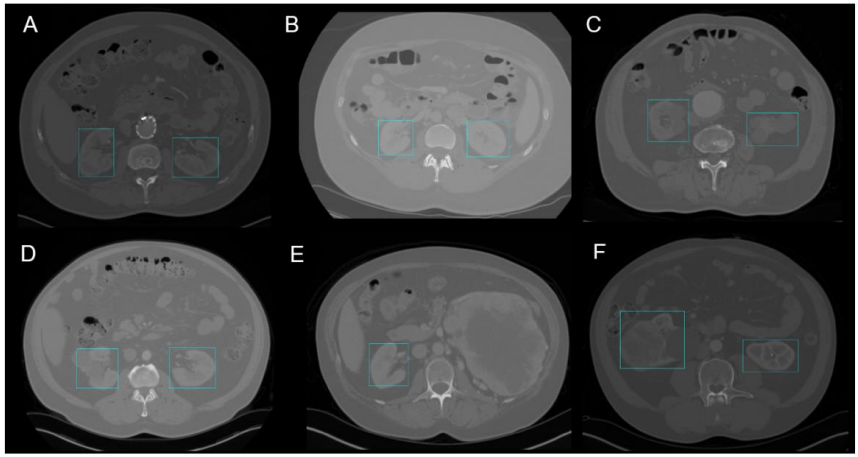

Faster R-CNN




YOLO

# YOLO

The basic idea



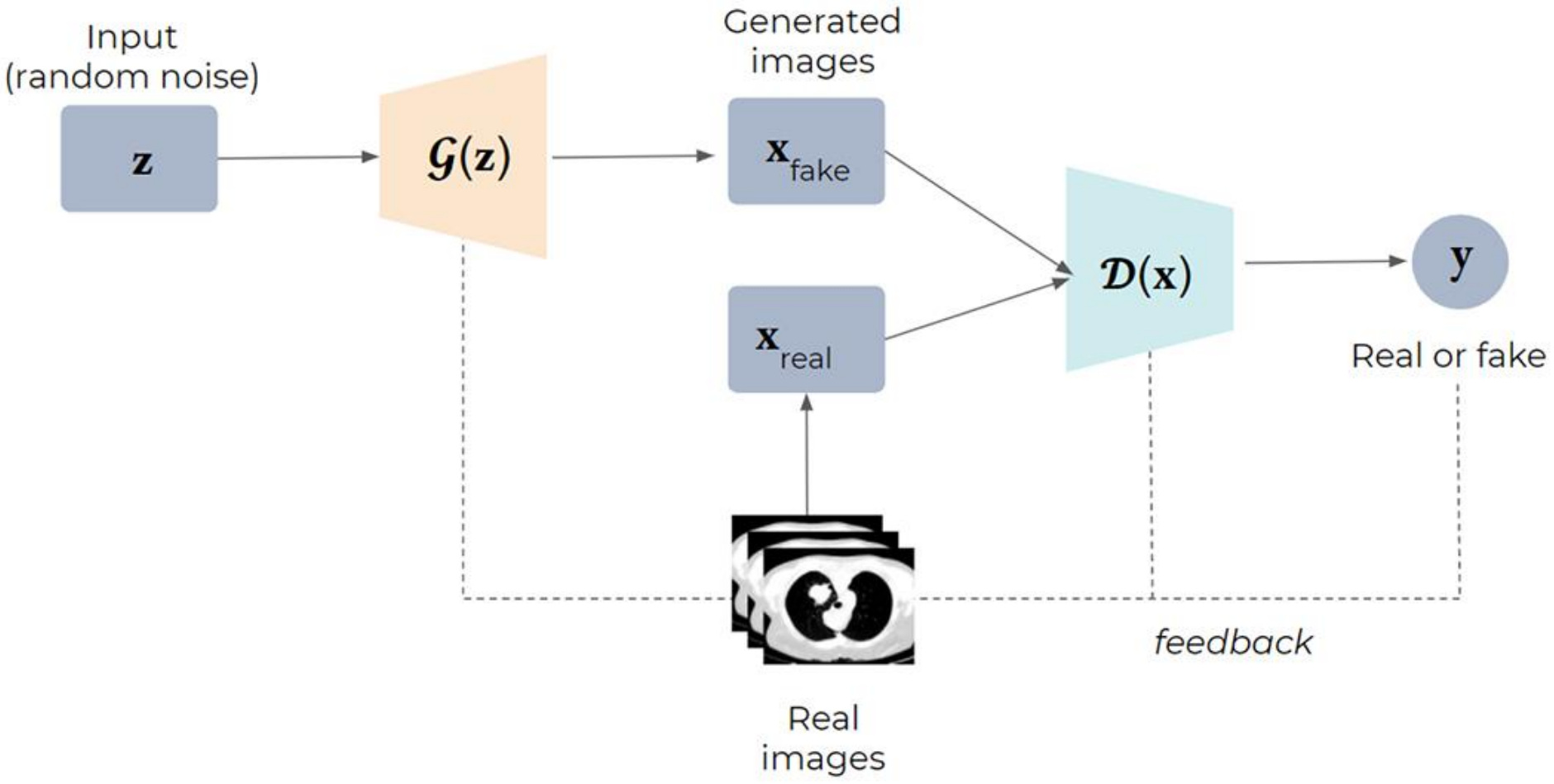

[1910.01268] Kidney Recognition in CT Using YOLOv3 (arxiv.org)

- Divide to Residual blocks
  - NxN grid cells – like in Captchas (hint, hint…)
- Bounding box regression
  - Select the boxes with objects and regress their sizes to the labeled object, until…
- The Intersections over Unions (IoU) is under a certain threshold
- Non-Max Suppression (NMS)
  - Keep only the boxes with the highest probability score of detection

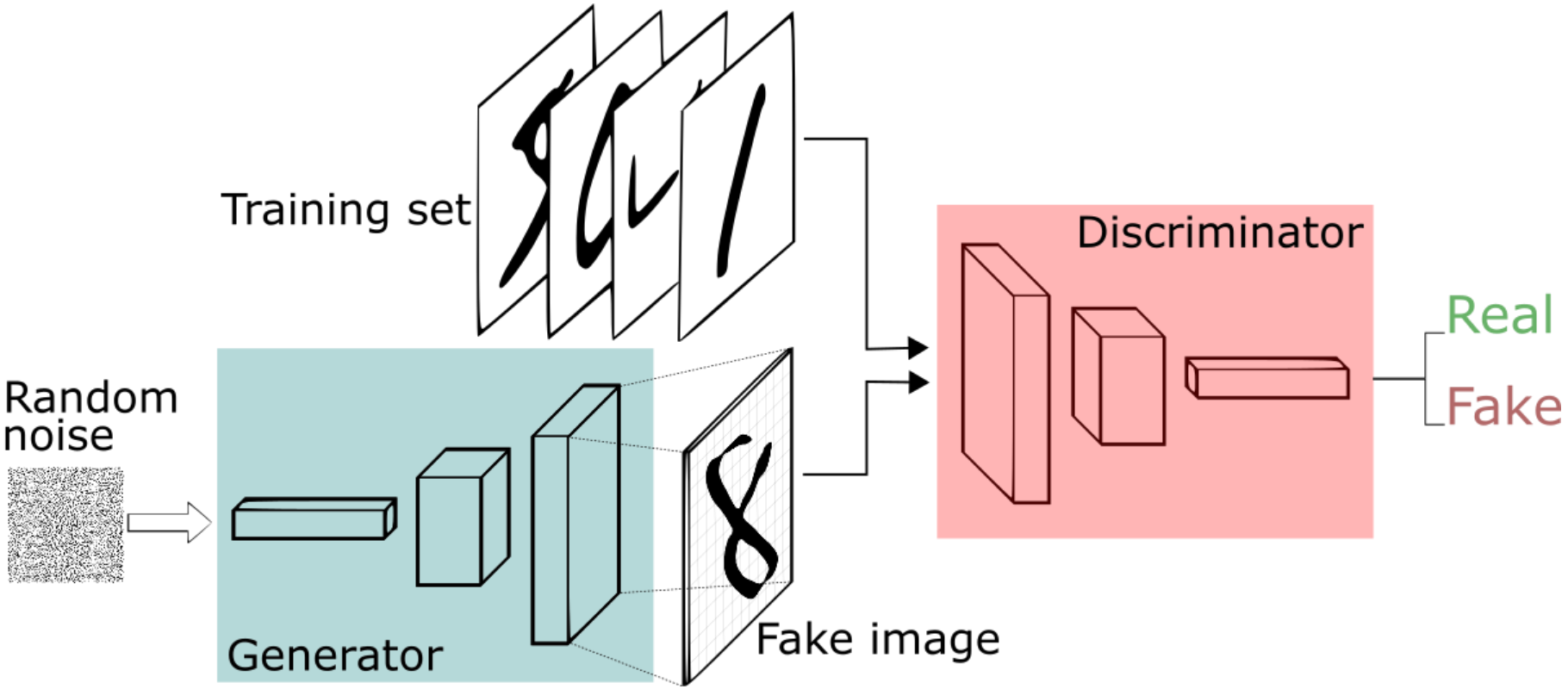
YOLO Object Detection Explained: A Beginner's Guide | DataCamp

# IMAGE SYNTHESIS

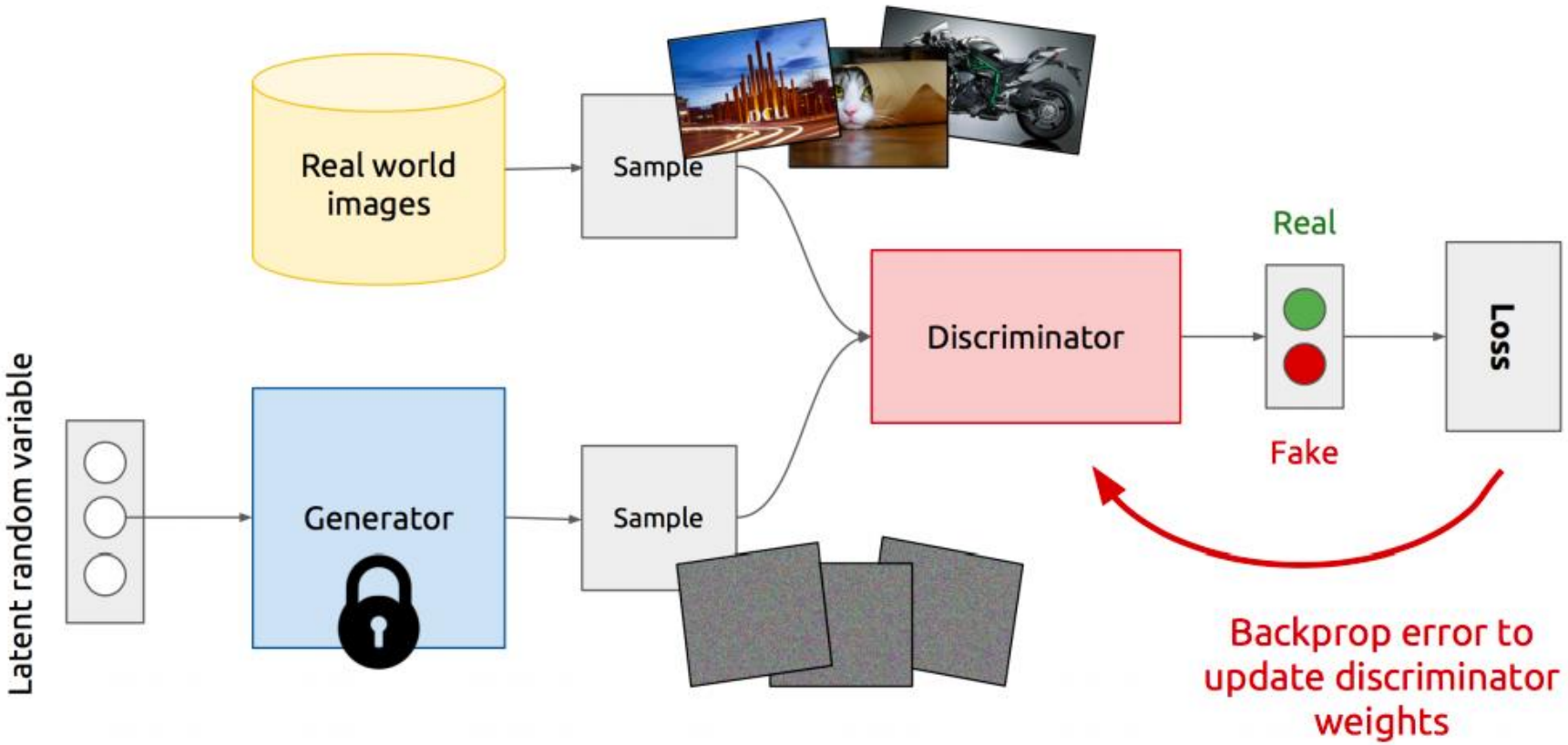
# Generative Adversarial Networks (GANs)



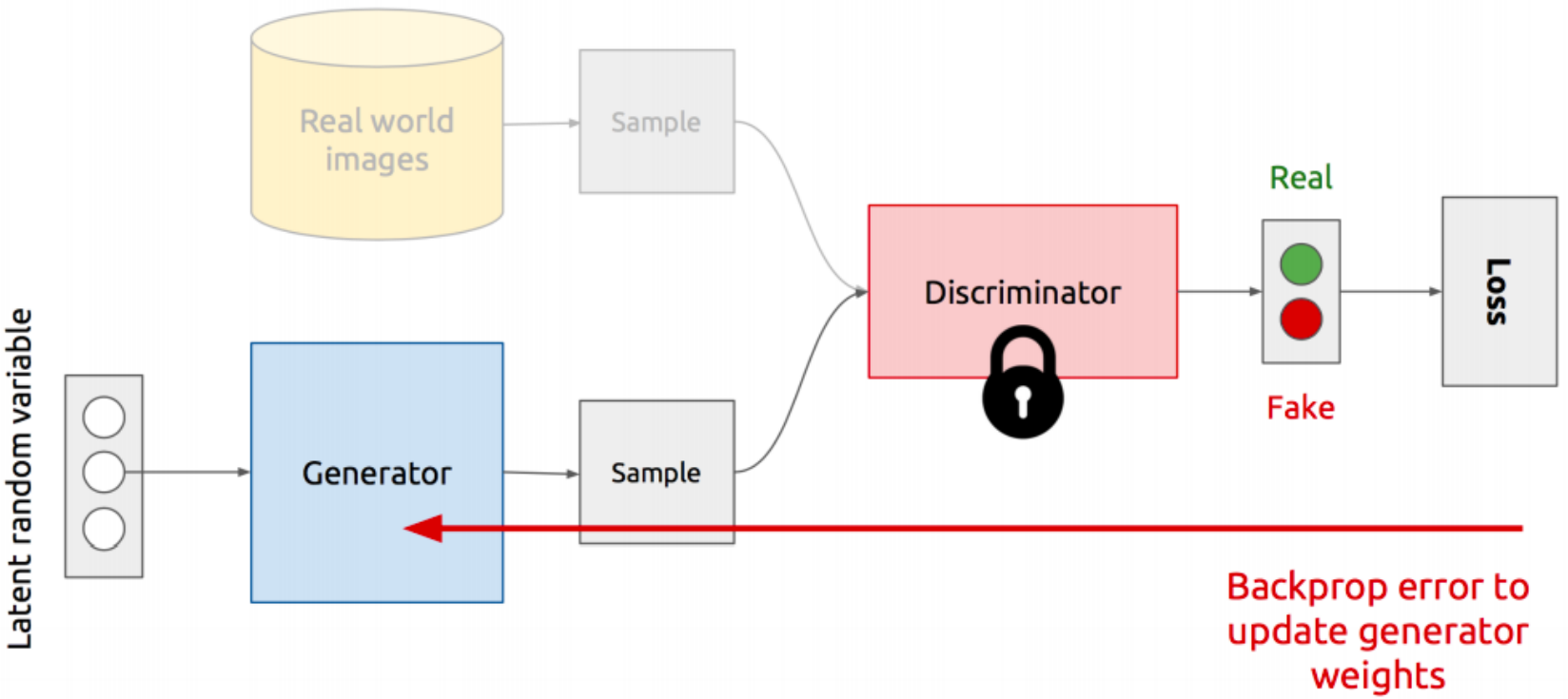
# Generative Adversarial Network



# Training Discriminator

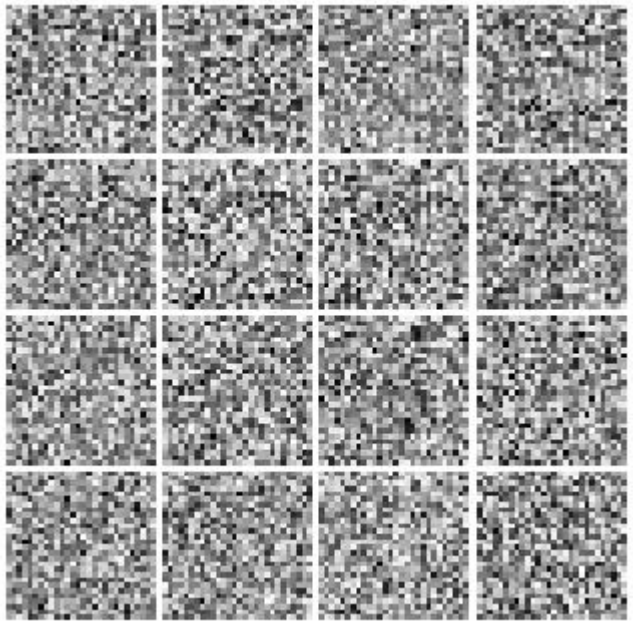


# Training Generator



# Training Process

## Fashion MNIST example



## Development over years

# Generating Human Poses

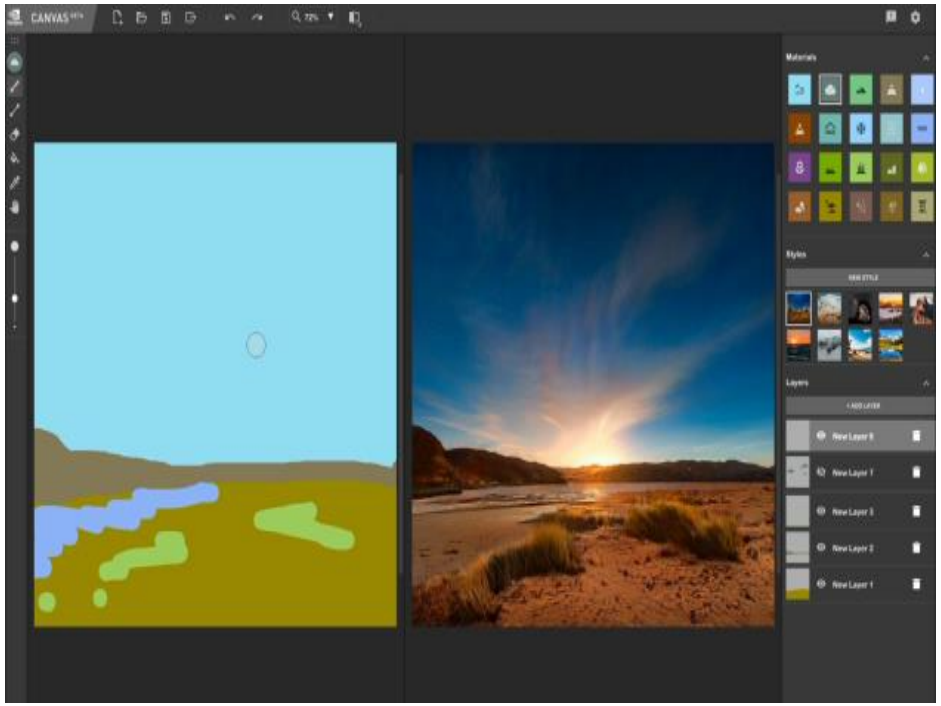
[1705.09368] Pose Guided Person Image Generation (arxiv.org)

# Image-to-Image translation


[1611.07004] Image-to-Image Translation with Conditional Adversarial Networks (arxiv.org)

# NVIDIA Canvas

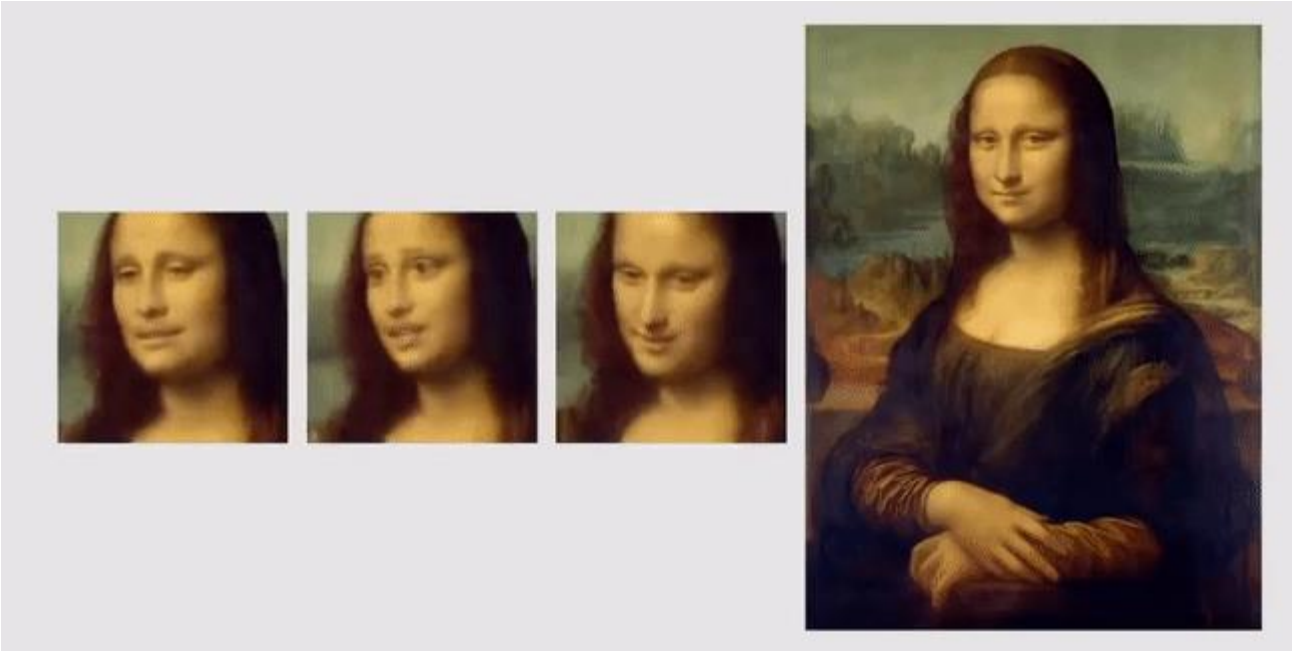
Image-to-image → on steroids



[NVIDIA Canvas: Turn Simple Brushstrokes into Realistic Images]

# GANs for a sequence of images

Because why stop at one?



[1905.08233] Few-Shot Adversarial Learning of Realistic Neural Talking Head Models (arxiv.org)

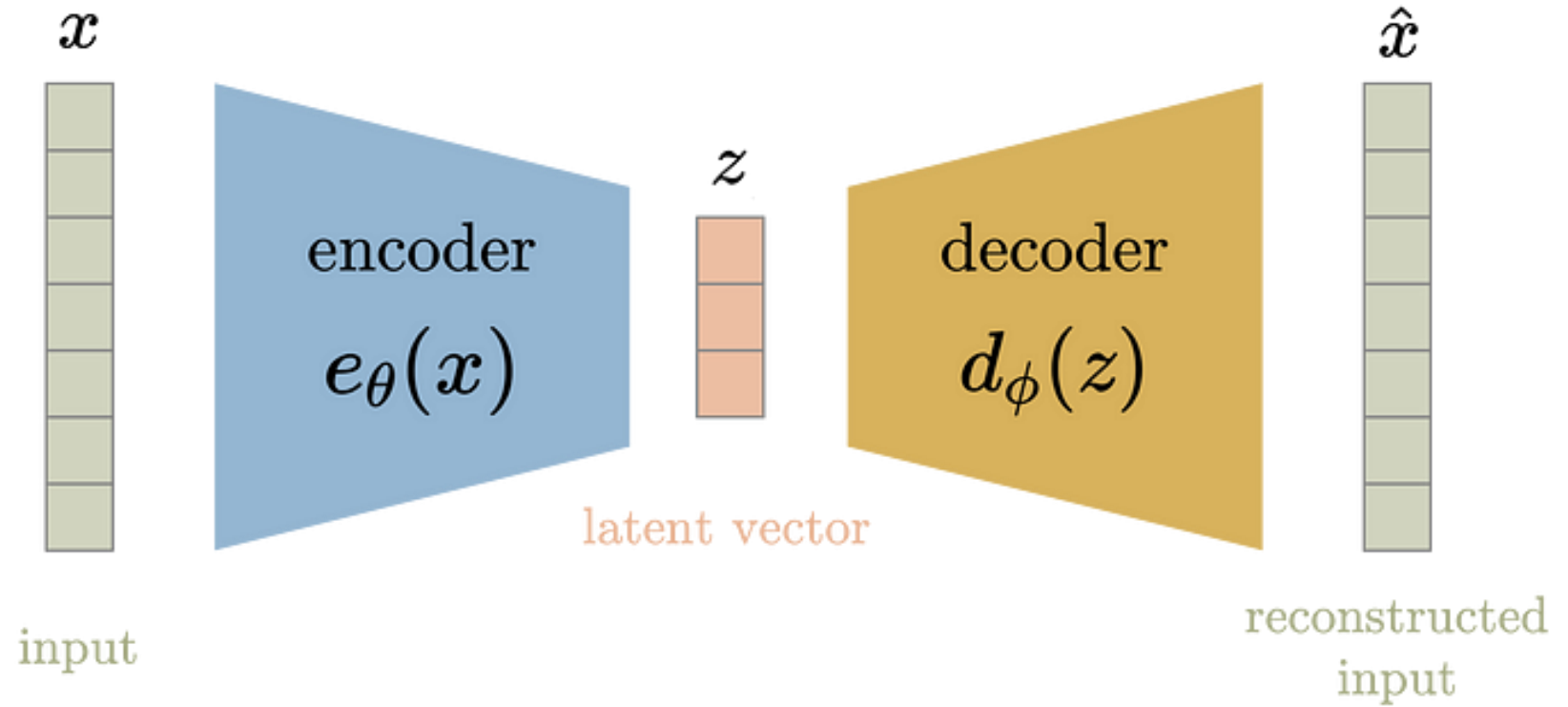
# AUTOENCODER

# Autoencoder

## Compressing Data

- For data compression, PCA uses linear algebra techniques to keep most of the explained variance in the data.
  - Can be used to encode data (e.g., an image) into a smaller vector (dimension reduction).
  - But – only supplies encoding with linearly uncorrelated features
- Can we do it with ML / DL?
  - CNN requires labeled data during training to compare the predicted output and calculated loss…

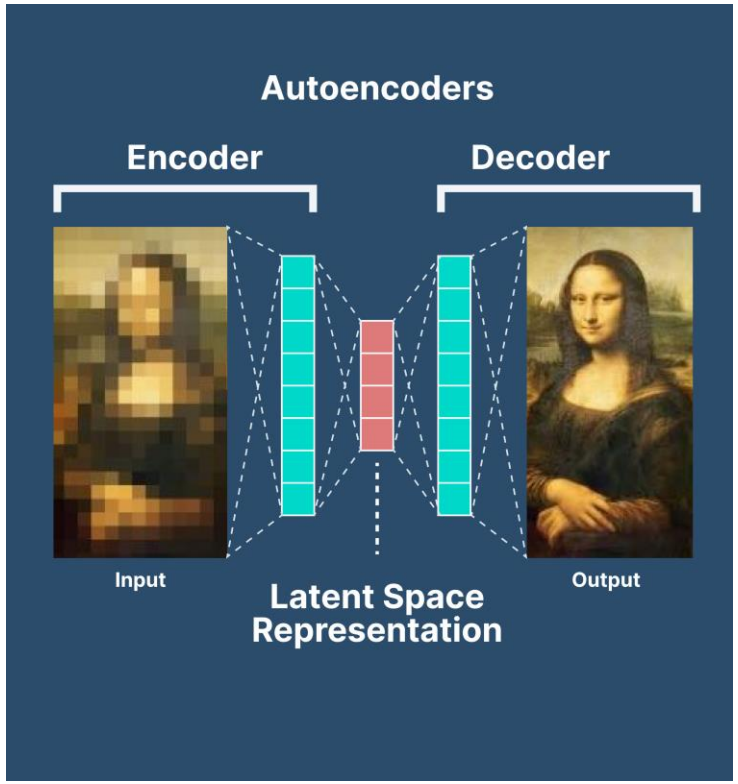
# Autoencoder

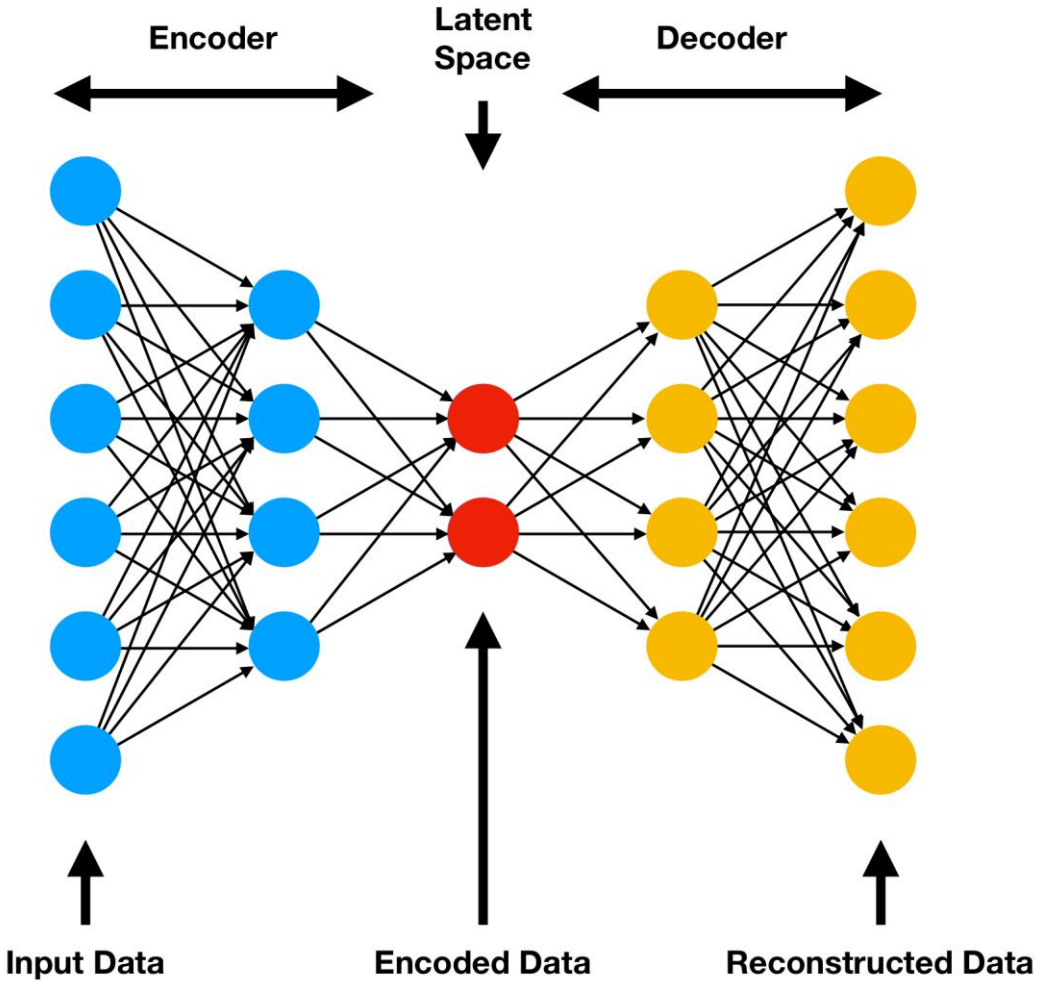



$$loss = \|x - \hat{x}\|_2 = \|x - d_\phi(z)\|_2 = \|x - d_\phi(e_\theta(x))\|_2$$

# Autoencoder

Image in, image out: Squeezing the network



# Autoencoder

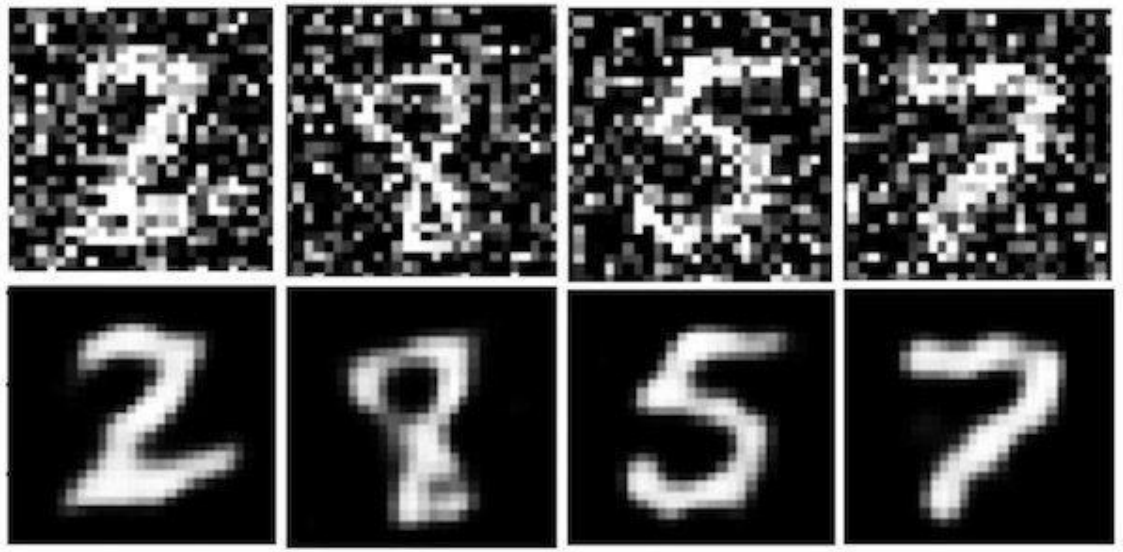
Notes

- Image in, the same image is used as an expected output
  - Loss: Comparing the difference using a reconstructive loss
- The network is conceptually divided into two parts:
  - Encoder – “compress” the image into a smaller representation
  - Decoder – “decompress” a vector back into the image
- The middle layer is significantly smaller than the input/output
  - Also called the “bottleneck”
  - Is used to represent the image on a lower dimensionality space
  - Preserve more semantic features than PCA

# Autoencoder

Many types exist

- Variational Autoencoders (VAE)
  - Regulates/contains the latent representation
- Denoising Autoencoder
  - Artificially noisy image in, original image out
- Deep Autoencoder
- Sparse Autoencoder

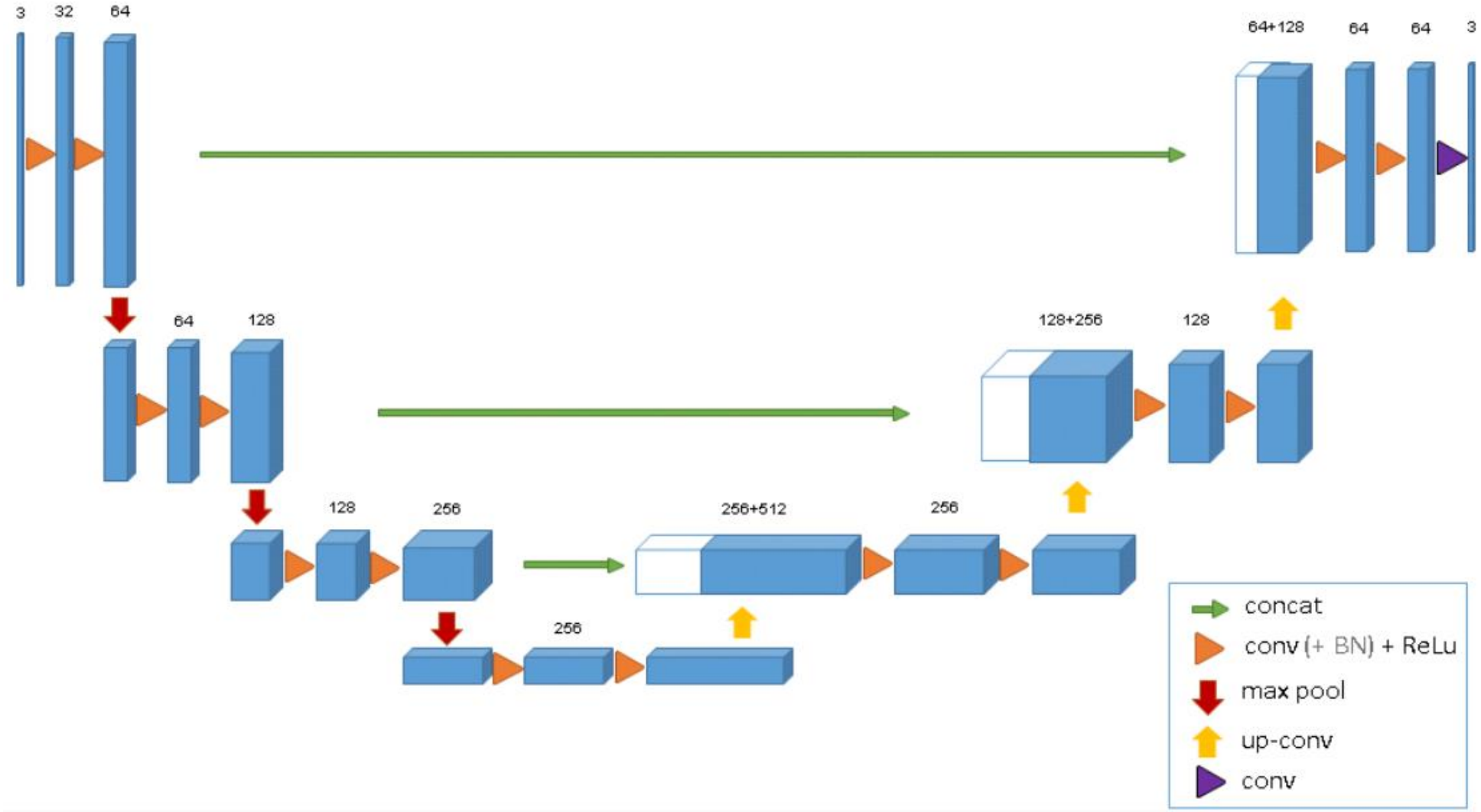


Autoencoders Explained | Baeldung on Computer Science

# Autoencoder

- Worked well for (almost same) image-in → same image-out
- Works less good for image-in → segmentation mapping-out

# U-Net


3
32
64
64+128
64
64
3
64
128
128+256
128
128
256
256+512
256
256
concat
conv (+ BN) + ReLu
max pool
up-conv
conv

# U-Net

The king of Segmentation

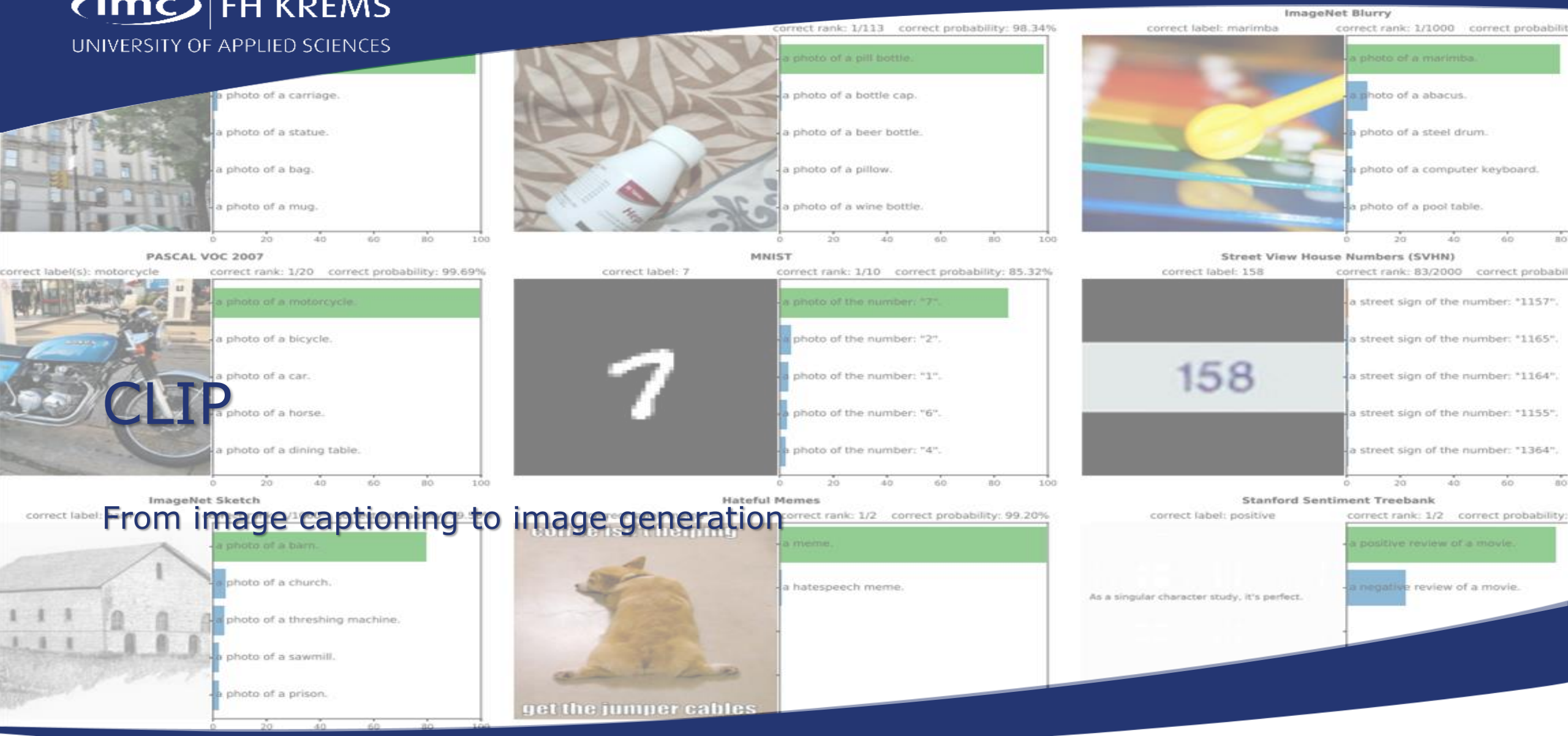
- Skip connections assist the network to “remember” what the original image on that stage (before compression) was

- Is often used for segmentation but can also be used for additional purposes.
  - E.g.:
    - Upscaling an image (to higher resolution),
    - Converting low-dosage X-Ray/CT/PET to high one
    - ‘generating’ PET scan from a CT scan, automatically

# U-Net

- Paper: [1505.04597] U-Net: Convolutional Networks for Biomedical Image Segmentation (arxiv.org)
- U-Net: Convolutional Networks for Biomedical Image Segmentation (uni-freiburg.de)
- Annotated Code: U-Net (labml.ai)

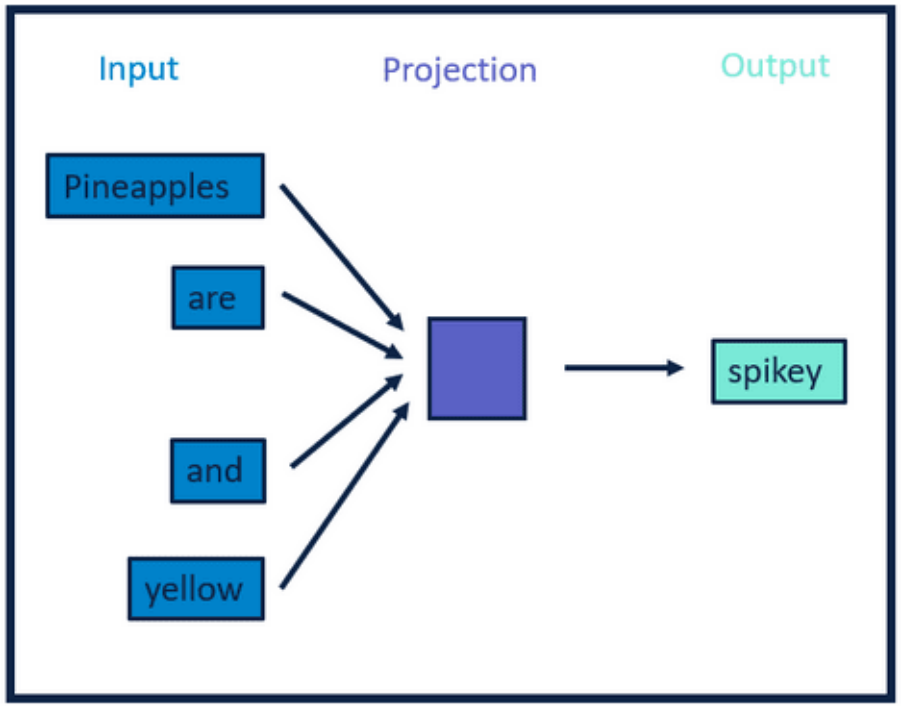
imc FH KREMS
UNIVERSITY OF APPLIED SCIENCES



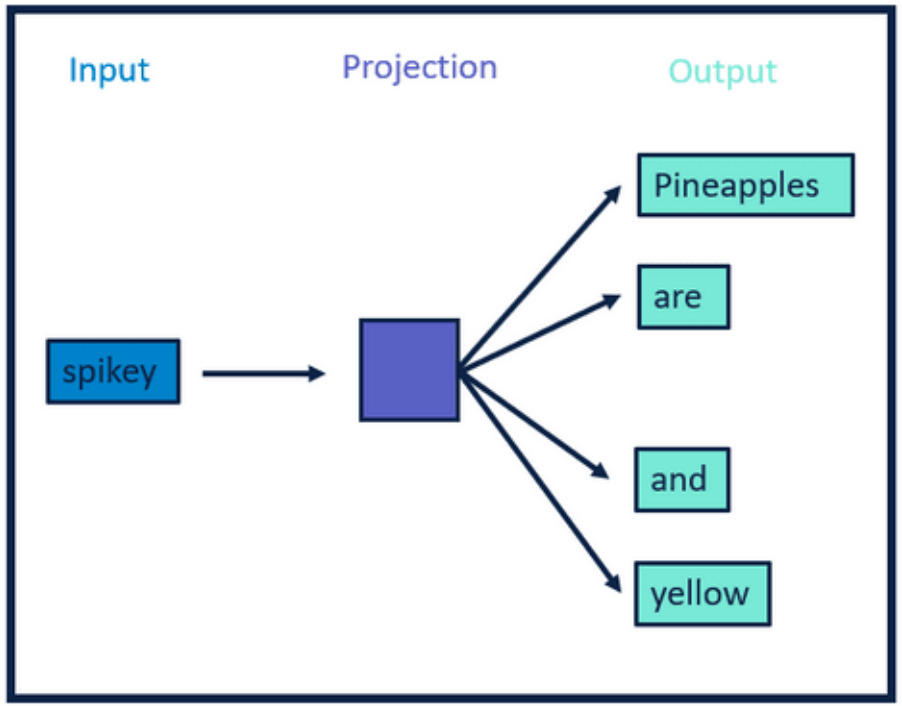

# CLIP

From image captioning to image generation

# Word2Vec



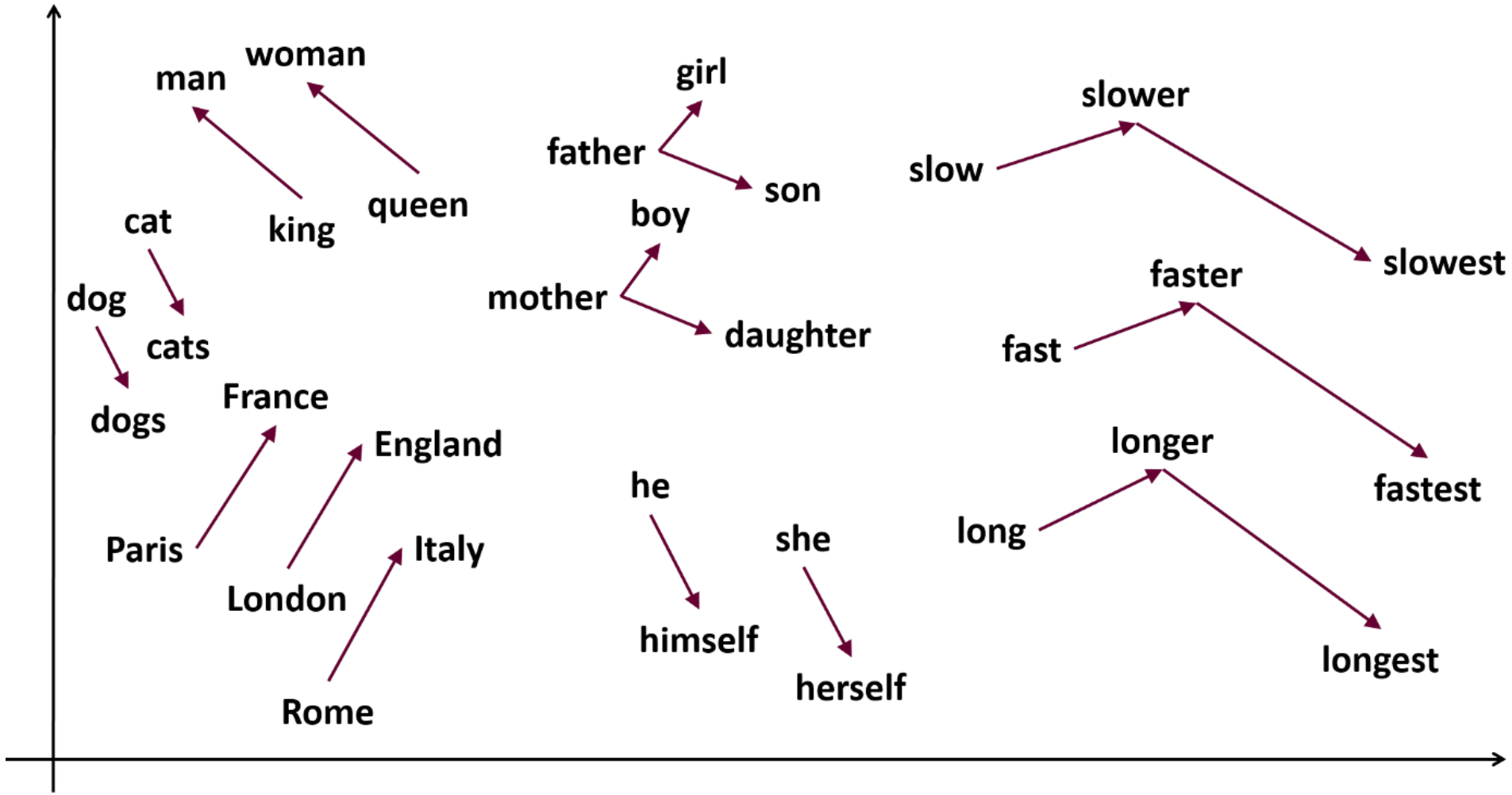

CBOW

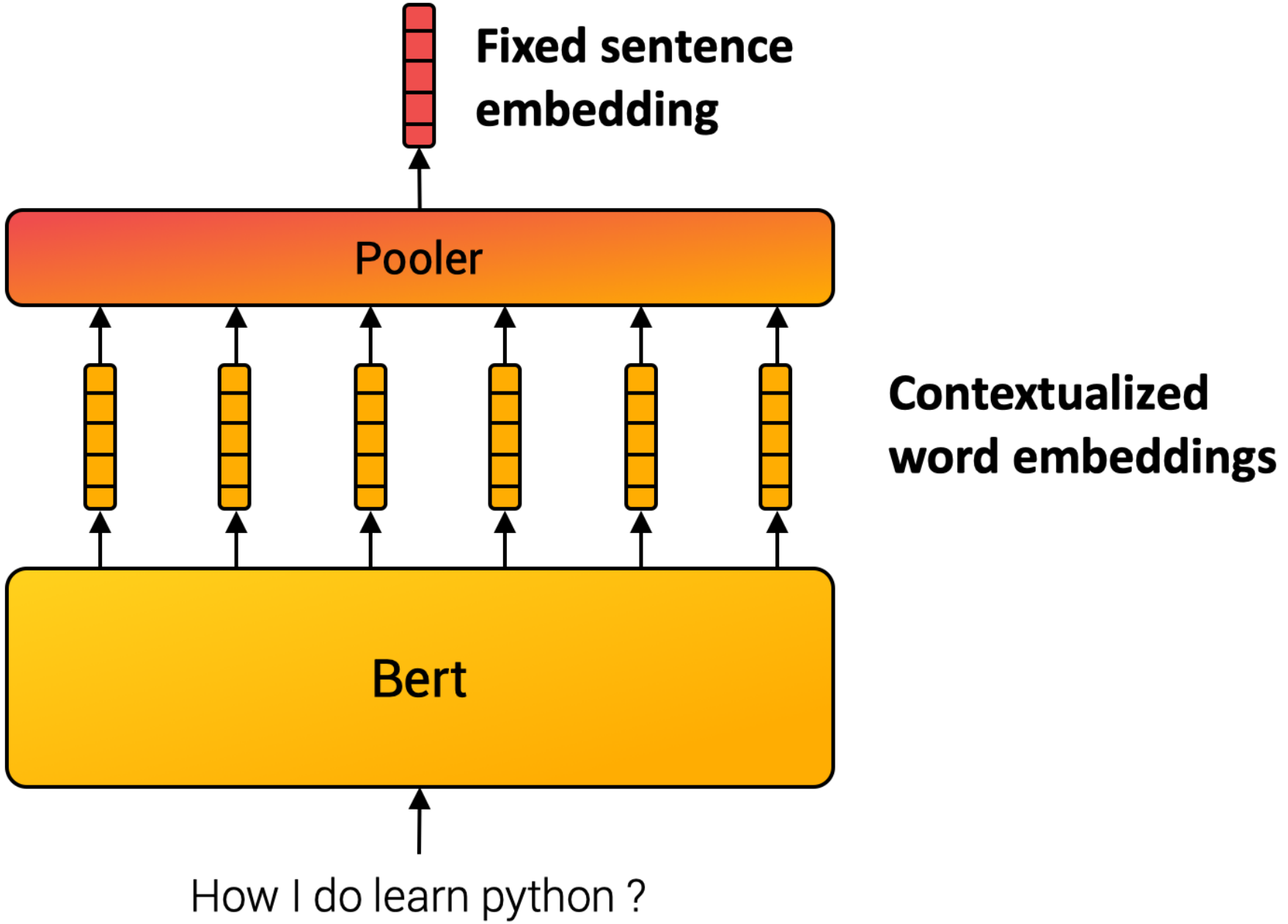



Skip-gram

# Word Embedding

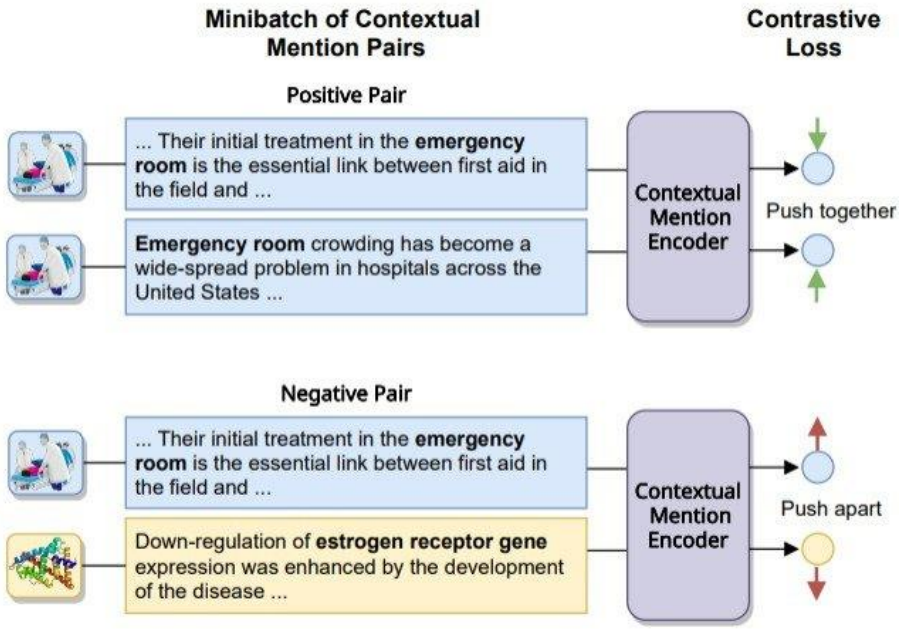
## Vector Arithmetic



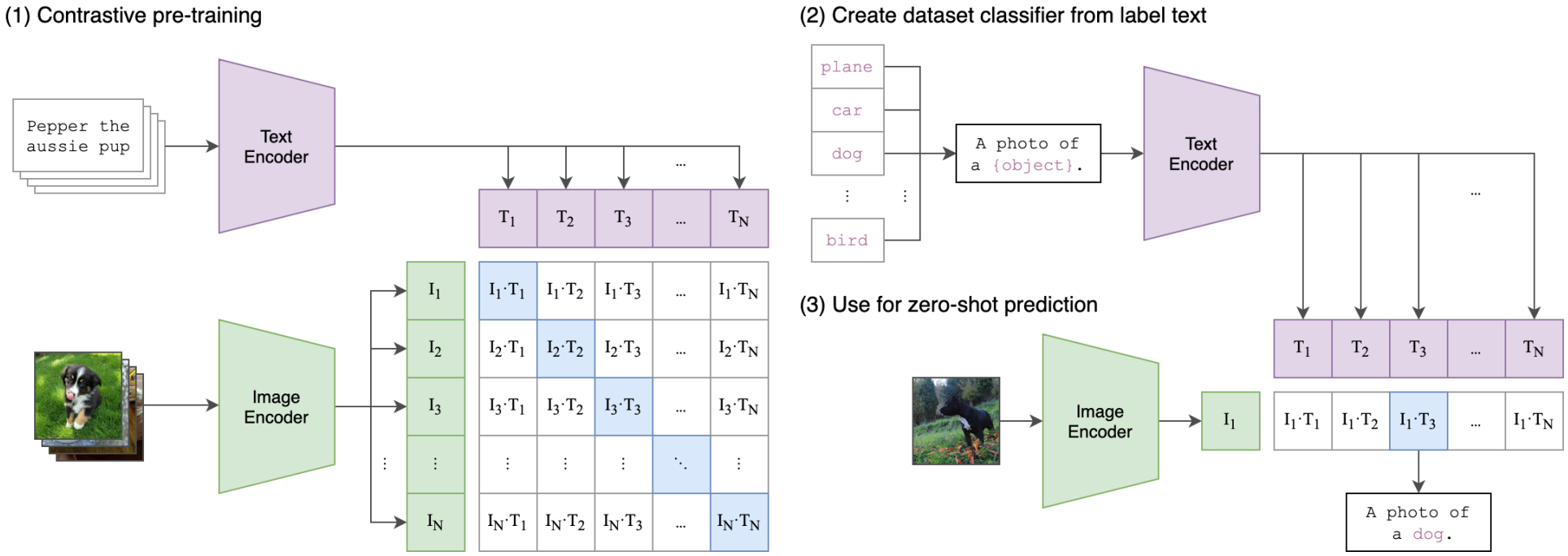

Fixed sentence embedding
Pooler
Contextualized word embeddings
Bert
How I do learn python ?

# Contrastive Learning

## Training vector representations



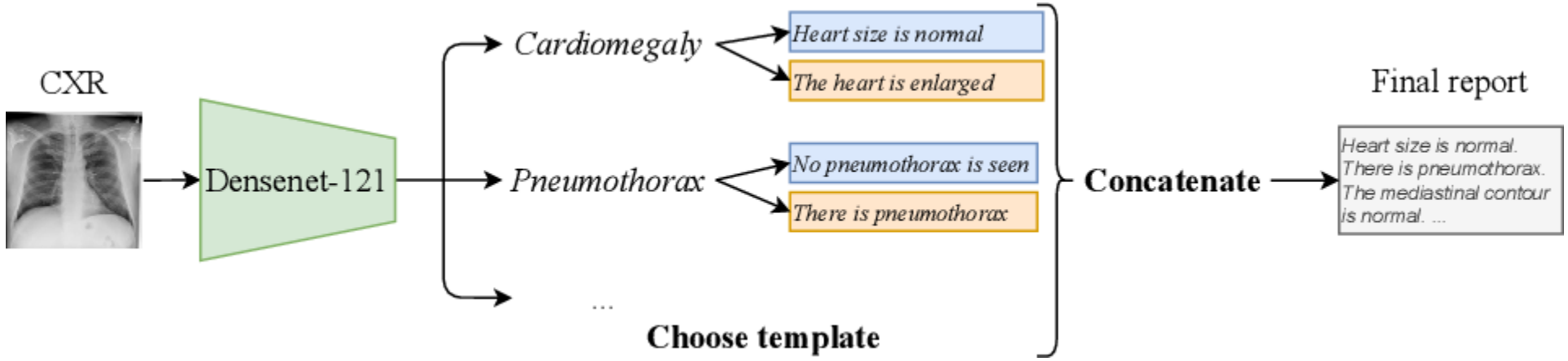
# Contrastive Learning

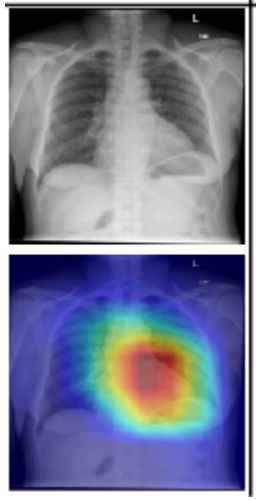



[CLIP: Connecting Text and Images (openai.com)]

# Automatic Captioning

## Convert a sentence vector representation back to text



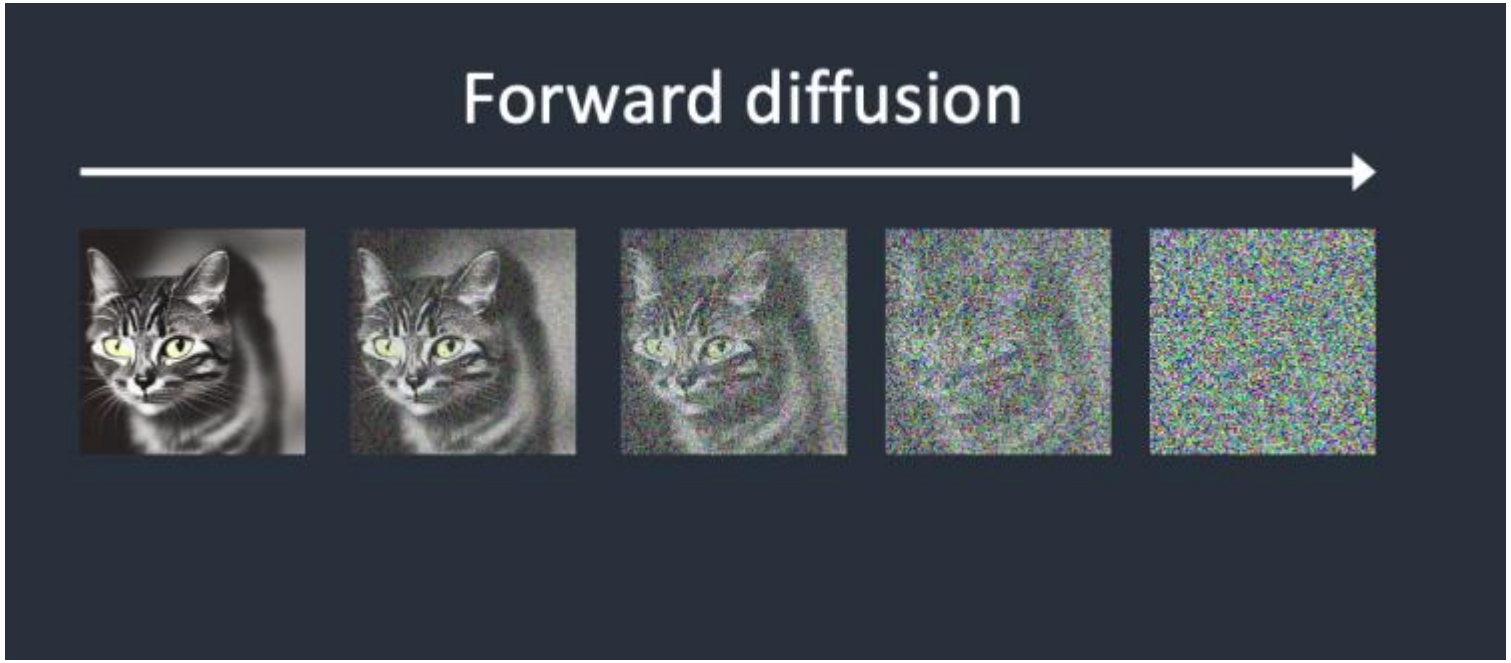

| CXR | Generated by CNN-TRG | Ground Truth |
|---|---|---|
|  | The heart is enlarged. The mediastinal contour is normal. No focal consolidation. The lungs are free of focal airspace disease. No atelectasis. No pleural effusion. No fibrosis. No pneumonia. No pneumothorax is seen. No pulmonary edema. No pulmonary nodules or mass lesions identified. No fracture is seen. | The heart is mildly enlarged. Left hemidiaphragm is elevated. There is no acute infiltrate or pleural effusion. The mediastinum is unremarkable. |

# Stable Diffusion

From text to image

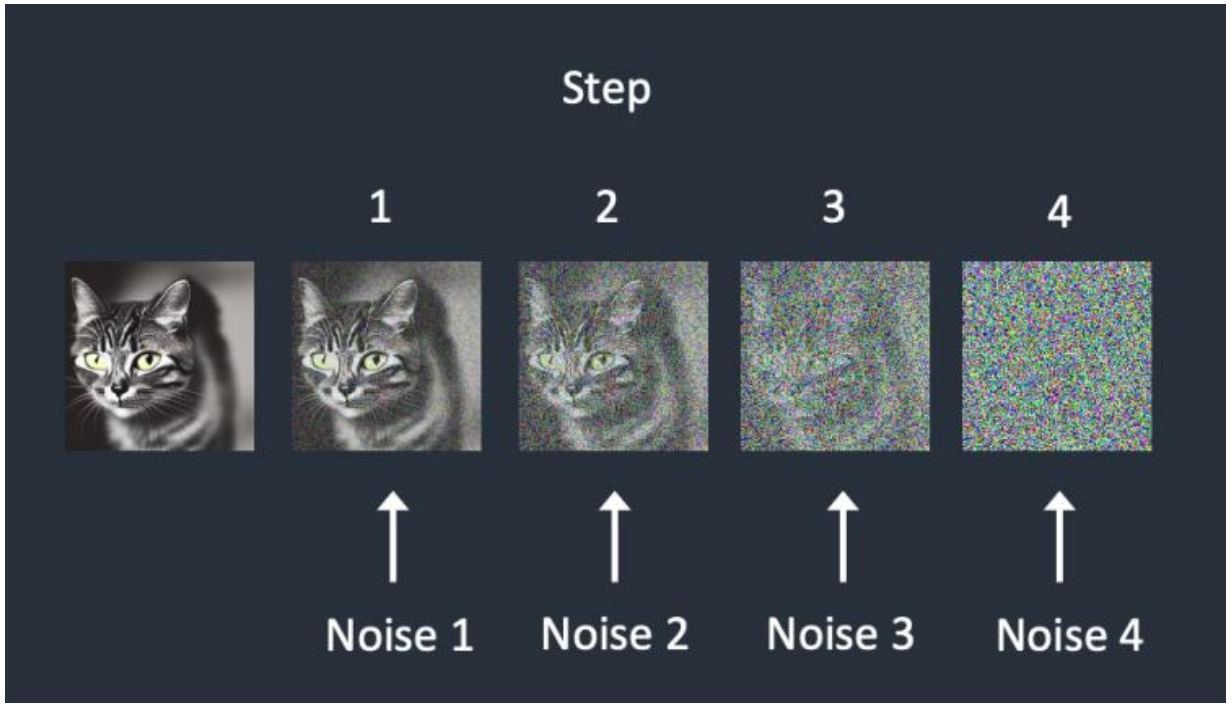
# Diffusion

Adding noise to an image until we can’t tell what it used to be



# Reverse Diffusion

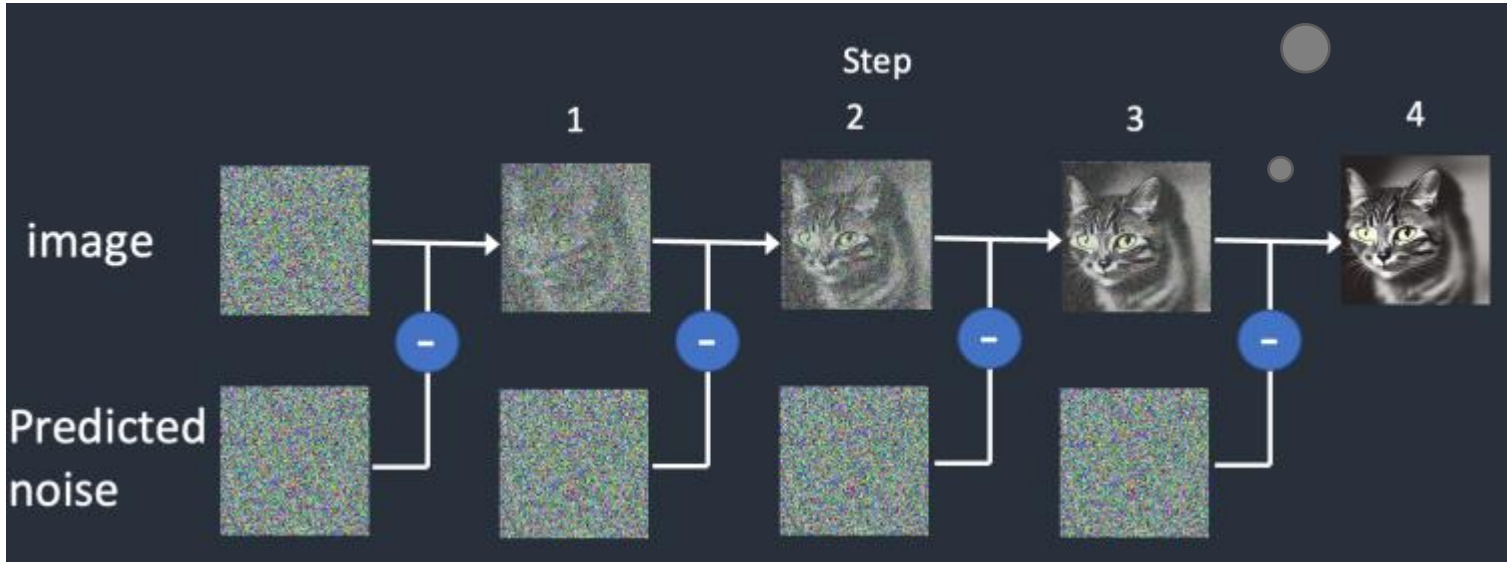
Predict the ***amount*** of added noise – using U-NET



But: this is expensive to perform on a whole image… We need to compress it!

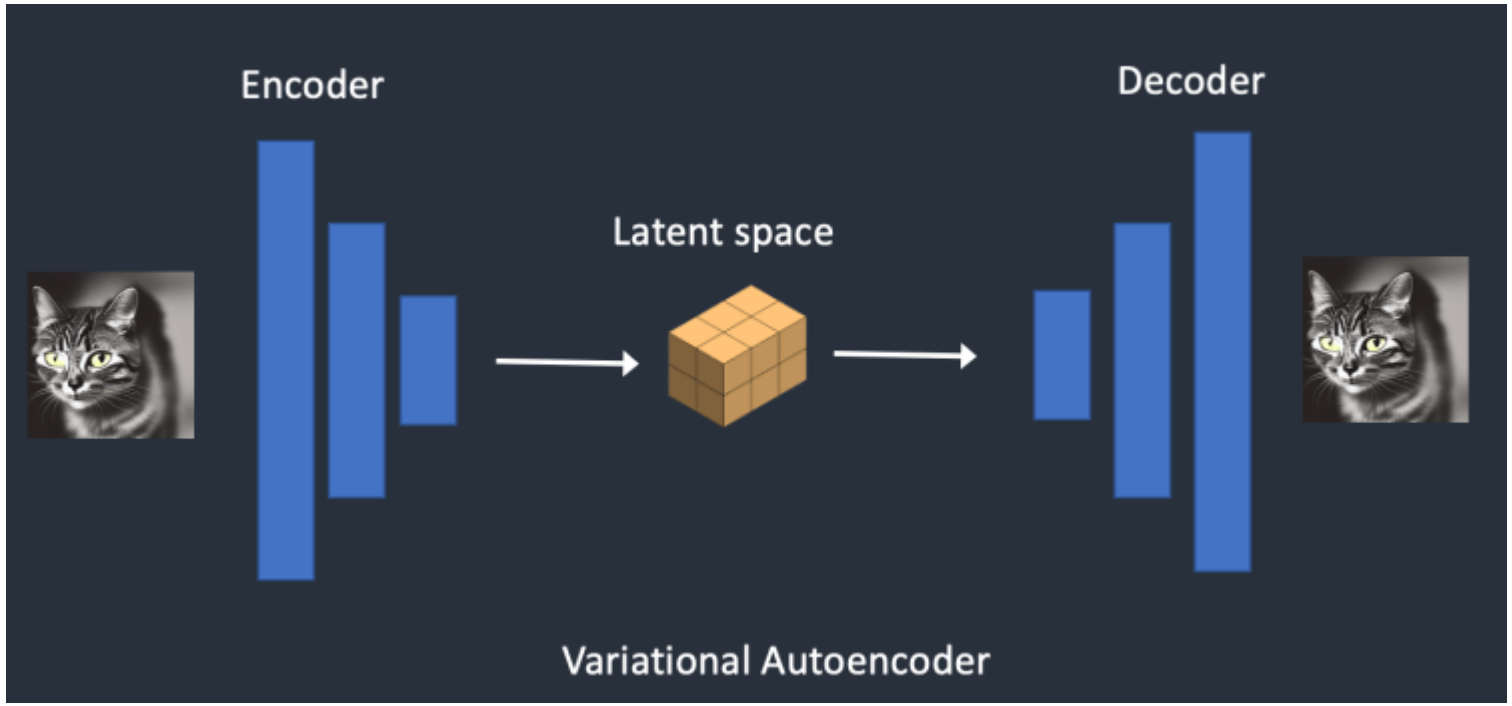
# Reverse Diffusion

Predict the ***amount*** of added noise – using U-NET – so that we can subtract it from a noisy image.



# Compressing the Image

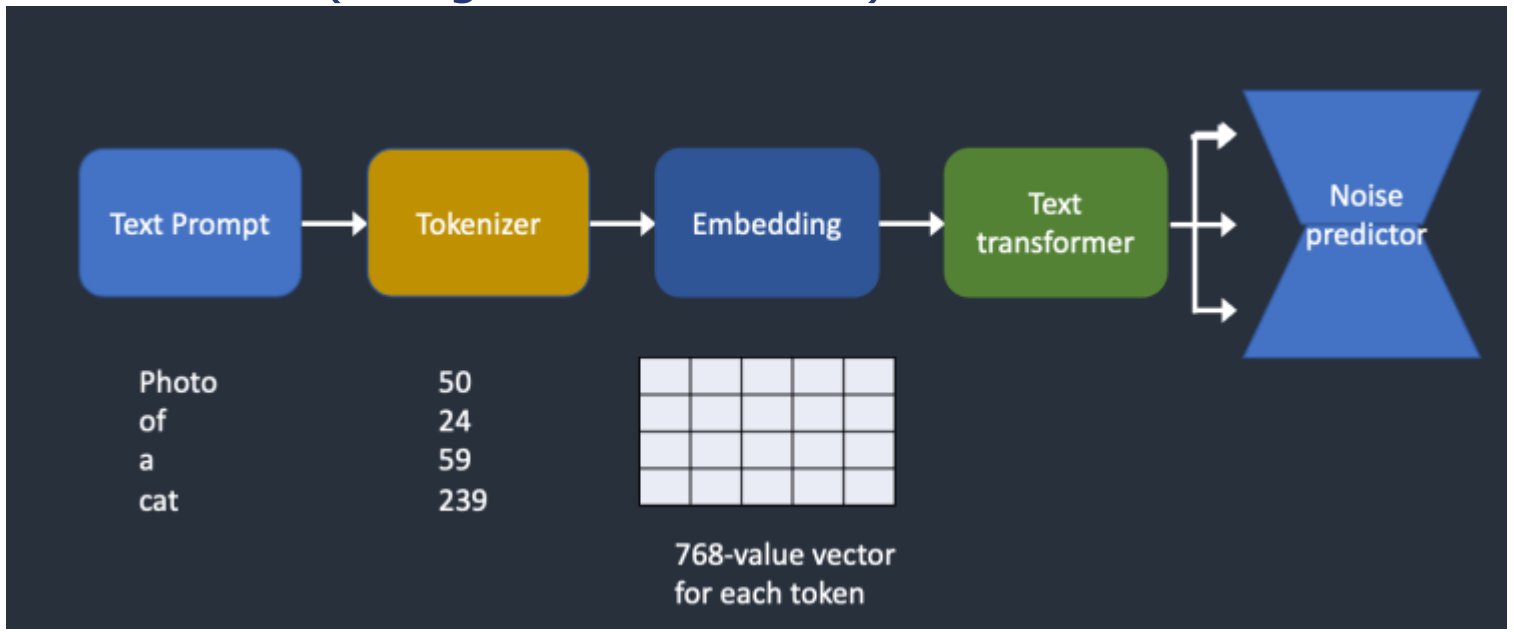
Into a latent representation:



# Conditioning using Text

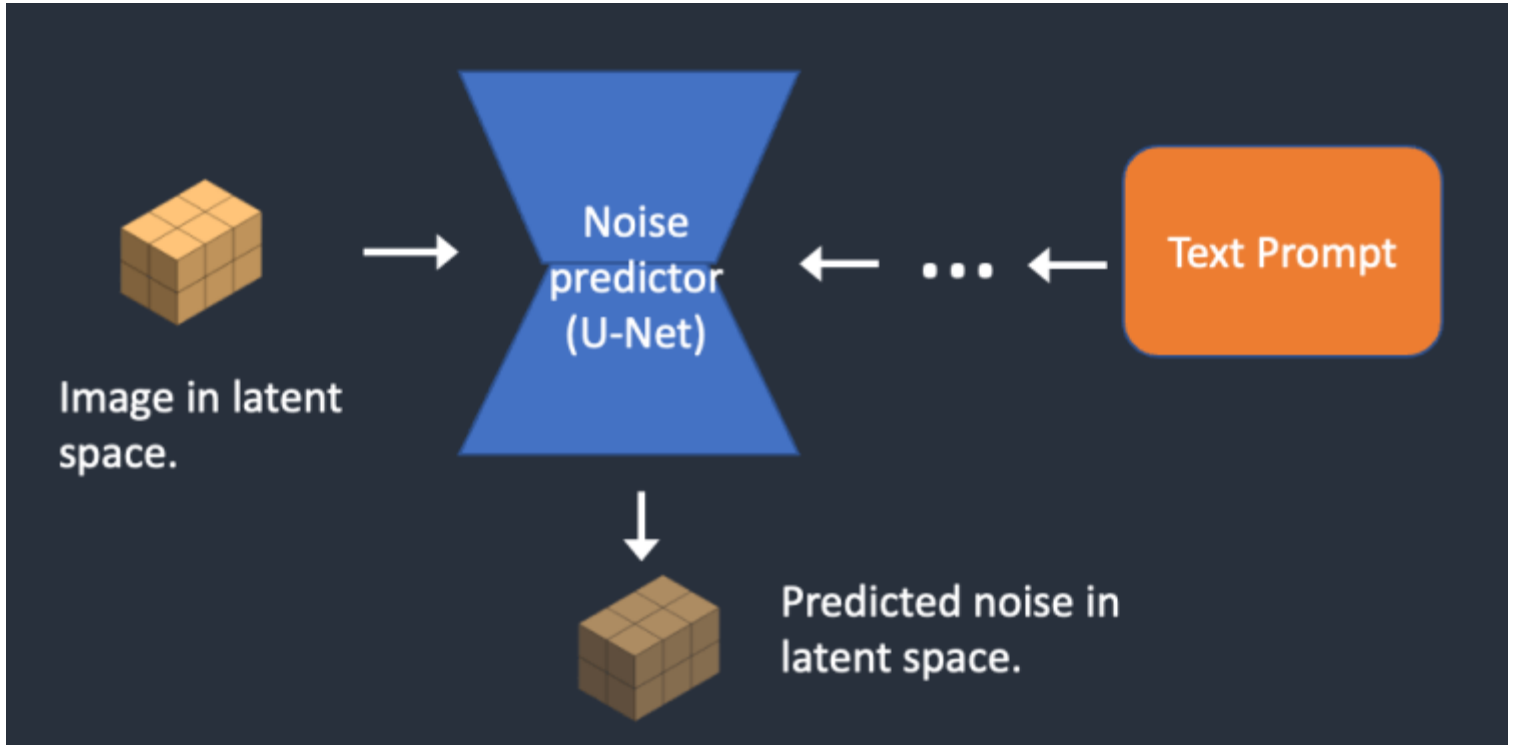
Convert the text to a vector (e.g., using BERT-like model)

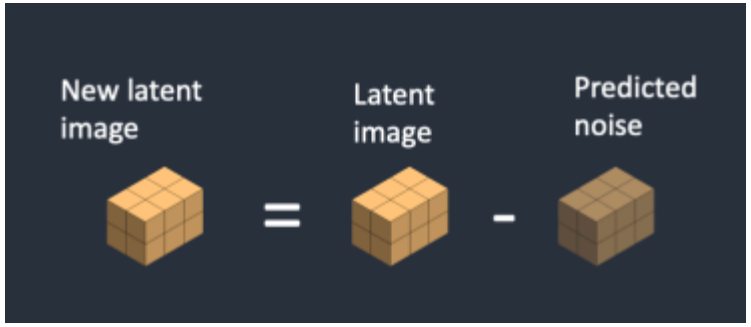
- Tokenizer – split the sentence into words (using CLIP tokenizer)
- Embedding – convert each word into a vector
- Transformer – "understand" the sentence meaning and convert it into a single vector.
- Noise-Predictor – U-NET

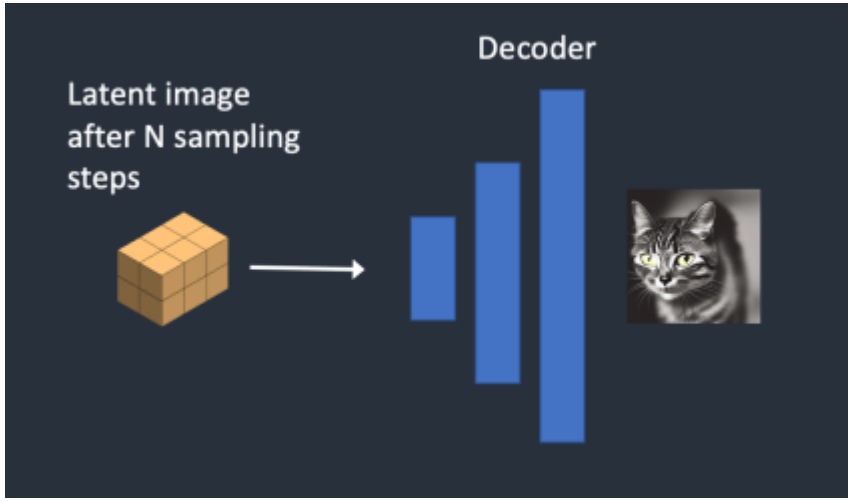


# Stable Diffusion

## The process

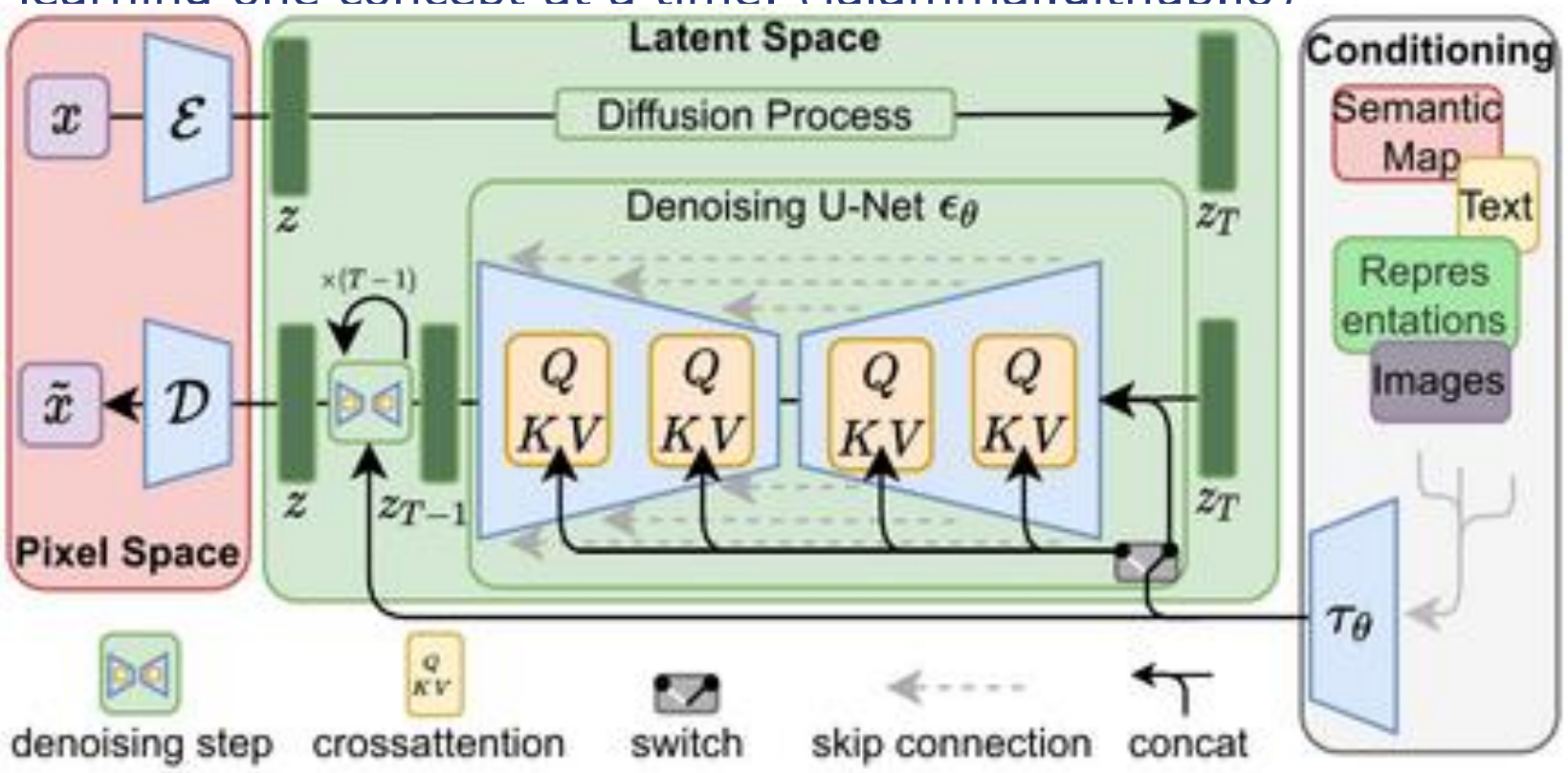


# Stable Diffusion – Recommended Read

The Illustrated Stable Diffusion – Jay Alammar – Visualizing machine learning one concept at a time. (jalammar.github.io)




# DALL-E

TEXT PROMPT an illustration of a baby daikon radish in a tutu walking a dog

AI-GENERATED IMAGES

Edit prompt or view more images↓

TEXT PROMPT an armchair in the shape of an avocado. . . .

AI-GENERATED IMAGES

Edit prompt or view more images↓

DALL·E: Creating Images from Text (openai.com)

# DALL-E (ver. 2)

ORIGINAL IMAGE

→

DALL·E 2 VARIATIONS

[DALL·E 2 (openai.com)]

# DALL-E (ver. 2)

DALL·E 2 (openai.com)

# DALLE-2

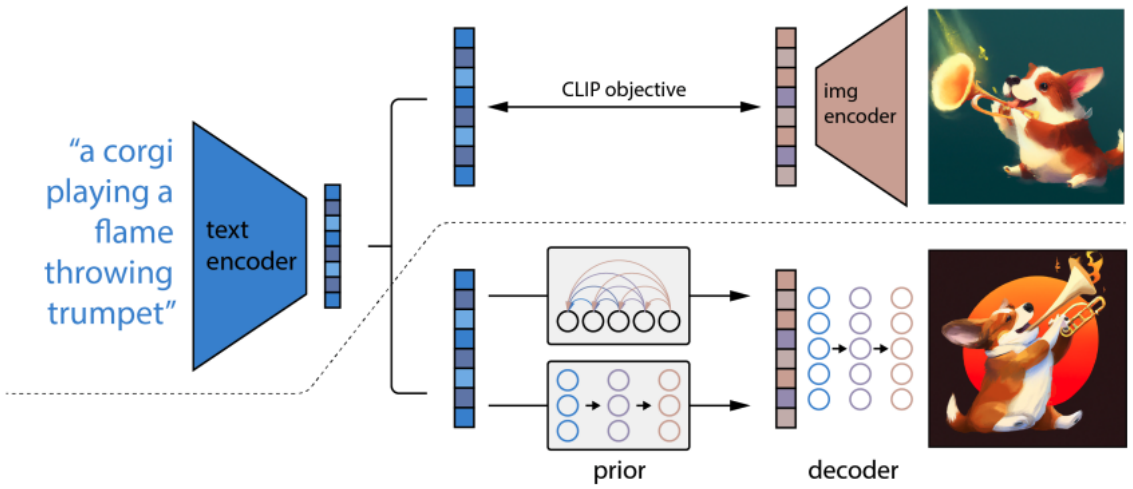



Figure 2: A high-level overview of unCLIP. Above the dotted line, we depict the CLIP training process, through which we learn a joint representation space for text and images. Below the dotted line, we depict our text-to-image generation process: a CLIP text embedding is first fed to an autoregressive or diffusion prior to produce an image embedding, and then this embedding is used to condition a diffusion decoder which produces a final image. Note that the CLIP model is frozen during training of the prior and decoder.

**Hierarchical Text-Conditional Image Generation with CLIP Latents**

**Aditya Ramesh*** OpenAI aramesh@openai.com
**Prafulla Dhariwal*** OpenAI prafulla@openai.com
**Alex Nichol*** OpenAI alex@openai.com
**Casey Chu*** OpenAI casey@openai.com
**Mark Chen** OpenAI mark@openai.com

imc FH KREMS
UNIVERSITY OF APPLIED SCIENCES

# Stable Diffusion for Medical Usage

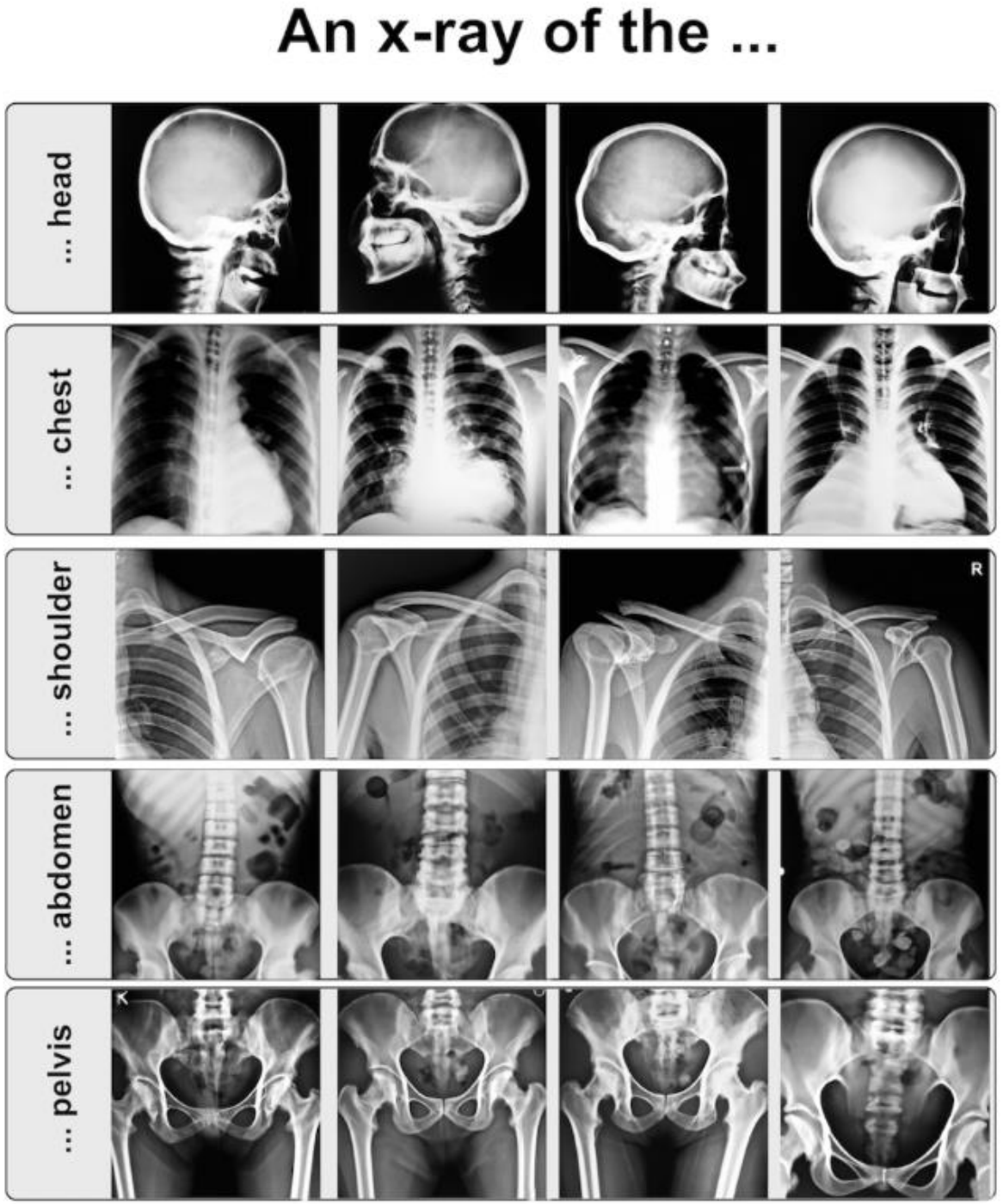
"pathology generation or CT, MRI, and ultrasound images are still limited"

WHAT DOES DALL-E 2 KNOW ABOUT RADIOLOGY?

A PREPRINT


**Lisa C. Adams***
Department of Radiology
Stanford University School of Medicine
Stanford, CA, USA
lcadams@stanford.edu

**Felix Busch***
Department of Radiology
Charité – Universitätsmedizin Berlin
Berlin, Germany
felix.busch@charite.de

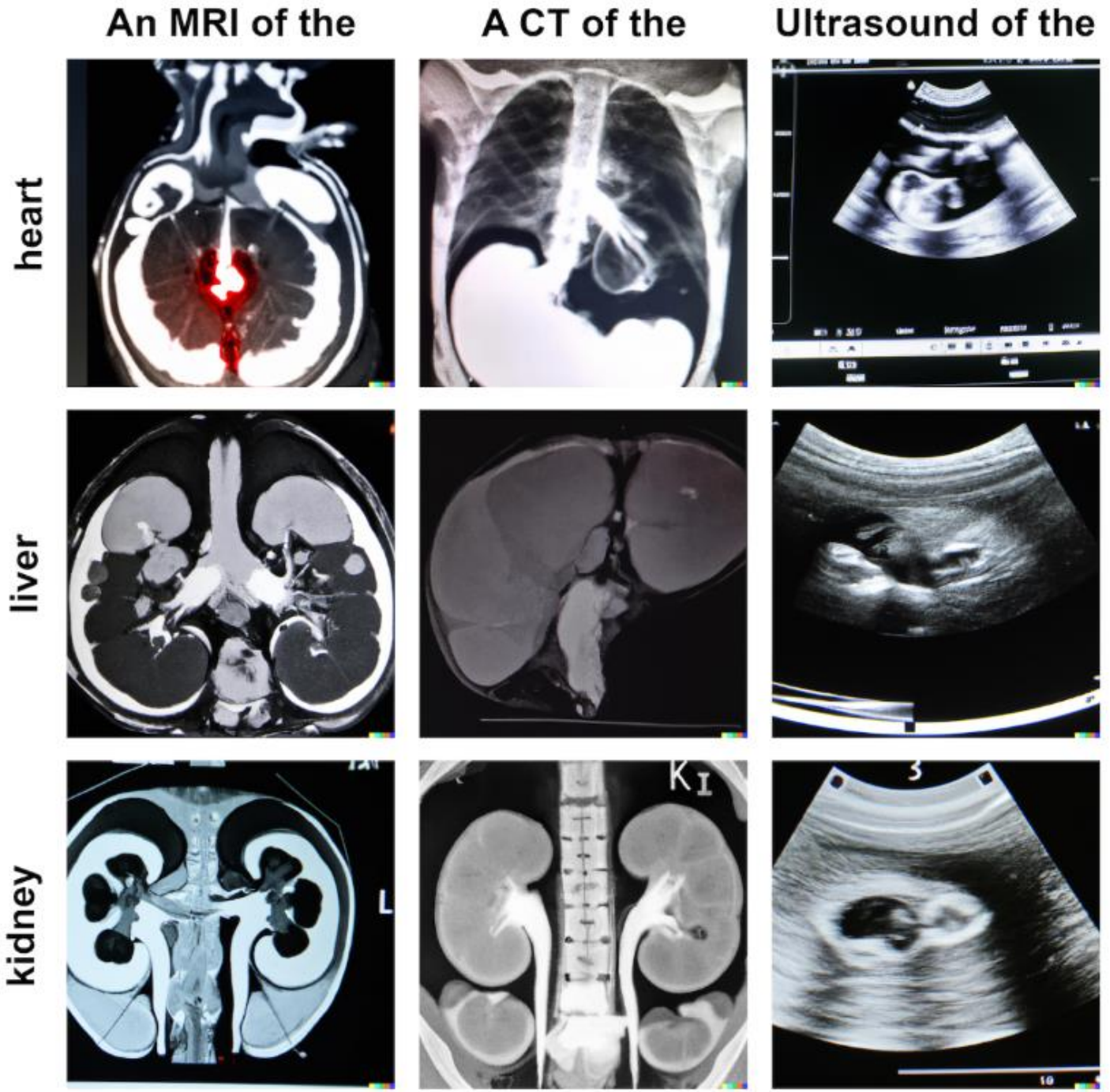
imc FH KREMS
UNIVERSITY OF APPLIED SCIENCES

# What Does DALL-E 2 Know About Radiology?

A Preprint


**Lisa C. Adams***
Department of Radiology
Stanford University School of Medicine
Stanford, CA, USA
lcadams@stanford.edu

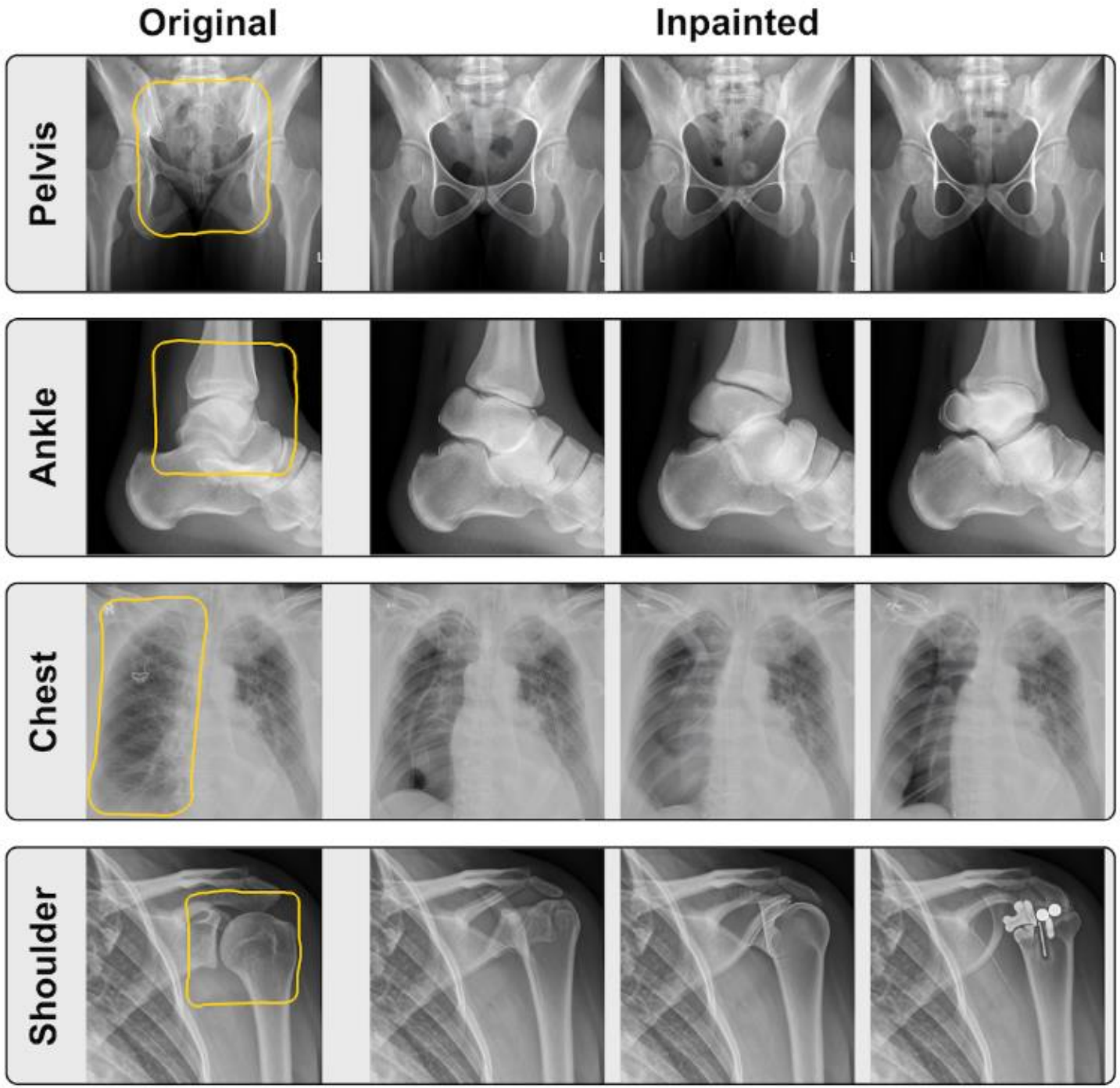
**Felix Busch***
Department of Radiology
Charité – Universitätsmedizin Berlin
Berlin, Germany
felix.busch@charite.de

imc FH KREMS
UNIVERSITY OF APPLIED SCIENCES

# Dalle-2 - Inpainting

Image-completion:
original vs Dalle-2

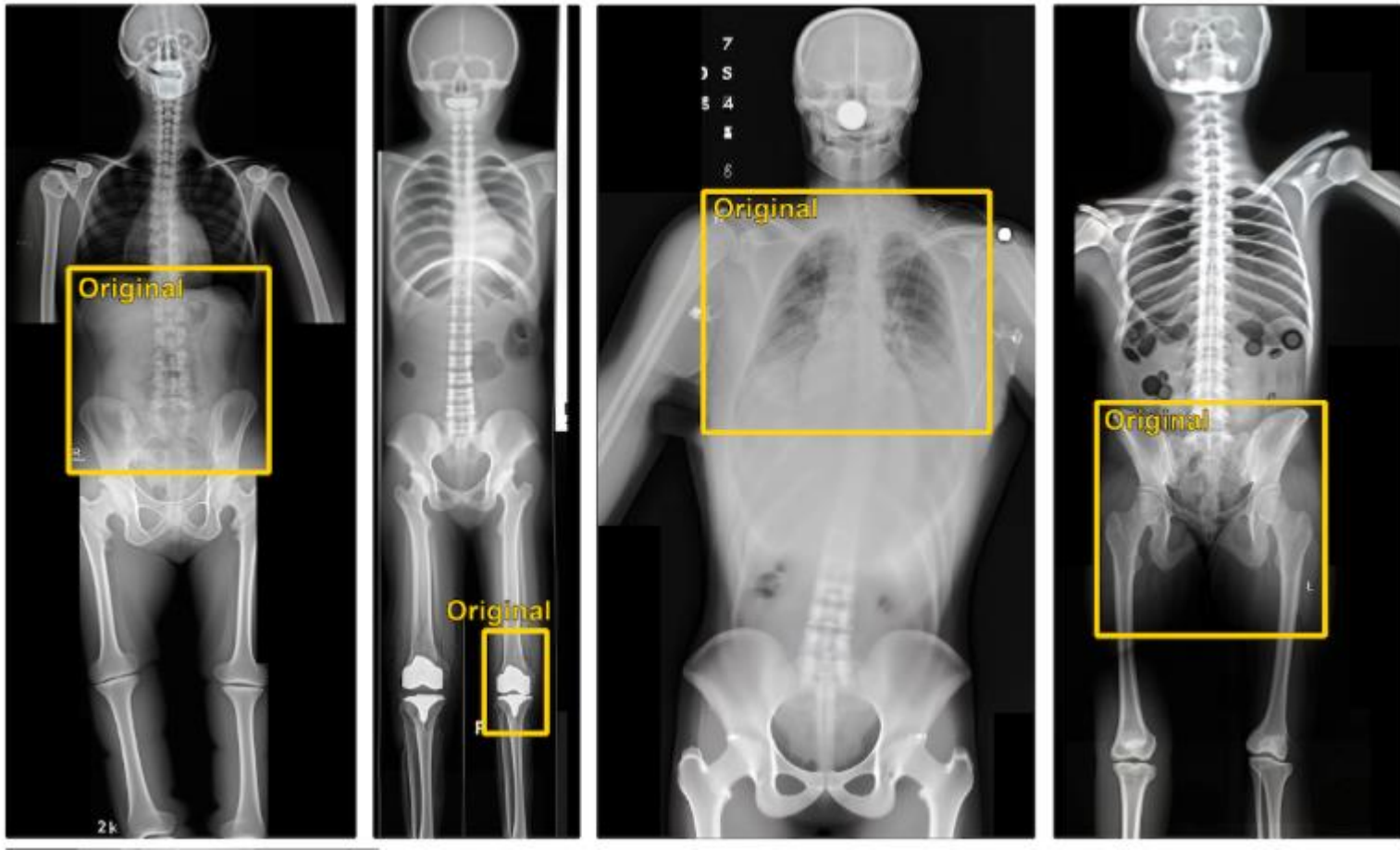


# Dalle-2

Image completion.

Yellow – the original image.

The rest is done by Dalle-2



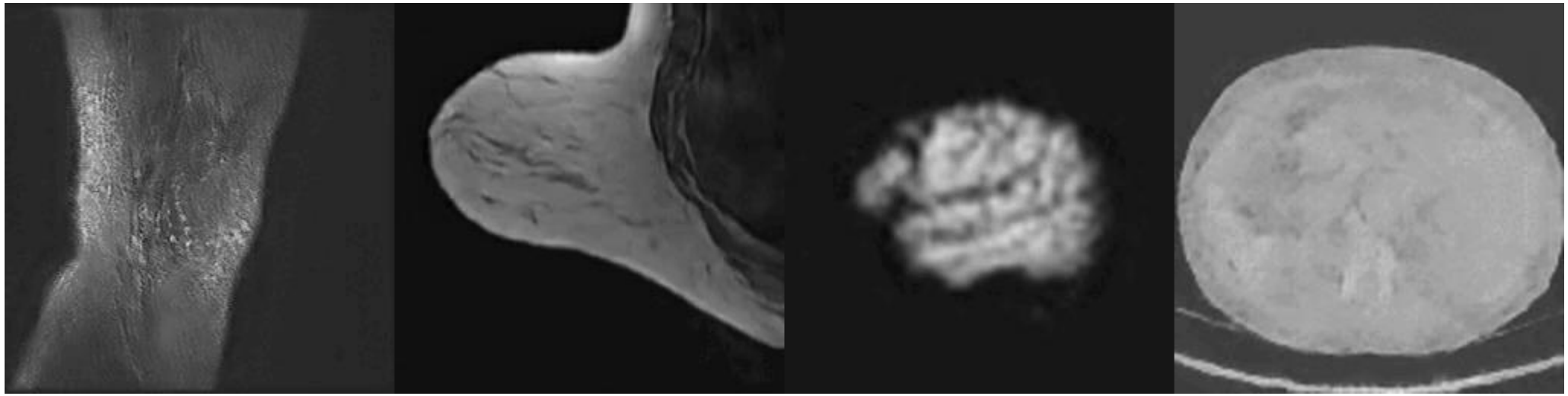
imc FH KREMS
UNIVERSITY OF APPLIED SCIENCES

Medical Diffusion: Denoising Diffusion Probabilistic Models for 3D Medical Image Generation

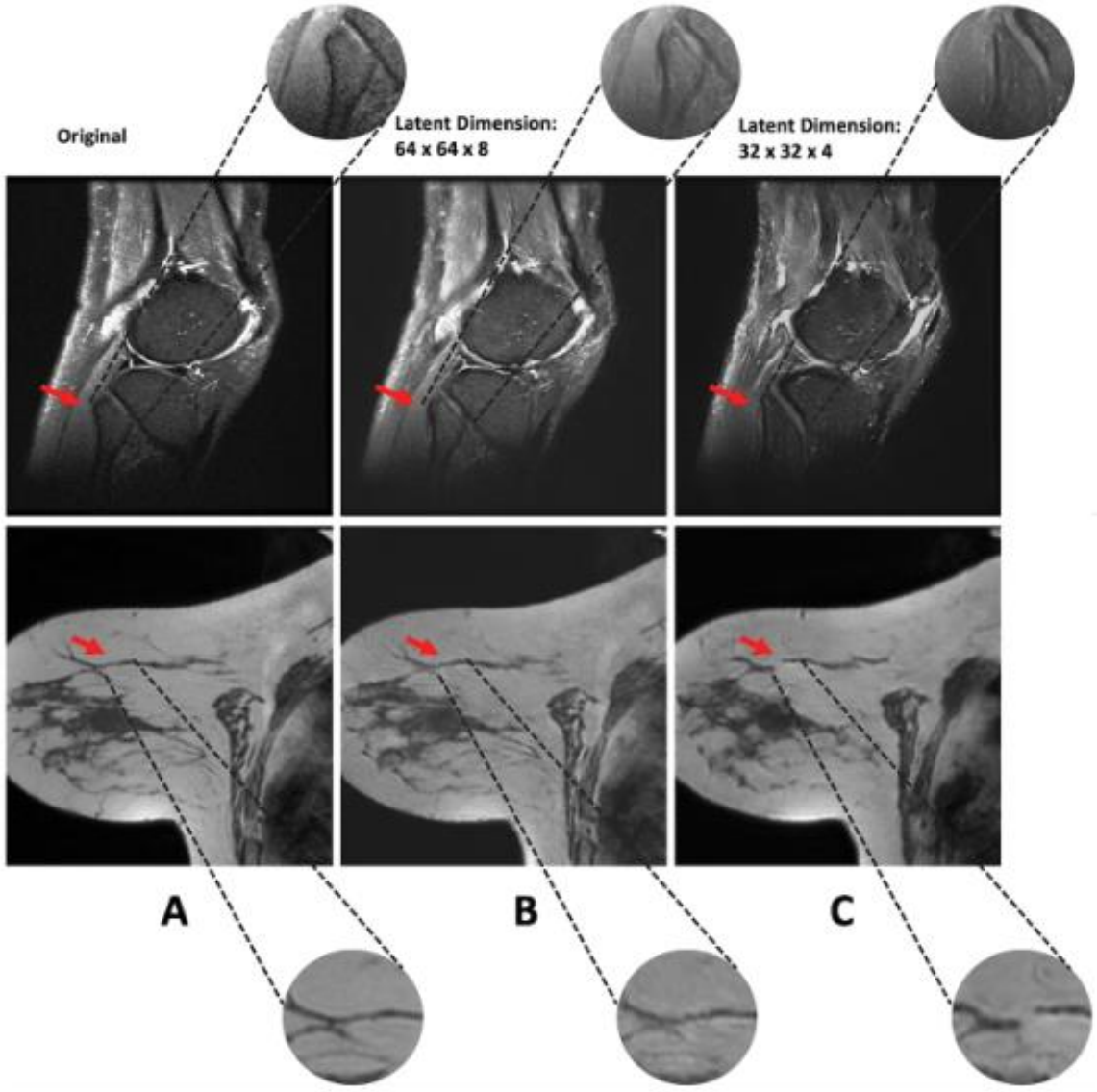
Firas Khader[1], Gustav Müller-Franzes[1], Soroosh Tayebi Arasteh[1], Tianyu Han[2], Christoph Haarburger[3], Maximilian Schulze-Hagen[1], Philipp Schad[1], Sandy Engelhardt[4], Bettina Baeßler[5], Sebastian Foersch[6], Johannes Stegmaier[7], Christiane Kuhl[1], Sven Nebelung[1], Jakob Nikolas Kather*[8,9,10,11], and Daniel Truhn*[1]

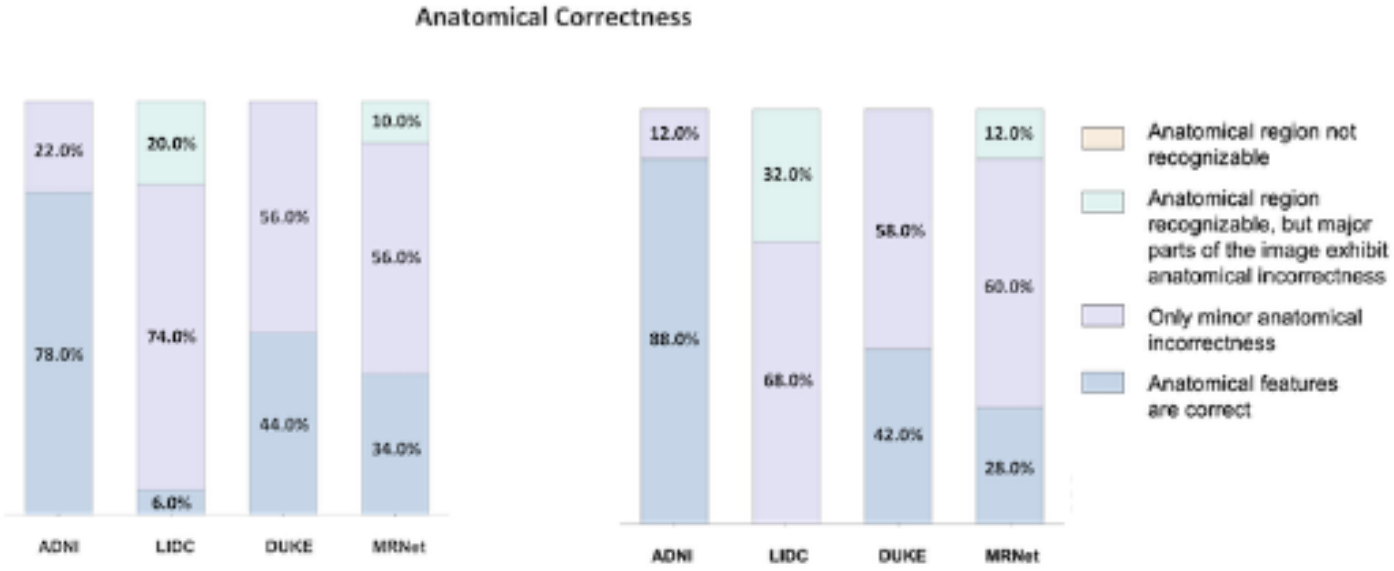
# Generated Images as Training Data

When data is scarce…



FirasGit/medicaldiffusion: Medical Diffusion: This repository contains the code to our paper Medical Diffusion: Denoising Diffusion Probabilistic Models for 3D Medical Image Synthesis (github.com)



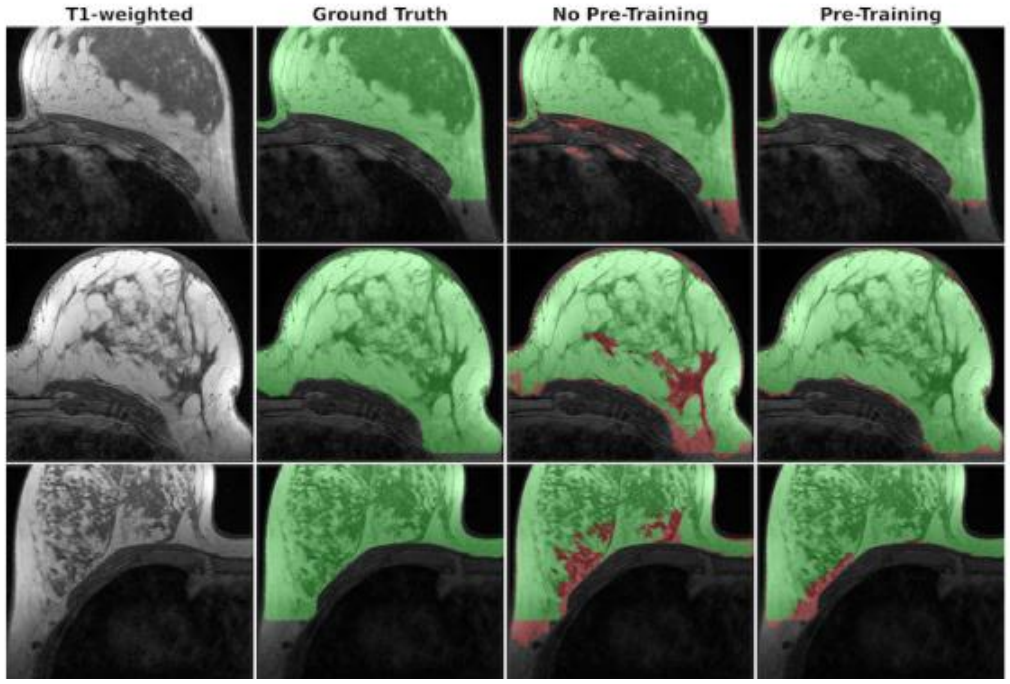




**Fig. 3.** Comparison of the reconstruction quality of the VQ-GAN autoencoder when using different compression factors for two different samples. A latent dimension of 64 x 64 x 8 (i.e., a compression factor of 4 in each dimension) allows for a detailed reconstruction of the original image that conserves anatomical consistency. A compression factor of 8 (i.e., a latent dimension of 32 x 32 x4) distorts the fibular bone in knee MRI and the fibroglandular tissue continuity in breast MRI.

# CNN - APPLICATIONS

# CNN – RL Example

Mass detection and classification in mammography

- Research directions may involve using **different** CNNs:
  (1) an **anomaly detector,** (2) a **mass localizer,** and a (3) **mass classifier**.

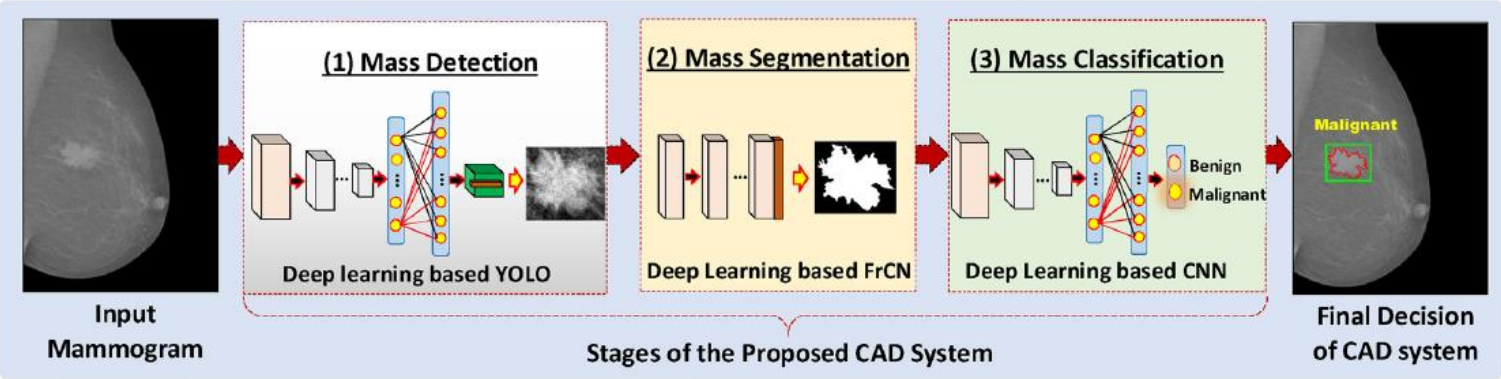



Fig. 1. Schematic diagram of the proposed computer-aided diagnosis (CAD) system based on deep learning to detect, segment, and classify breast cancer masses from input digital X-ray mammograms.

# CNN – RL Example #2

## Mass Detection and Classification

- A different approach for the same task is taking advantage of powerful architectures such as **YOLO**, **U-Net,** and **Faster R-CNN** to simultaneously **segment** and **classify** masses in an end2end manner.

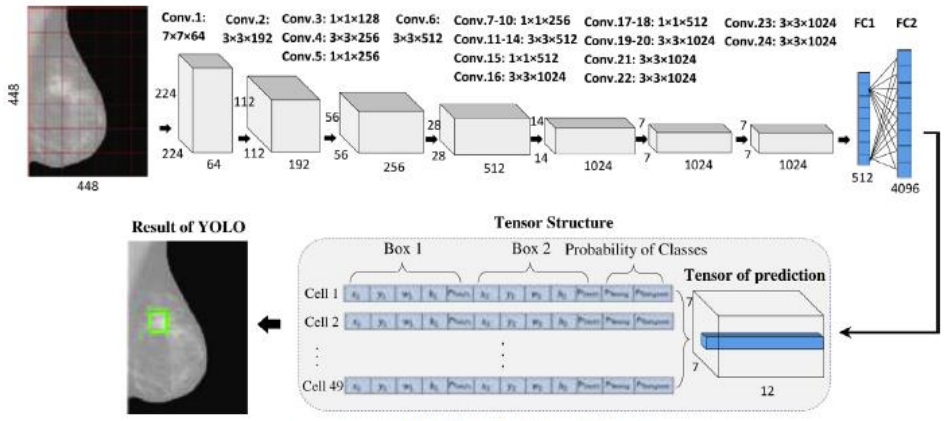



Fig. 2. The structures of proposed YOLO-based CAD system.

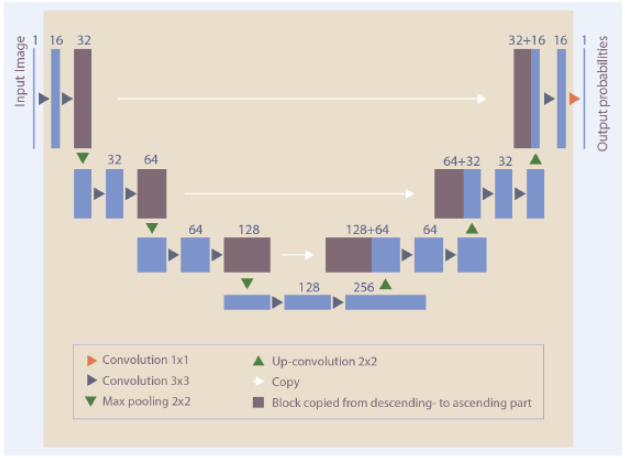



Figure 2. The u-net architecture, where we have doubled the number of filters in the each block and applied batch normalization. Each convolution $3 \times 3$ was followed by a rectified linear unit activation layer, while the final convolution $1 \times 1$ was followed by a pixel-wise sigmoid activation layer.

# DREAM Challenge

- **Faster R-CNN** achieved **2nd** place in the Digital Mammography **DREAM Challenge** (2016 – 2017)

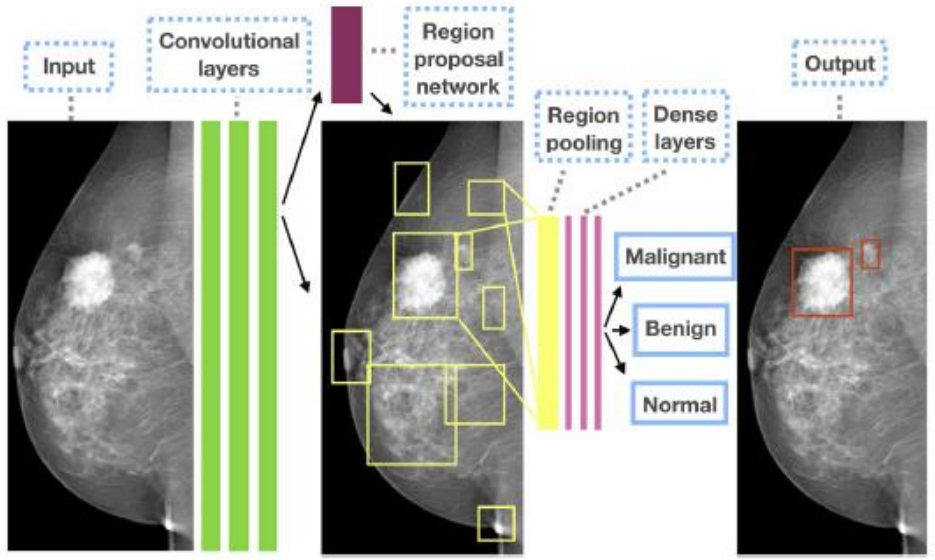



**Figure 1.** The outline of the Faster R-CNN model for CAD in mammography.

# Image Classification with CNN

unbalanced Data

- **RetinaNet** has been used to address the issue of **unbalanced data** (malignant vs benign)
- It uses a **focal loss** function instead of **cross-entropy**.

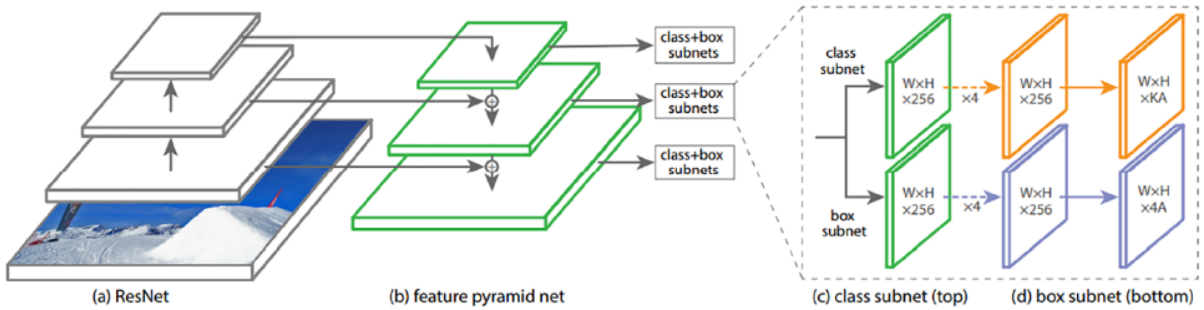



**Fig 1. The network architecture of RetinaNet.** RetinaNet uses the Feature Pyramid Network (FPN) [34] on top of the convolutional neural network ResNet [35] as a backbone network (a) to generate a rich convolutional feature pyramid (b). The class subnet (c) is for classifying anchor boxes, and the box subnet (d) is for regressing from anchor boxes to ground-truth object boxes. (Lin, Tsung-Yi, et al., 2017 [32].)

# Additional Methods Used in Research

- **Attention-guided** models (e.g., AUNet)
- **Transfer learning** (e.g., AlexNet, VGG16, ResNet50, Inception-V3)
  - In some cases, **pre-training** was performed using one mammogram database (e.g., INBreast) and **fine-tuning** was performed using a **different** mammogram database (e.g., OPTIMAM).
- Incorporating **context** around mass regions has shown to **improve** classification accuracy.
- **Unsupervised learning** based on auto-encoders for **feature extraction** and/or mass **segmentation** are also common.

# Recommended Watch

- DeepMind@UCL lecture series on YouTube

# References – Deep Learning

- **2019** - Deep convolutional neural networks for mammography: advances, challenges and applications (nice review)
- **2020** - Survey of deep learning in breast cancer image analysis (focuses mostly on **feature extraction**; also discusses **histopathology**)
- **2021** - Convolutional neural networks for breast cancer detection in mammography: A survey (nice review – has list of **FDA approved** CAD systems)
- **2021** - Breast Cancer Segmentation Methods: Current Status and Future Potentials (nice review of past/current methods; has some **language issues**)

# References – Deep Learning

Good intro papers of AI/DL in radiology ("AI in Medicine" section)

- **2017** - Deep Learning A Primer for Radiologists (nice **intro to CNNs** in radiology)
- **2020** - Artificial Intelligence A Primer for Breast Imaging Radiologists (AI+radiology, risk prediction, tomosynthesis)