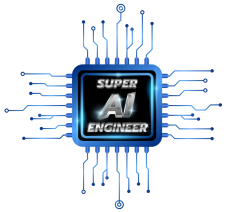


Super AI Engineer Development Program SS4

Modern Image Search

AI Lecture

January 30, 2024

Romrawin Chumpu - Jinpu - จิ้นปู้
Super AI Engineer SS1

Table of Content

Our Journey for Today

First Half (1 hour)

1. How Does Computer Understand Images: A Review
2. What is **Image Search**?
3. Types of **Image Search** in Modern Machine Learning
4. **Hands-On #1** Image Search Image (CBIR)

Second Half (1 hour)

5. State-Of-The-Art Deep Learning with Image Data
6. **Hands-On #2** Image-Text Search (CLIP)
7. Multimodal Models: Image-Text Focus
8. **Hands-On #3** Image-Text Segmentation Search (SAM & CLIP)

1 How Does Computer Understand Images: A Review

Digital Images

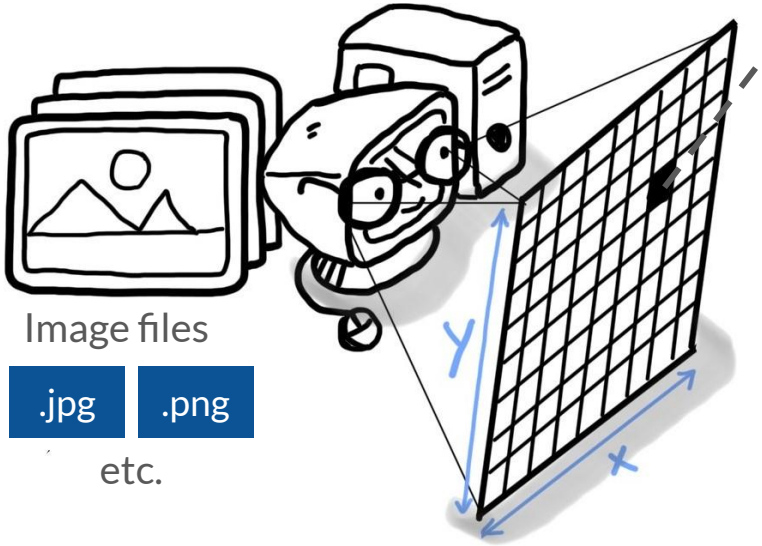


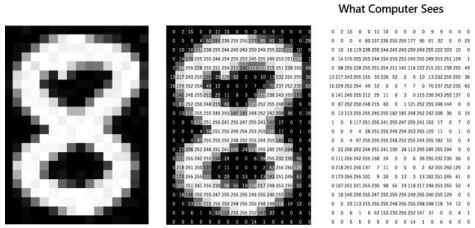
Image files

.jpg .png

etc.

Pixel

The smallest element of an image

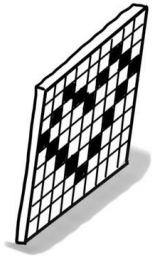


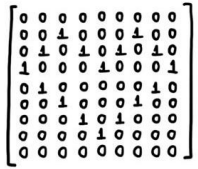
Computer sees an image as **a grid of pixels** or **a 2D matrix of pixels**

Pixel(x, y) → **Value**

Image types and resolution

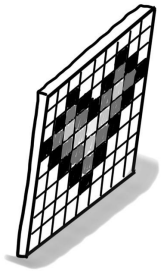
Pixel(x, y) → **Value**

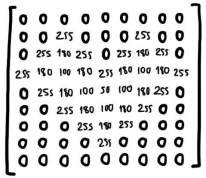




Binary Image

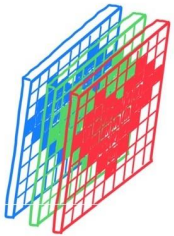
1 | 0

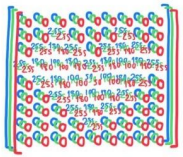




Gray Image

0-255





Color Image

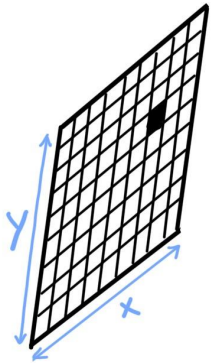
0-255 | 0-255 | 0-255

Shape

(x, y, 1)

(x, y, 3)

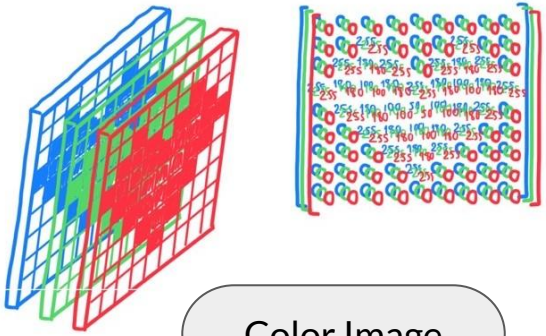
(x, y, c)

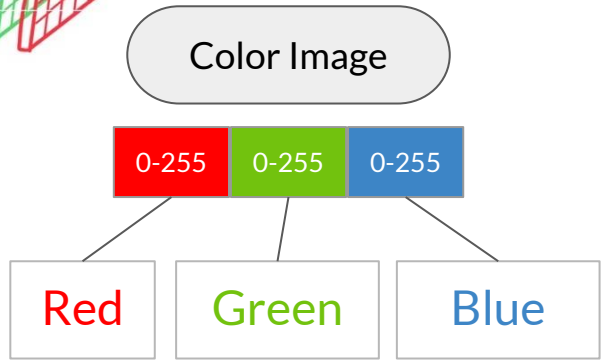


Resolution

	x	y
1080p	1920	1080
4K	3840	2160
8K	7680	4320

Image bit depth





Why 0-255?

Bit depth

1 bit (2^1) = 2 tones
2 bits (2^2) = 4 tones
3 bits (2^3)= 8 tones
4 bits (2^4) = 16 tones
8 bits (2^8) = 256 tones
16 bits (2^{16}) = 65,536 tones
24 bits (2^{24}) = 16.7 million tones

8 Bit Quantization

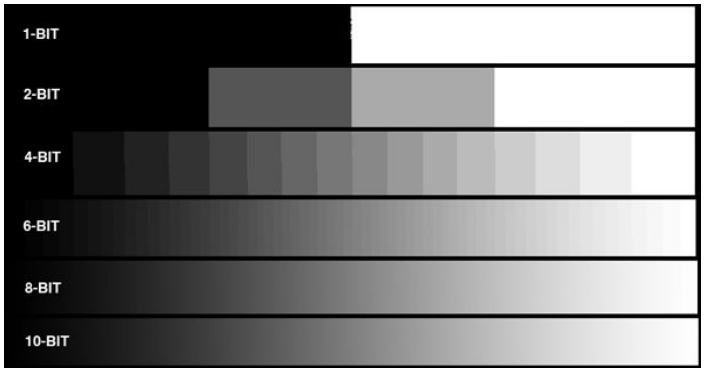


Image Processing and Computer Vision

Image Processing

enhancing and transforming images

Transforming Pixels into Perfection

- Image Enhancement
- Filtering
- Transformation

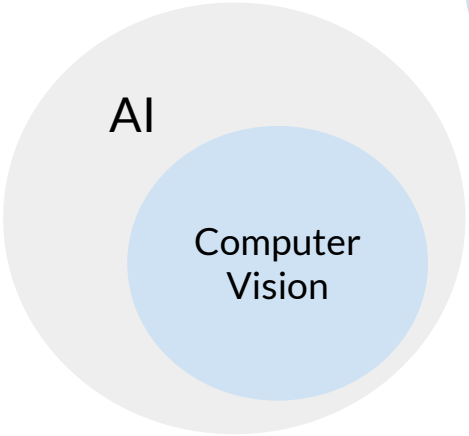
Computer Vision

extract meaning from images

Decoding the Visual World

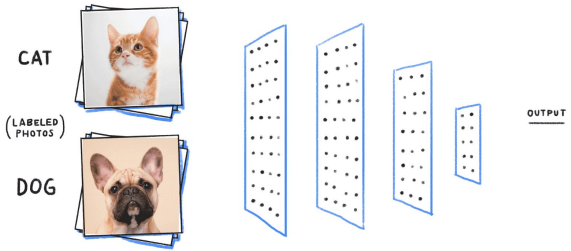
- Pattern Recognition
- Object Detection
- Deep Learning

Computer Vision Tasks



Computer vision advancements help us ground up image understanding.

Image Classification



Object Detection

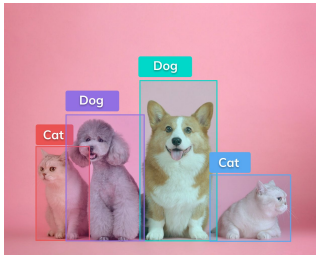


Image Segmentation

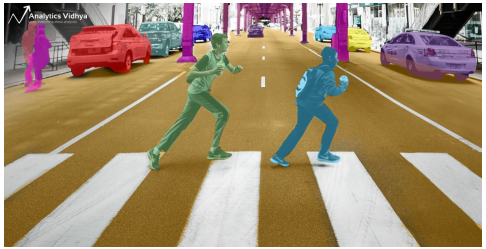


Image Captioning

"man in black shirt is playing guitar."

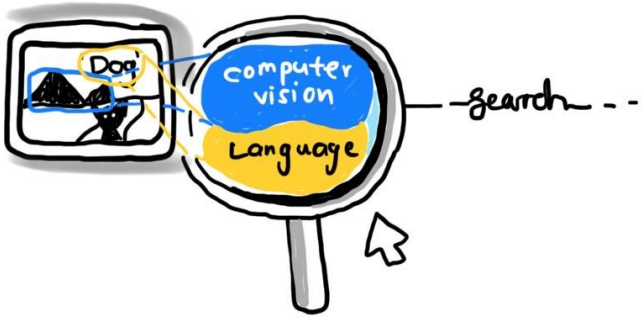
"construction worker in orange safety vest is working on road."

"two young girls are playing with lego toy."

Computer Vision to Image Search

What sets image search apart from other applications?

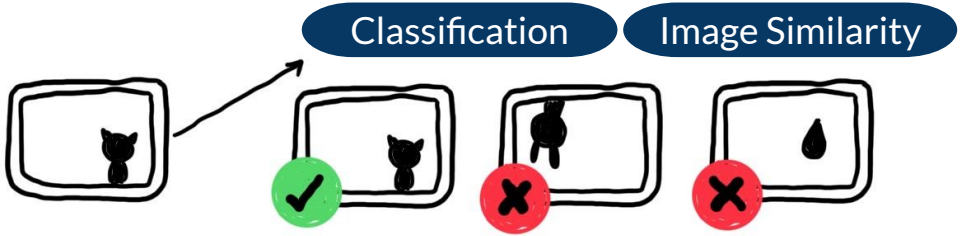
- A combination of multiple computer vision and natural language processing techniques



Examples

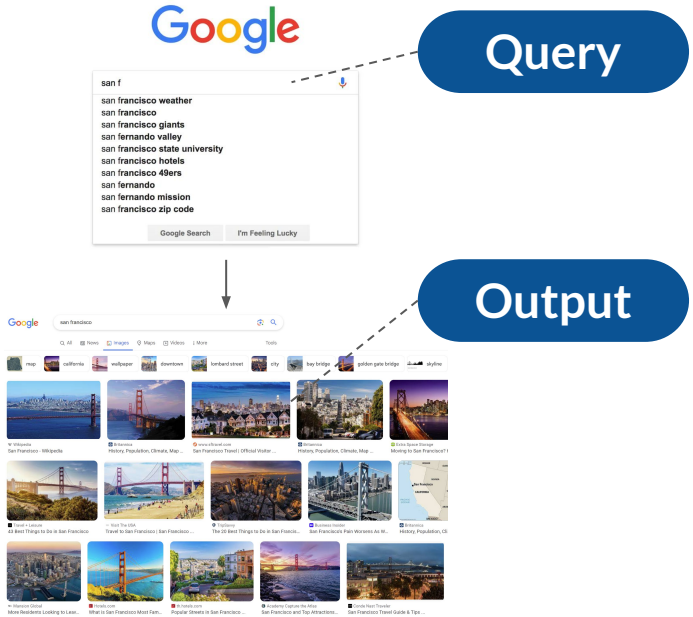
Text-Search-Image

Image-Search-Image



2 What is **Image Search**?

Searching: Google Search

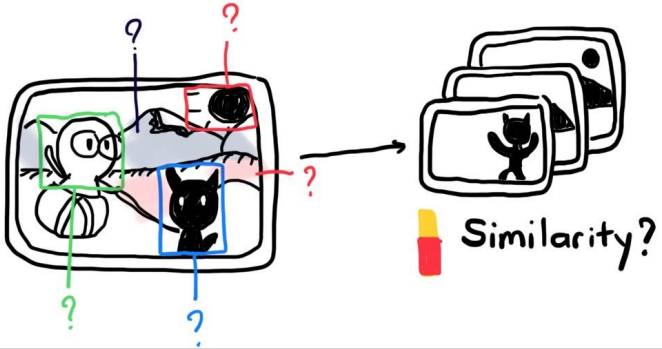


SOTA Similar Terms

Image Retrieval | Image Search

Formal Definition: one of computer vision tasks that involves finding images similar to a provided query from a large database.

Short Definition: find similar images



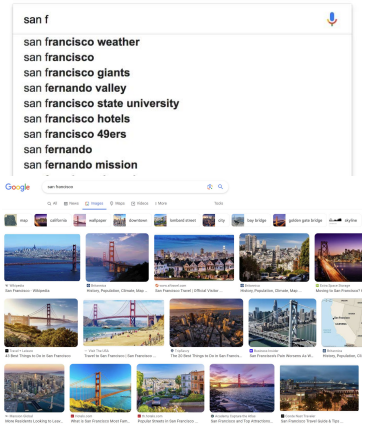
Challenges - Understand what is in the images and what is the best way to compare them

3 Types of **Image Search** in Modern Machine Learning

Image Meta Search

Metadata Keywords, text, …

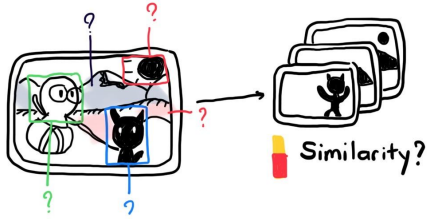
Google



Content-based image retrieval

CBIR

Computer Vision



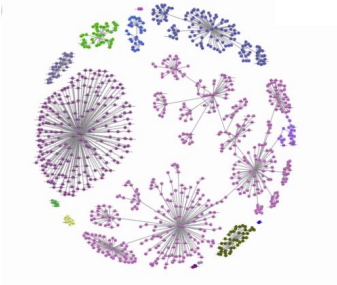
looks at the content (pixels) of images and return results that match a particular query

color, texture, shape/object, etc.

Reverse image search

Image collection exploration

- Mechanism for Explore large digital image repositories
- Solve semantic gap from CBIR



Summarization

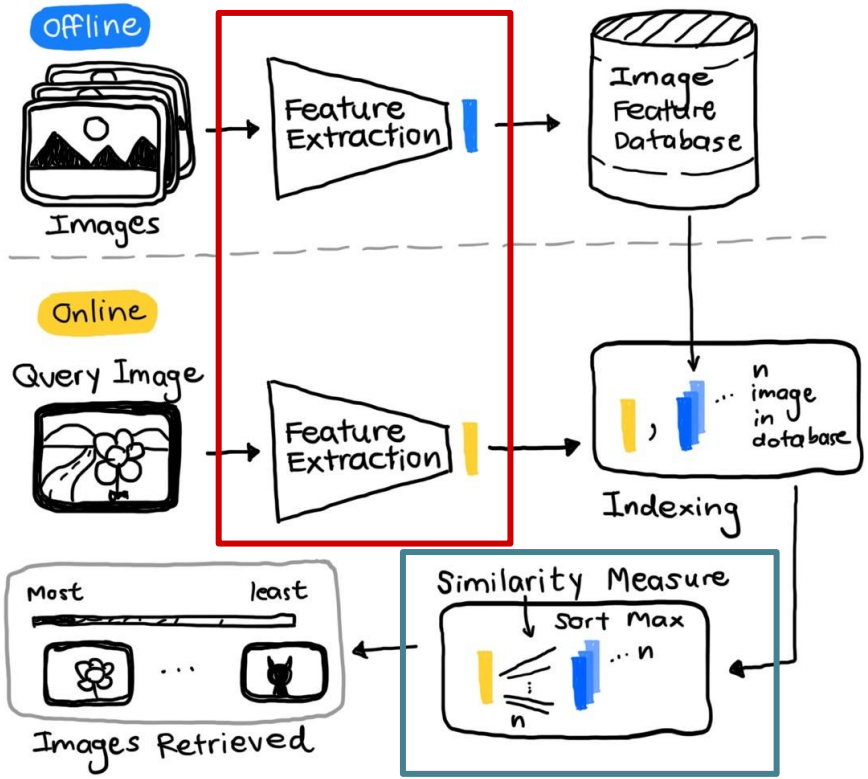
Visualization

Interaction

CBIR

Content-based image retrieval

Goal: search for similar images



Reference:
Alkhawlani, Mohammed & Elmogy, Mohammed & El-Bakry, Hazem. (2015). Text-based, Content-based, and Semantic-based Image Retrievals: A Survey. International Journal of Computer and Information Technology. 4. 58-66.

Feature Extraction

Feature extraction is a process in machine learning and data analysis that involves identifying and extracting relevant features from raw data.

Extract Meaningful Data

Solve **Curse of Dimensionality**

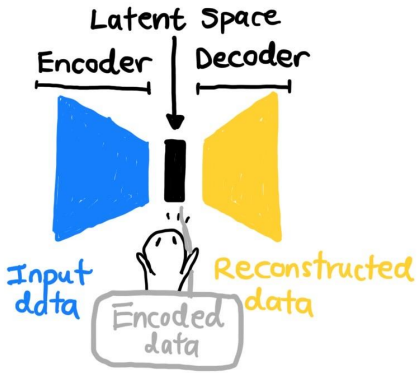
Traditional Methods

Principal Component Analysis (PCA)	Independent Component Analysis (ICA)
Linear Discriminant Analysis (LDA)	Locally Linear Embedding (LLE)
t-distributed Stochastic Neighbor Embedding (t-SNE)	

Source: [Feature Extraction Techniques. An end to end guide on how to reduce a... | by Pier Paolo Ippolito | Towards Data Science]

Modern ML Method

Autoencoder



Latent Space: Pulling out the middle layers of Pretrained Neural Networks

Similarity Measure

Metric for measuring how similar the template (x) is to the target (y)

Distance Equations

Examples of metrics intended for **real-valued vector spaces:**

Euclidean distance

```
sqrt(sum((x - y)^2))
```

Minkowski distance

```
sum(w * |x - y|^p)^(1/p)
```

Manhattan distance

```
sum(|x - y|)
```

SEuclidean distance

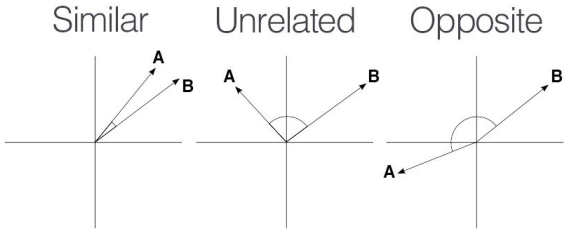
```
sqrt(sum((x - y)^2 / V))
```

Chebyshev distance

```
max(|x - y|)
```

Mahalanobis distance

```
sqrt((x - y)'V^-1(x - y))
```



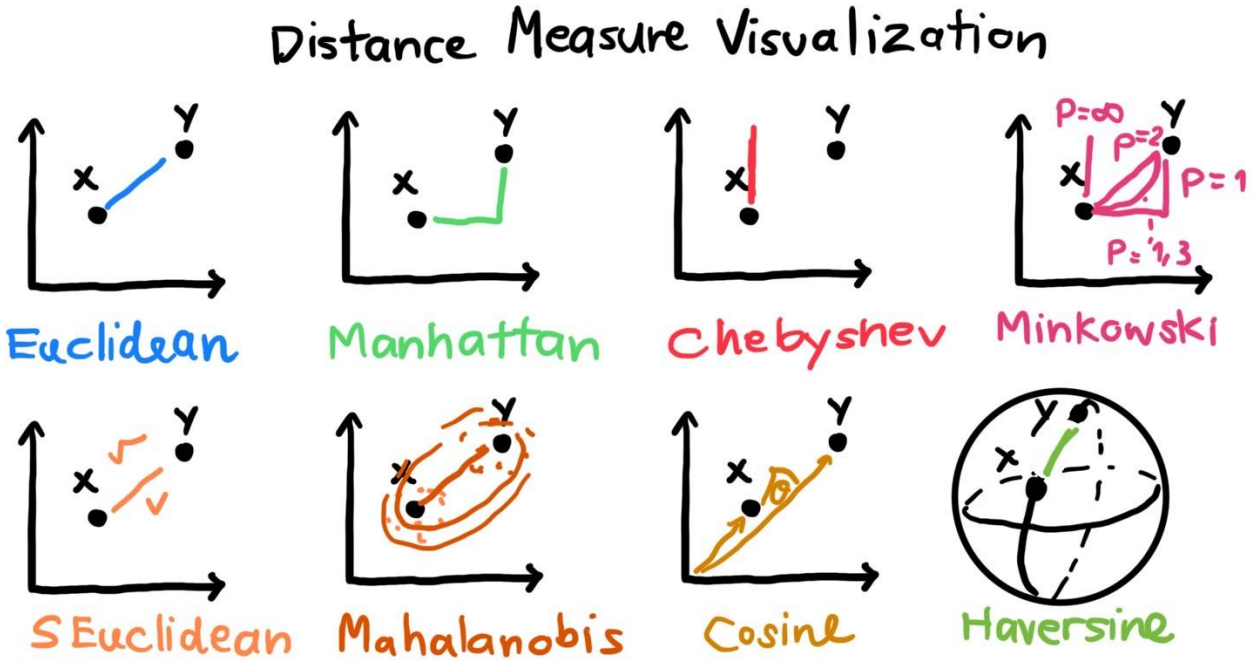
Cosine Similarity

$$\cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}}$$

Source: [sklearn.metrics.DistanceMetric — scikit-learn 1.4.0 documentation]

Metric for measuring how similar the template (x) is to the target (y)

Similarity Measure



4

Hands-On #1
Image Search Image
(CBIR)

colab

Hugging Face

https://bit.ly/3HCBY9K

Food-101 Dataset

Q&A

First Half Break
(10 mins)

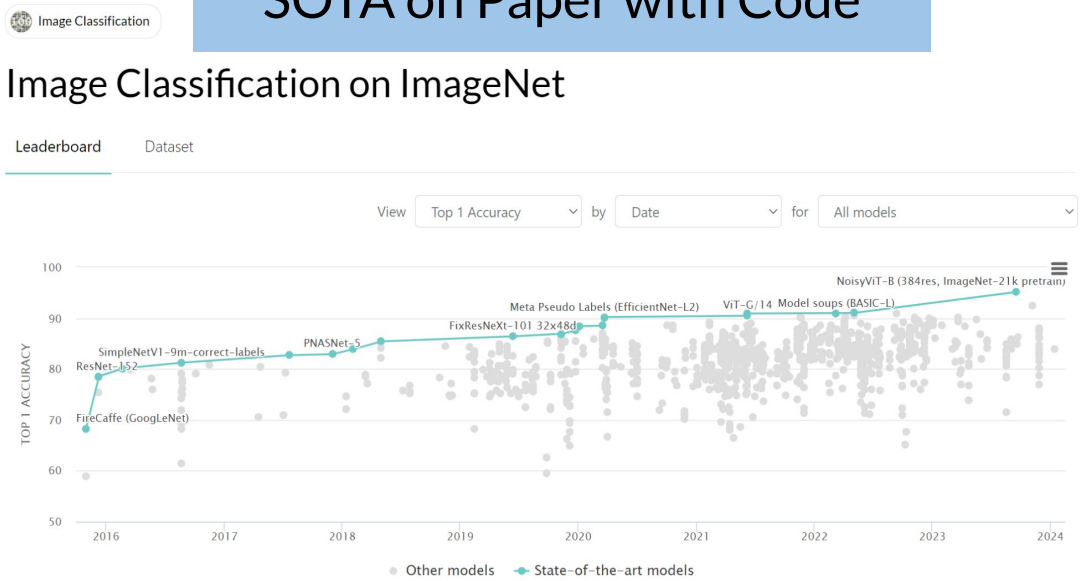
5 State-Of-The-Art Deep Learning with Images

2012 **ImageNet1K Dataset**

1000 Object Classes

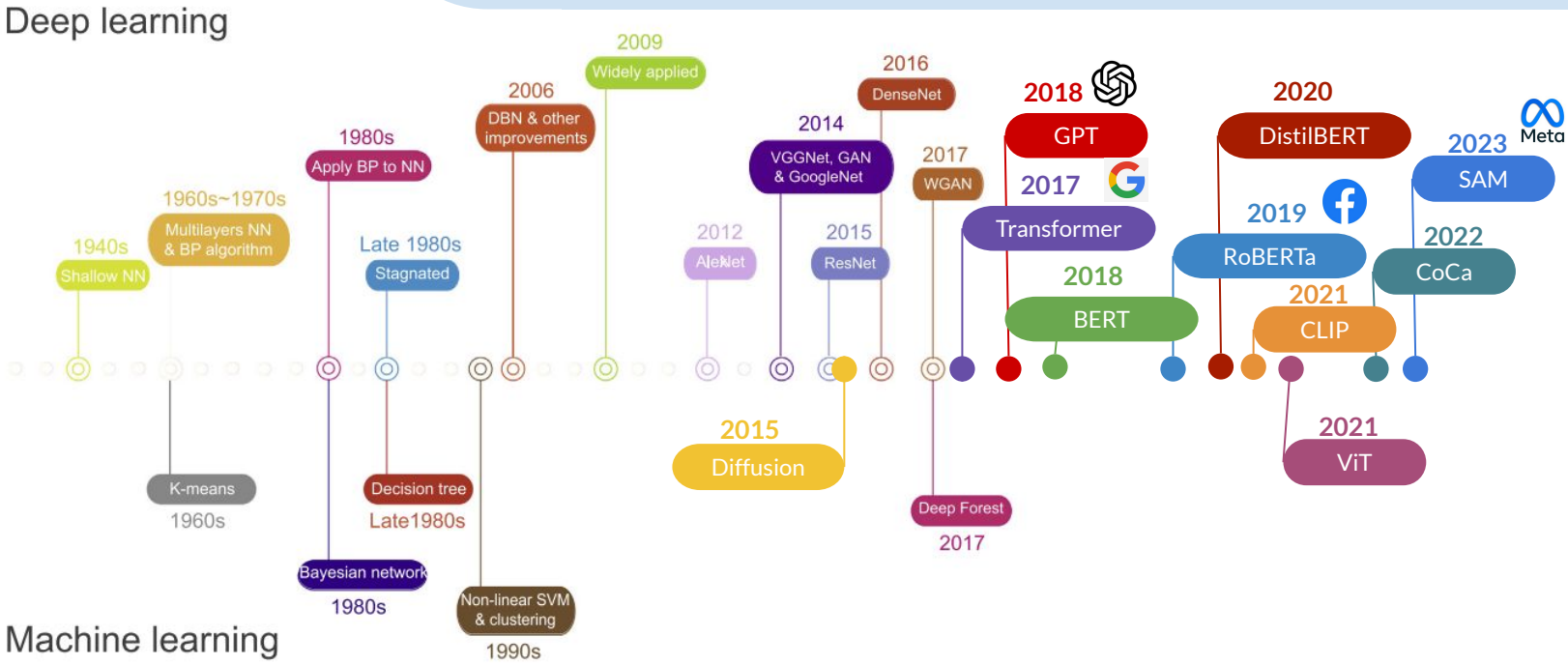
Russakovsky, Olga, et al. ImageNet Large Scale Visual Recognition Challenge. arXiv:1409.0575, arXiv, 29 Jan. 2015. arXiv.org, https://doi.org/10.48550/arXiv.1409.0575.

SOTA on Paper with Code



High computational power → Larger model parameters

Timeline of Breakthrough Image Models

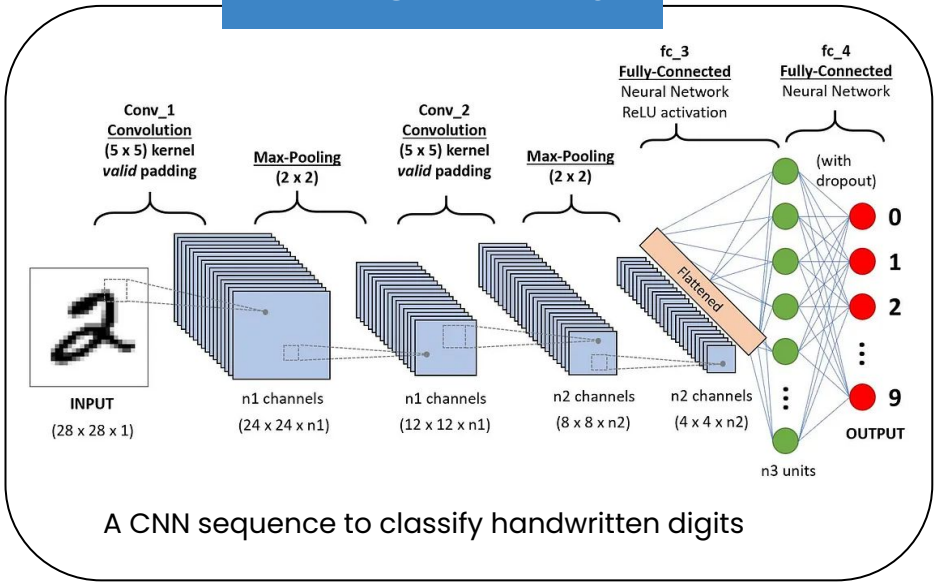


Cao, Chensi & Liu, Feng & Tan, Hai & Song, Deshou & Shu, Wenjie & Li, Weizhong & Zhou, Yiming & Bo, Xiaochen & Xie, Zhi. (2018). Deep Learning and Its Applications in Biomedicine. Genomics, Proteomics & Bioinformatics. 16. 10.1016/j.gpb.2017.07.003.

2012-2017

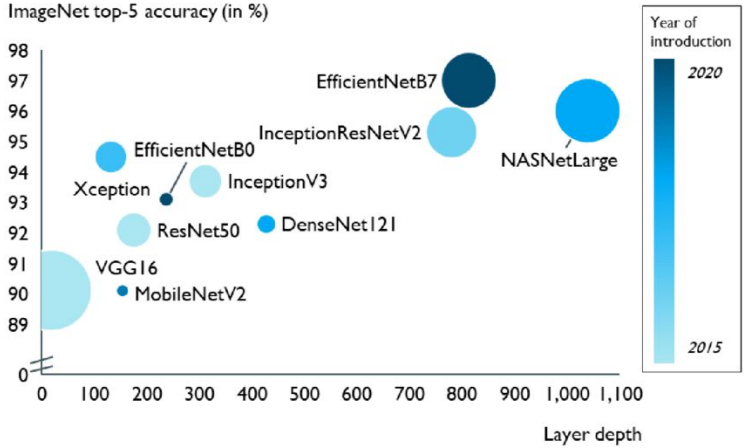
Era of Convolutional Neural Networks

How CNN works



A CNN sequence to classify handwritten digits

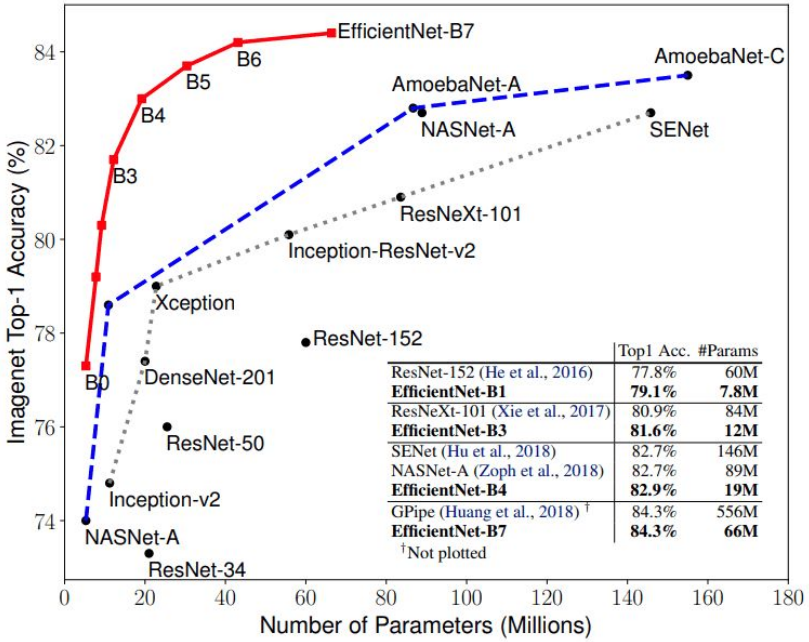
Comparison of Convolutional Neural Network Architectures in Terms of Size and Performance on Traditional ImageNet Benchmark



Tetzlaff, Keno & Hartmann, Jochen & Heitmann, Mark. (2022). Performance of automated image classification.

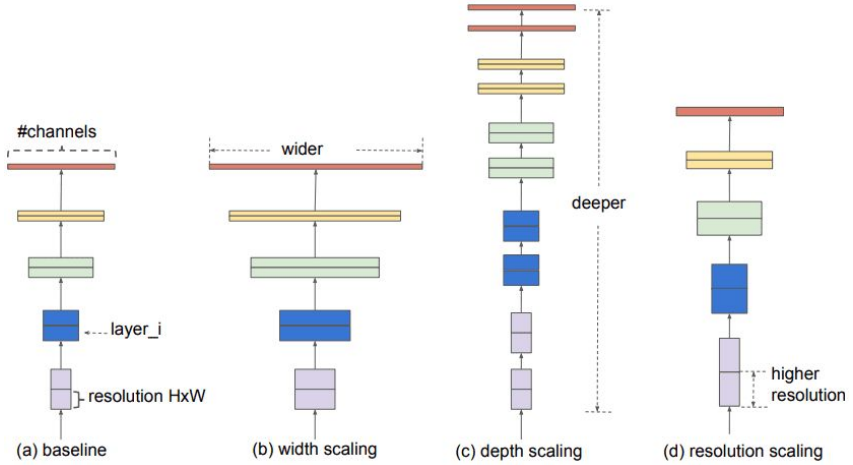
2019

EfficientNet Networks



	Top1 Acc.	#Params
ResNet-152 (He et al., 2016)	77.8%	60M
EfficientNet-B1	**79.1%**	**7.8M**
ResNeXt-101 (Xie et al., 2017)	80.9%	84M
EfficientNet-B3	**81.6%**	**12M**
SENet (Hu et al., 2018)	82.7%	146M
NASNet-A (Zoph et al., 2018)	82.7%	89M
EfficientNet-B4	**82.9%**	**19M**
GPipe (Huang et al., 2018) †	84.3%	556M
EfficientNet-B7	**84.3%**	**66M**

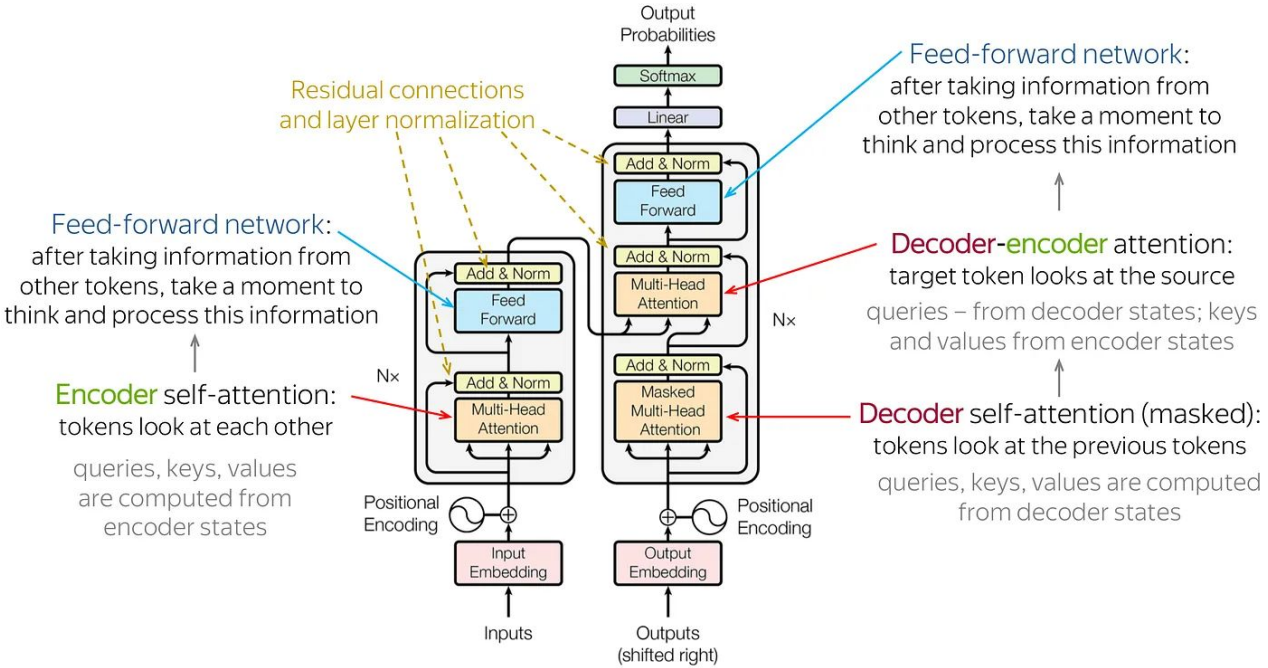
†Not plotted



Tan, Mingxing, and Quoc V. Le. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. arXiv:1905.11946, arXiv, 11 Sept. 2020. arXiv.org, https://doi.org/10.48550/arXiv.1905.11946.

2017 - Now

Transformer-Based Networks



Original paper:
Attention Is All You Need

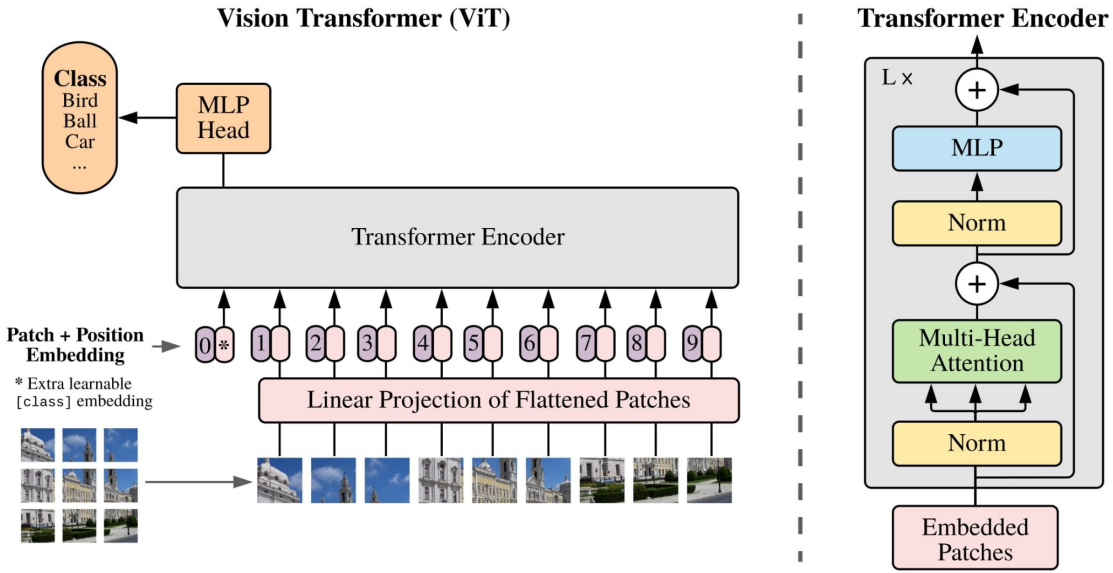
Vaswani, Ashish, et al. Attention Is All You Need. arXiv:1706.03762, arXiv, 1 Aug. 2023. arXiv.org, http://arxiv.org/abs/1706.03762.

Source:
https://stats.stackexchange.com/questions/512242/why-does-transformer-has-such-a-complex-architecture

2017 - Now

Transformer-Based Networks

Vision Transformer (ViT)



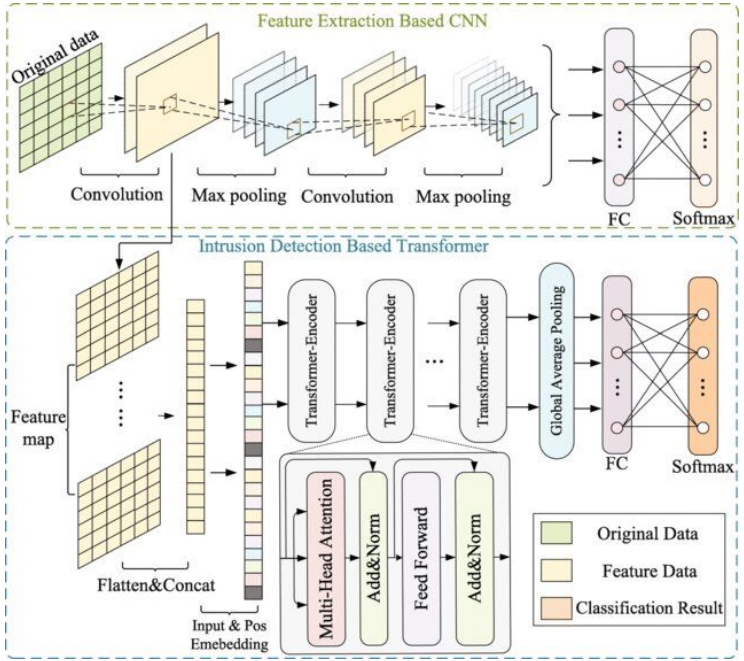
Dosovitskiy, Alexey, et al. An Image Is Worth 16x16 Words: Transformers for Image Recognition at Scale. arXiv:2010.11929, arXiv, 3 June 2021. arXiv.org, https://doi.org/10.48550/arXiv.2010.11929.

2017 - Now

Transformer/Convolution Hybrid Networks

Vanilla is heavy, adapt for a lighter weight

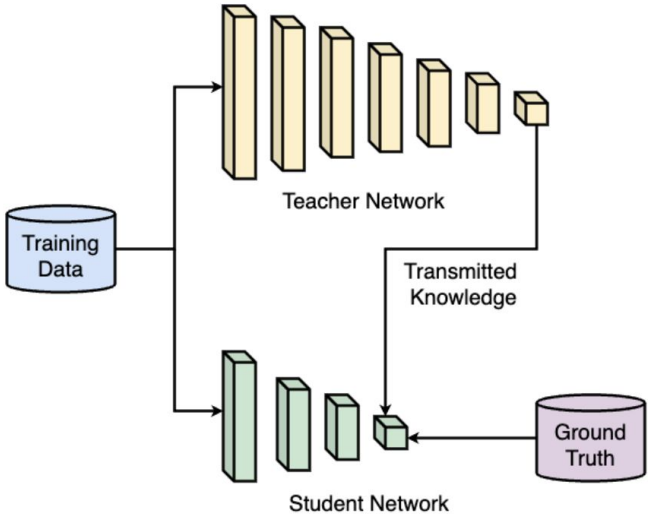
CNN-transformer hybrid



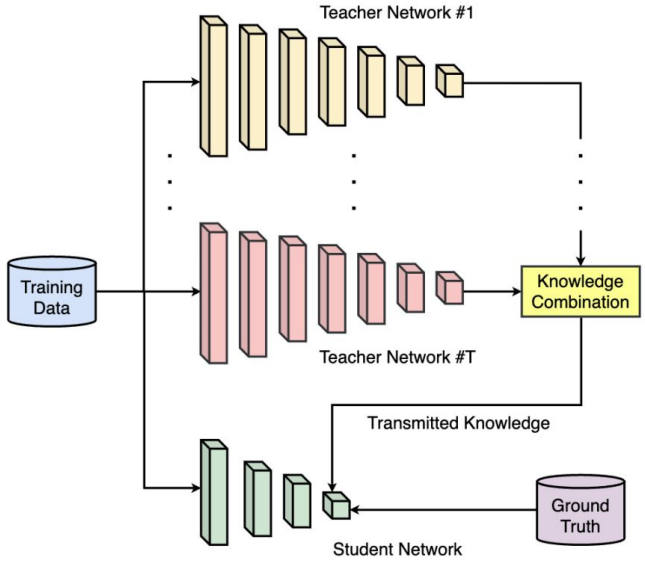
Yao, Ruizhe & Wang, Ning & Chen, Peng & Ma, Di & Sheng, Xianjun. (2022). A CNN-transformer hybrid approach for an intrusion detection system in advanced metering infrastructure. Multimedia Tools and Applications. 82. 10.1007/s11042-022-14121-2.

2017 - Now

Teacher-Student Networks



(a) Knowledge learning from single teacher to single student.



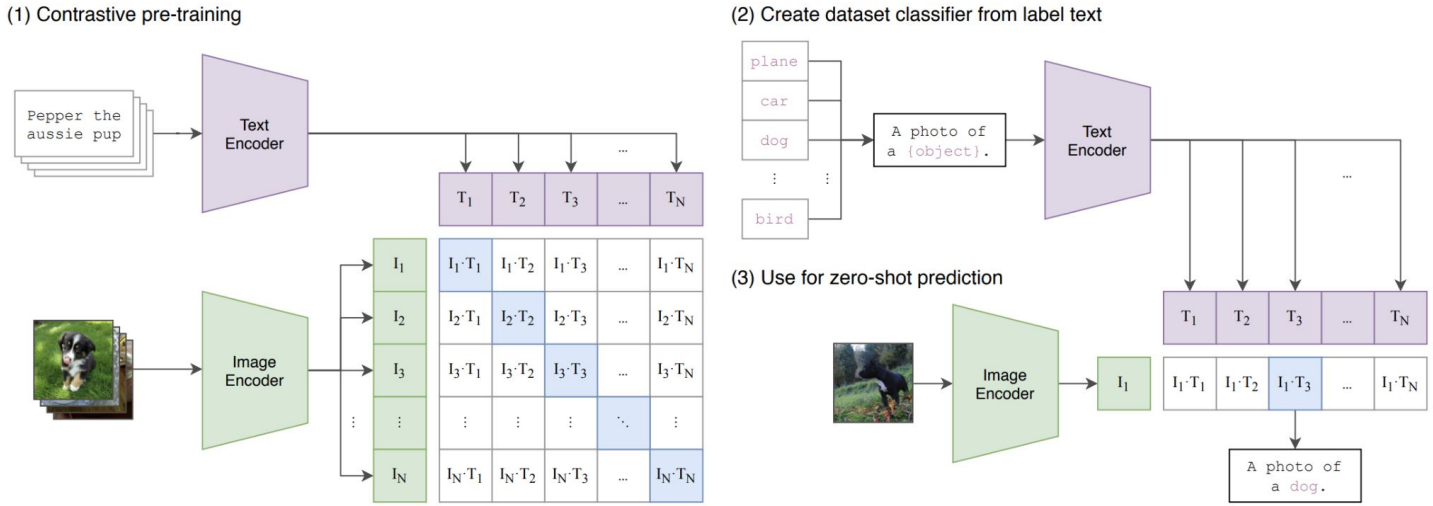
(b) Knowledge learning from multiple teachers to a single student.

Hu, Chengming, et al. Teacher-Student Architecture for Knowledge Learning: A Survey. arXiv:2210.17332, arXiv, 27 Oct. 2022. arXiv.org, http://arxiv.org/abs/2210.17332.

Contrastive Learning Networks

2021 - Now

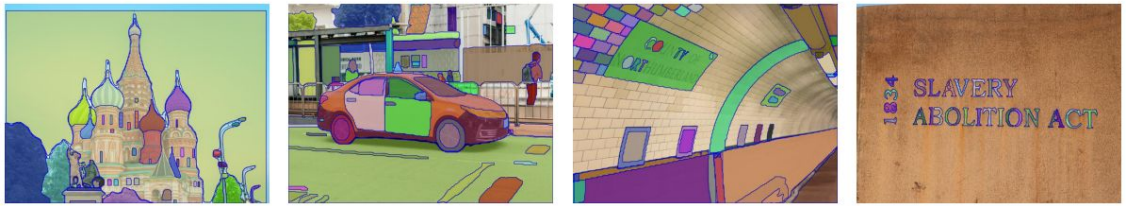
CLIP (Contrastive Language-Image Pretraining)



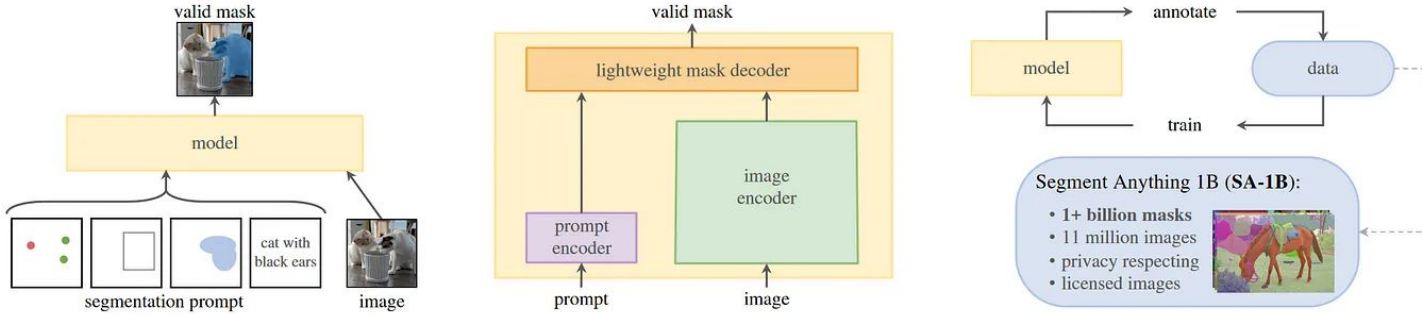
Radford, Alec, et al. Learning Transferable Visual Models From Natural Language Supervision. arXiv:2103.00020, arXiv, 26 Feb. 2021. arXiv.org, http://arxiv.org/abs/2103.00020.

2023 - Now

SAM (Segment Anything)



Source:
https://medium.com/syncedreview/fastsam-drastically-reduces-cost-to-provide-real-time-solution-for-segment-anything-model-466532e86e24



Kirillov, Alexander, et al. Segment Anything. arXiv:2304.02643, arXiv, 5 Apr. 2023. arXiv.org, https://doi.org/10.48550/arXiv.2304.02643.

6

Hands-On #2
Image-Text Search (CLIP)

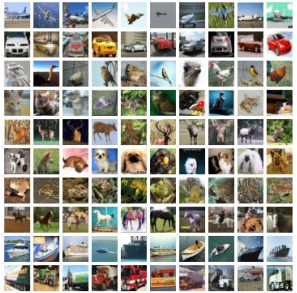
colab

https://bit.ly/3Ug171i

7 Multimodal Models: Image-Text Focus

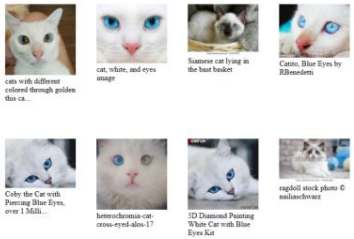
Multimodal Deep Learning is a subset of deep learning that deals with the fusion and analysis of data from multiple modalities, such as text, images, video, audio, and sensor data.

Image-to-Text Retrieval Dataset



MSCOCO

Flickr30k



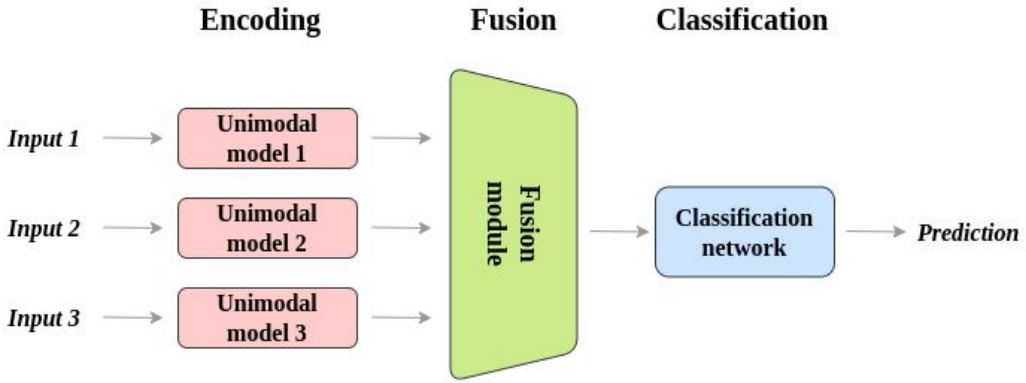
LAION-400M

WHOOPS! A Vision-and-Language Benchmark of Synthetic and Compositional Images

Source: [facebookresearch/multimodal: TorchMultimodal is a PyTorch library for training state-of-the-art multimodal multi-task models at scale. (github.com)]

Multimodal Models: Image-Text Focus

A general multimodal workflow



Multimodal Deep Learning combines the strengths of different modalities to create a more complete representation of the data, leading to better performance on various machine learning tasks.

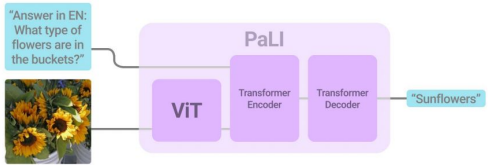
Source: Multimodal Models and Computer Vision: A Deep Dive (roboflow.com)

Significant Progress of Multimodal Models

Visual question answering

Combine visual Input and NLP

PaLI (Pathways Language and Image model)



Text-to-image and Image-to-text task

DALL-E

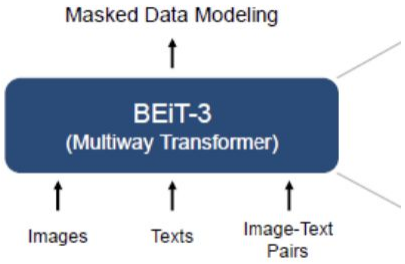
Stable Diffusion

Midjourney

Natural language for visual reasoning

Microsoft

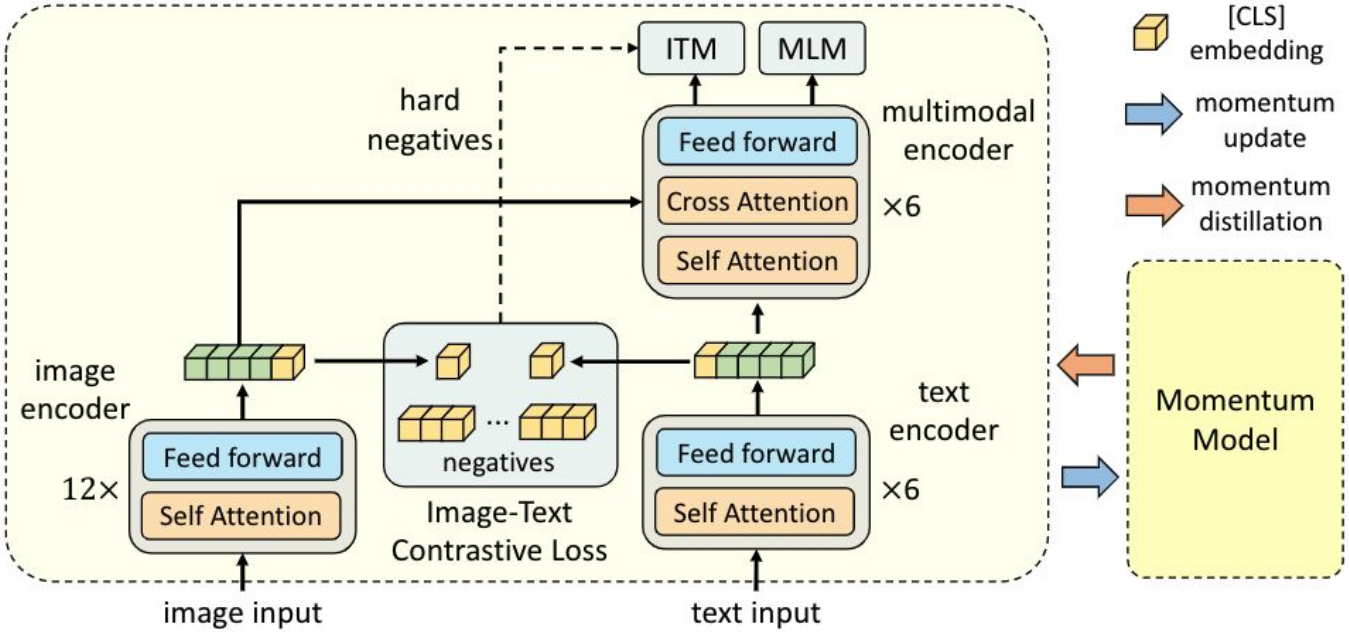
BEiT-3



ALBEF

Align before Fuse: Vision and Language Representation Learning with Momentum Distillation

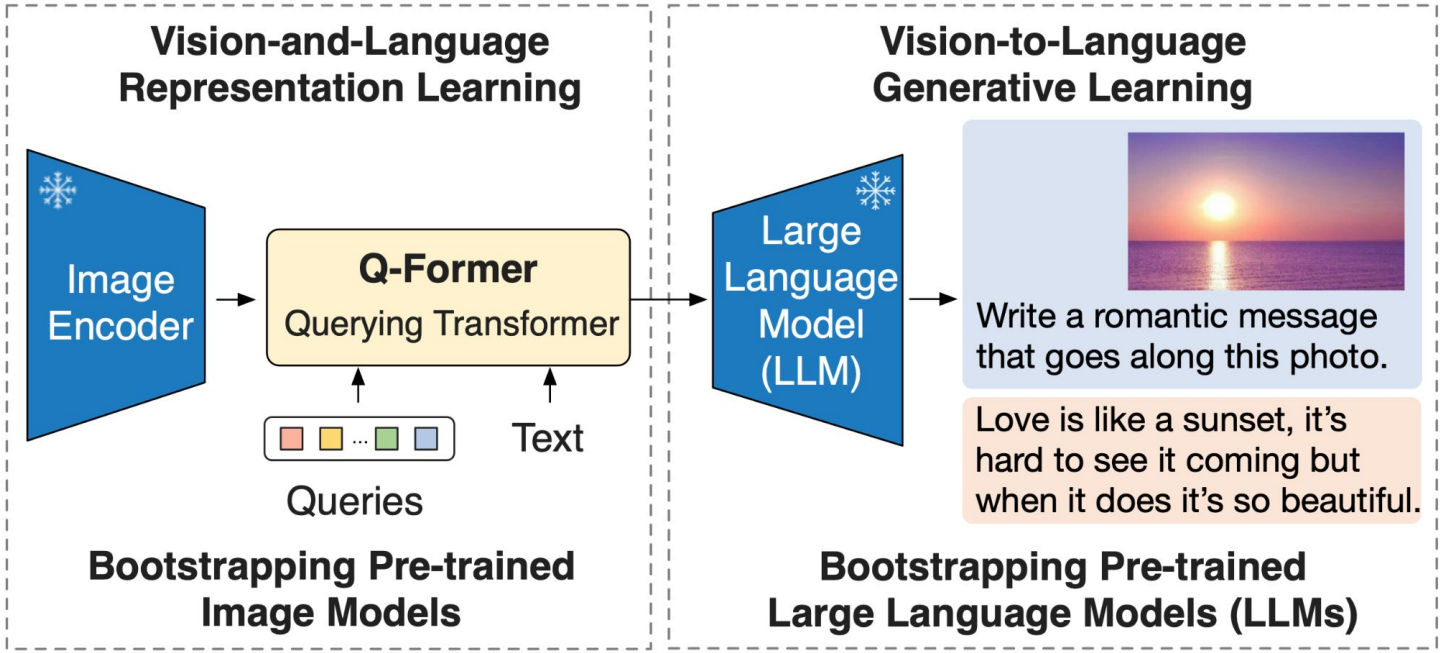
Junnan Li, Ramprasaath R. Selvaraju, Akhilesh D. Gotmare
Shafiq Joty, Caiming Xiong, Steven C.H. Hoi



BLIP-2

BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models

Junnan Li, Dongxu Li, Silvio Savarese, Steven Hoi



BEiT-3

Image as a Foreign Language: BEiT Pretraining for All Vision and Vision-Language Tasks

Wenhui Wang, Hangbo Bao, Li Dong, Johan Bjorck, Zhiliang Peng, Qiang Liu, Kriti Aggarwal, Owais Khan Mohammed, Saksham Singhal, Subhojit Som, Furu Wei

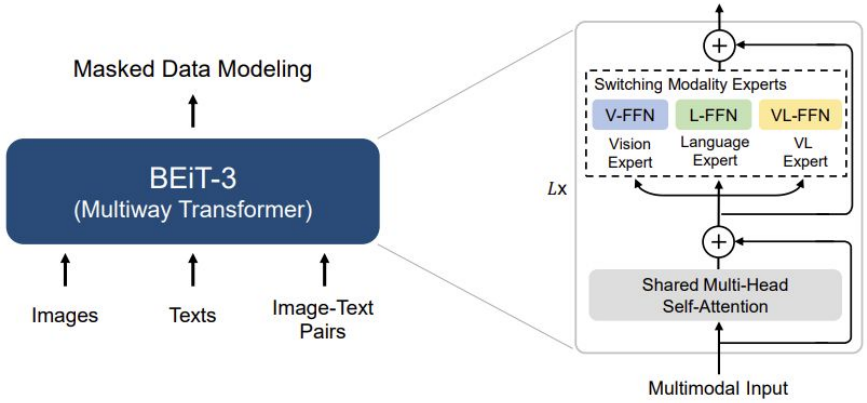
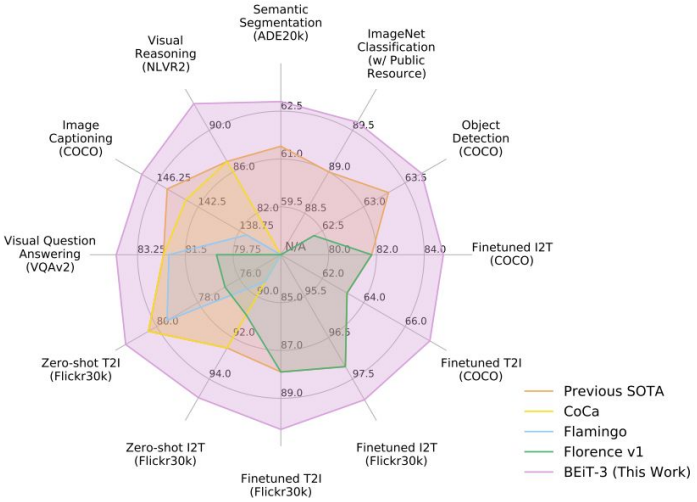


Figure 2: Overview of BEIT-3 pretraining. We perform masked data modeling on monomodal (i.e., images, and texts) and multimodal (i.e., image-text pairs) data with a shared Multiway Transformer as the backbone network.



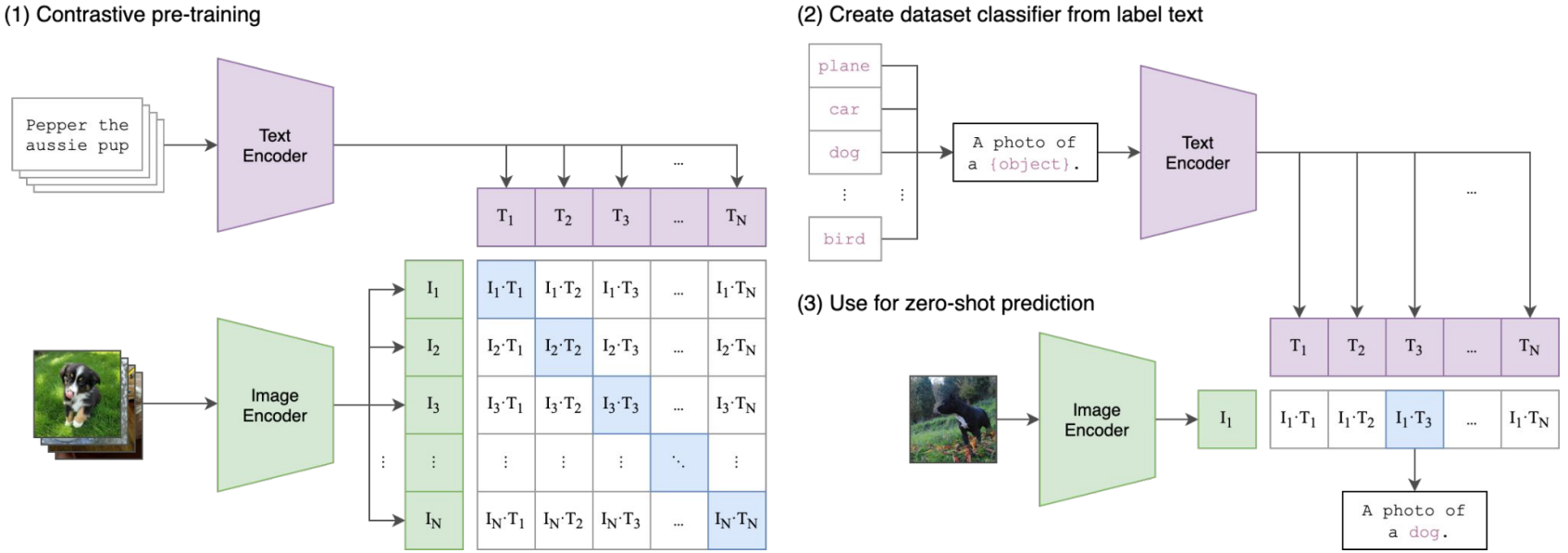
Github: https://github.com/microsoft/unilm/tree/master/beit3

CLIP

Learning Transferable Visual Models From Natural Language Supervision

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, Ilya Sutskever

26 Feb 2021

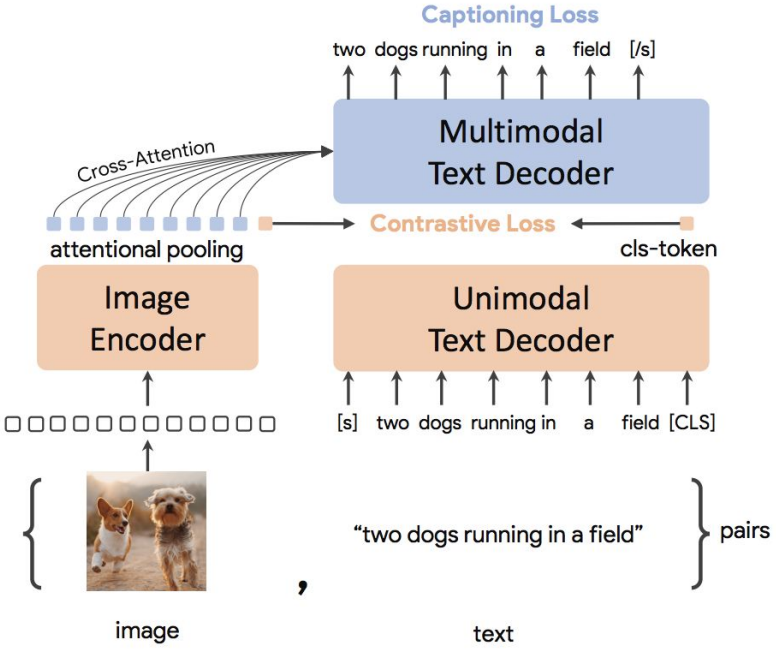


CoCa: Contrastive Captioners are Image-Text Foundation Models

Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, Yonghui Wu

CoCa

14 June 2022



Algorithm 1 Pseudocode of Contrastive Captioners architecture.

```
# image, text.ids, text.labels, text.mask: paired {image, text} data
# con_query: 1 query token for contrastive embedding
# cap_query: N query tokens for captioning embedding
# cls_token_id: a special cls_token_id in vocabulary

def attentional_pooling(features, query):
  out = multihead_attention(features, query)
  return layer_norm(out)

img_feature = vit_encoder(image) # [batch, seq_len, dim]
con_feature = attentional_pooling(img_feature, con_query) # [batch, 1, dim]
cap_feature = attentional_pooling(img_feature, cap_query) # [batch, N, dim]

ids = concat(text.ids, cls_token_id)
mask = concat(text.mask, zeros_like(cls_token_id)) # unpad cls_token_id
txt_embs = embedding_lookup(ids)
unimodal_out = lm_transformers(txt_embs, mask, cross_attn=None)
multimodal_out = lm_transformers(
    unimodal_out[:, :-1, :], mask, cross_attn=cap_feature)
cls_token_feature = layer_norm(unimodal_out)[:, -1:, :] # [batch, 1, dim]

con_loss = contrastive_loss(con_feature, cls_token_feature)
cap_loss = softmax_cross_entropy_loss(
    multimodal_out, labels=text.labels, mask=text.mask)
```

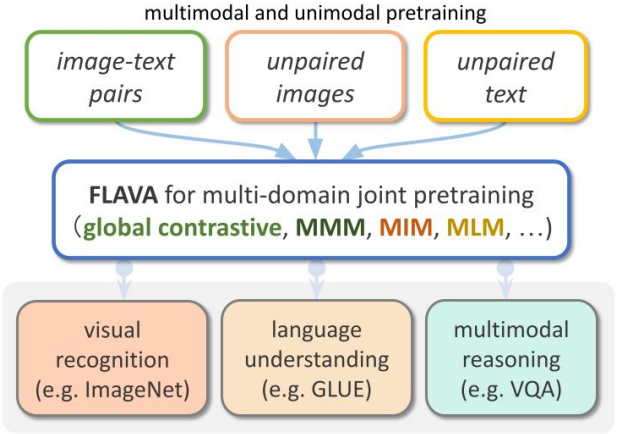
`vit_encoder`: vision transformer based encoder; `lm_transformer`: language-model transformers.

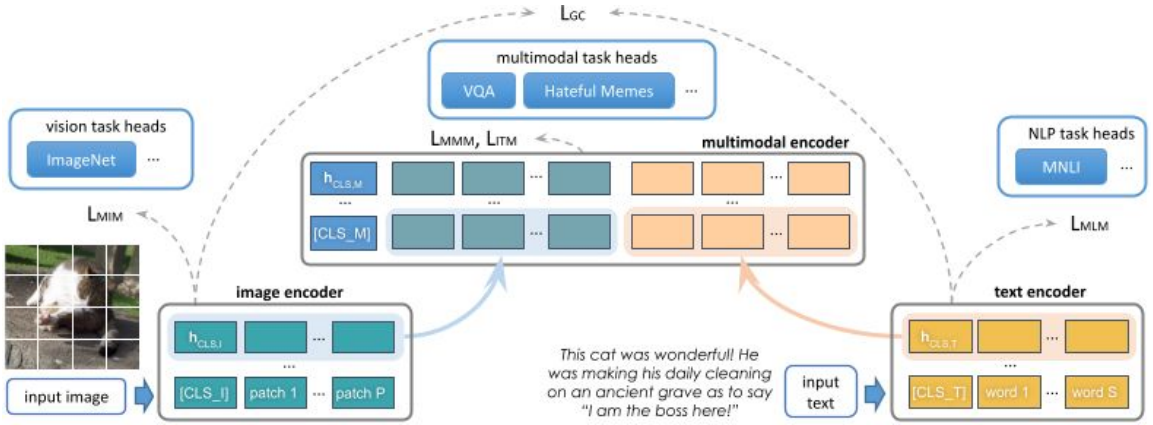
FLAVA

FLAVA: A Foundation Language And Vision Alignment Model

Amanpreet Singh, Ronghang Hu, Vedanuj Goswami, Guillaume Couairon, Wojciech Galuba, Marcus Rohrbach, Douwe Kiela

29 March 2022





Flamingo

Flamingo: a Visual Language Model for Few-Shot Learning

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, and others

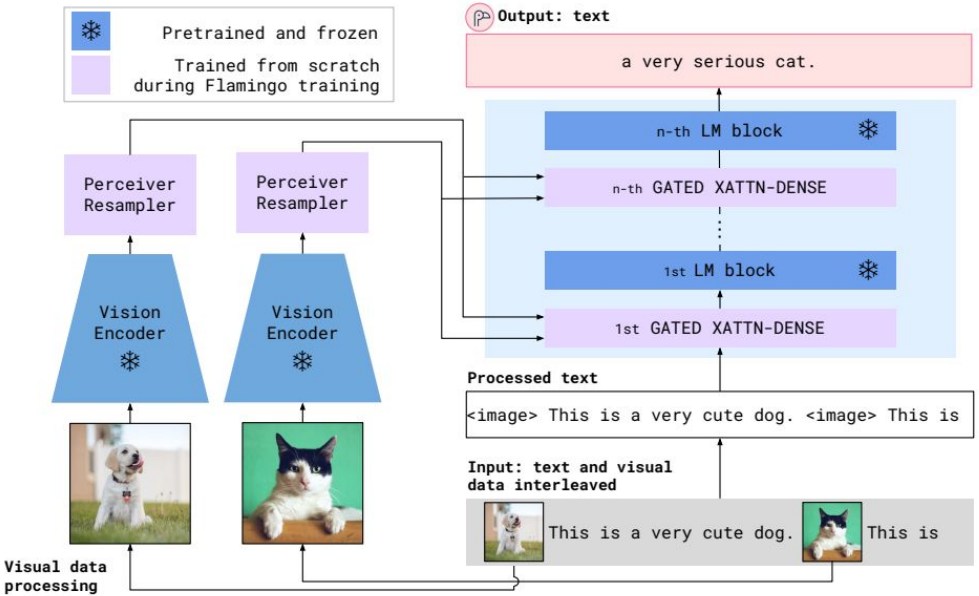


Figure 3 | **Overview of the Flamingo model.** The Flamingo models are a family of visual language model (VLM) that can take as input visual data interleaved with text and can produce free-form text as output. Key to its performance are novel architectural components and pretraining strategies described in Section 3.

MDETR

MDETR -- Modulated Detection for End-to-End Multi-Modal Understanding

Aishwarya Kamath, Mannat Singh, Yann LeCun, Gabriel Synnaeve, Ishan Misra, Nicolas Carion

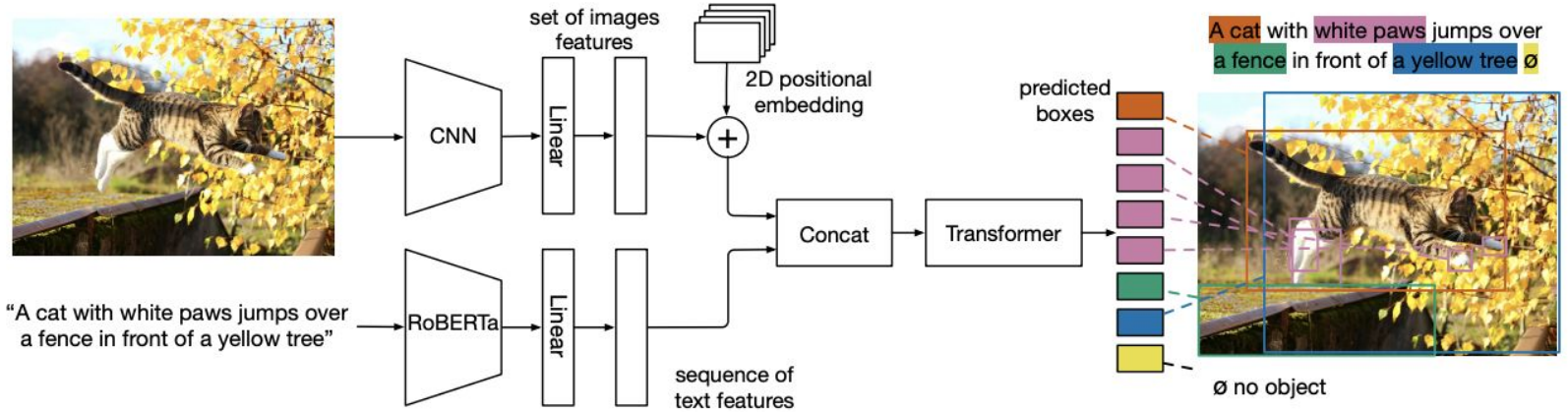


Figure 2: MDETR uses a convolutional backbone to extract visual features, and a language model such as RoBERTa to extract text features. The features of both modalities are projected to a shared embedding space, concatenated and fed to a transformer encoder-decoder that predicts the bounding boxes of the objects and their grounding in text.

8

Hands-On #3
Image-Text Segmentation Search (SAM & CLIP)

Special pipeline, but incomplete!

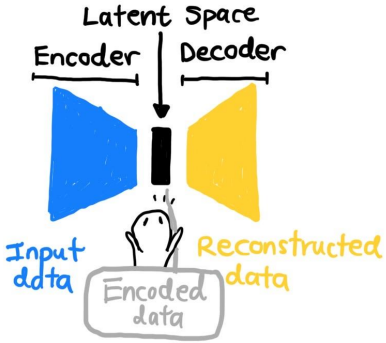
colab

https://bit.ly/47S6yXA

Summary

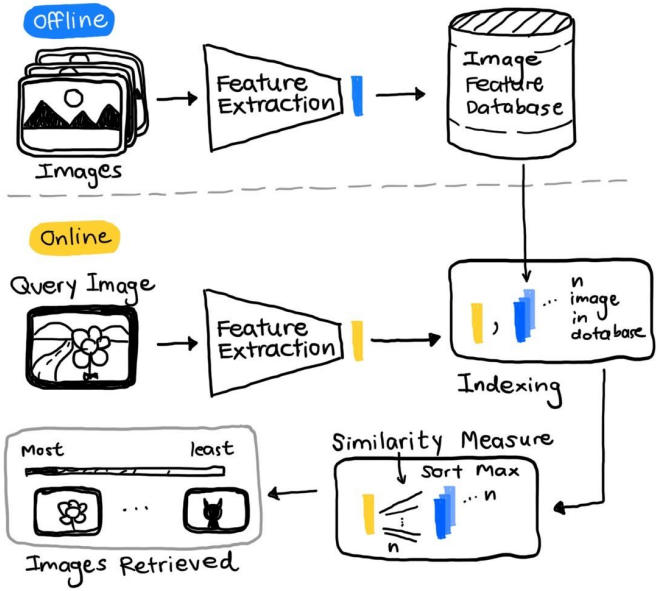
Overview of What We Learned

Image search is finding similar images



Feature Extraction:
Latent Space

Pipeline of image search



You can use all model keywords in this lecture to tackle the upcoming hackathon!

Q&A

Good luck
on the hackathon!