Multimodal Embeddings

Omar Hassan

Jan Kels

24.05.2023

Outline

1. The Multimodality Paradigm
   What & Why?

2. The Modalites
   Language & Vision

3. Multimodal Learning
   Tasks & Architectures

1. The Multimodality Paradigm | What?

- Definition
  - Depends on context

  - Loosely speaking:
    "Different forms of data representing semantically related knowledge"

  - Examples:
    - Images
    - Texts
    - Videos
    - Audios
    - …

  - Images → Vision
    Texts → Language

1. The Multimodality Paradigm | Why?

- “Humans tends to learn with multimodal approach”
- Better contextual representation about concepts
- Symbol Grounding Problem (Harnad 1990)
- Other reasons

1. The Multimodality Paradigm | Why?

- Other Reasons:
  - Complementary Information

    Different modalities provide distinct and complementary information about a given concept or instance.

  - Enhanced Understanding

    Incorporating multiple modalities can lead to a better understanding of the data.

  - Robustness to Missing Data

    Multimodal embeddings can handle incomplete data in one modality by relying on the available information from other modalities.

2. The Modalities | Language

Main Objective:

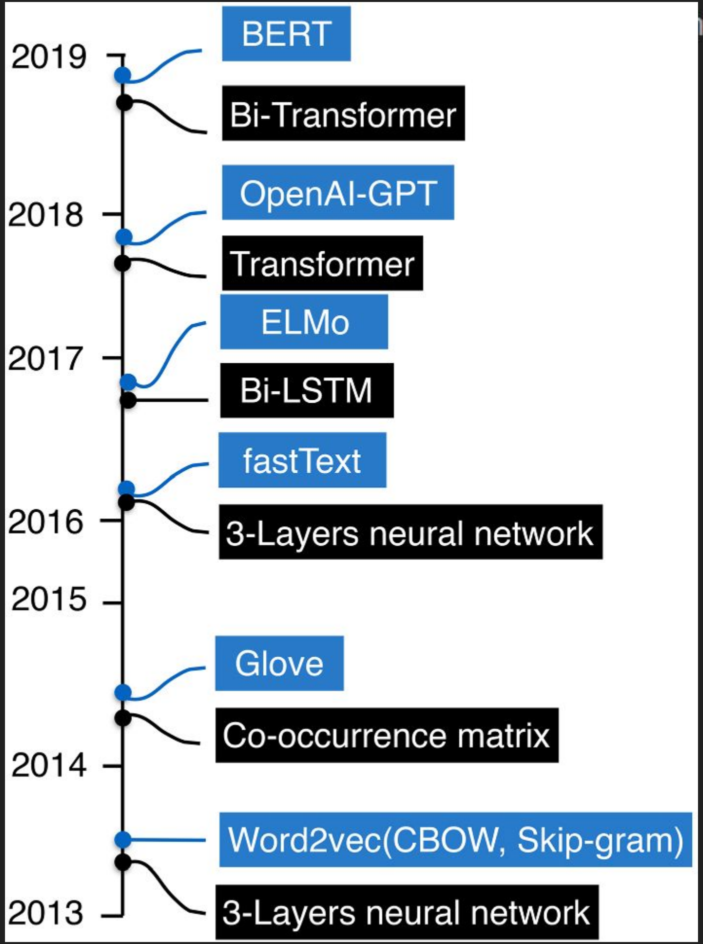
"How can the machine understand texts?"

2013 - 2023

10 Years of impressive progress

From word2vec to GPT-4

*After 2019: GPT-3, **CLIP**, GPT-4*



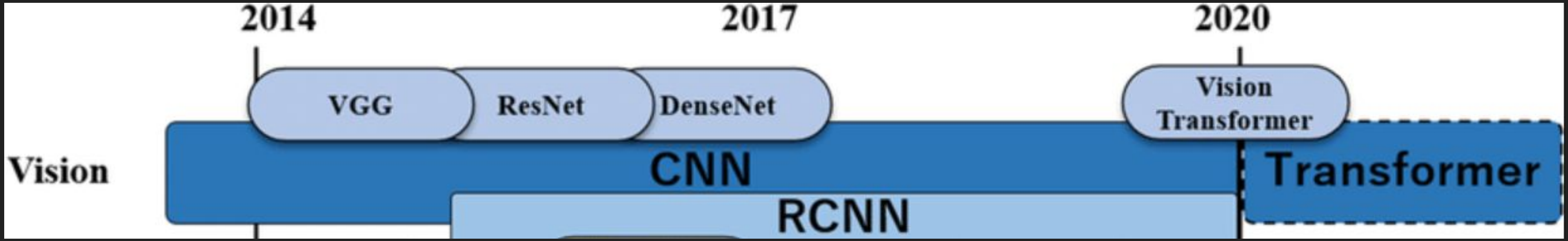
Wang, Shirui & Zhou, Wenan & Jiang, Chao. (2020). A survey of word embeddings based on deep learning.

2. The Modalities | Vision

Main Objective:

“How can the machine understand images?”

From AlexNet to Vision Transformer



Shin, Andrew & Ishii, Masato & Narihira, Takuya. (2022). Perspectives and Prospects on Transformer Architecture for Cross-Modal Tasks with Language and Vision. International Journal of Computer Vision.

Vision Transformers

- Introduced in 2020

  "An Image is Worth 16*16 Words: Transformers for Image Recognition at Scale"

- Beat CNN SOTA models by 4X

- Architecture:
  1. Split an image into patches
  2. Flatten the patches
  3. Flattened patches → Embeddings
  4. Add positional embeddings
  5. Feed the sequence as an input to a standard transformer encoder
  6. Pretrain the model with image target labels (fully supervised on a huge dataset)
  7. Finetune on the downstream dataset for image classification

- Interactive Architecture

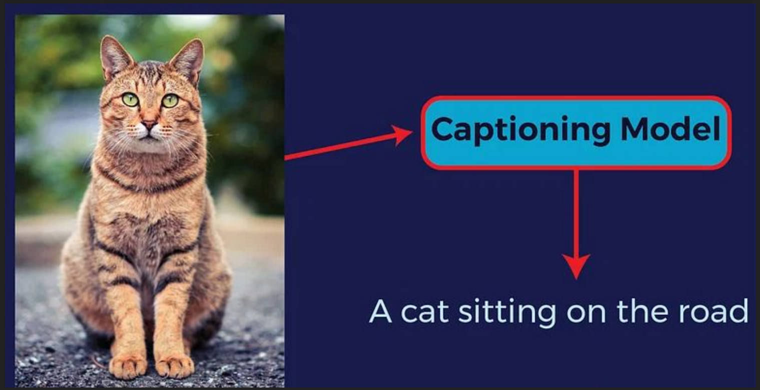
Ever wondered, why transformers are called transformers?!

Vaswani:

- “Attention was a key to transformers”
- “But Attention-Net didn’t sound very exciting.”
- “Then a senior software engineer on the team, came up with the name Transformer.”
- “He argued we were transforming representations”

3. Multimodal Learning | Tasks

- Generation tasks
  - Visual Captioning (VC)



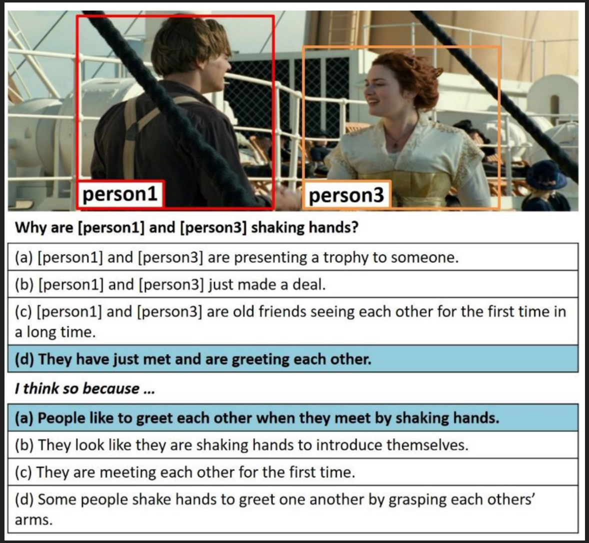
https://www.projectpro.io/article/image-captioning-deep-learning-project/717

Visual Question Answering (VQA)

https://visualqa.org/

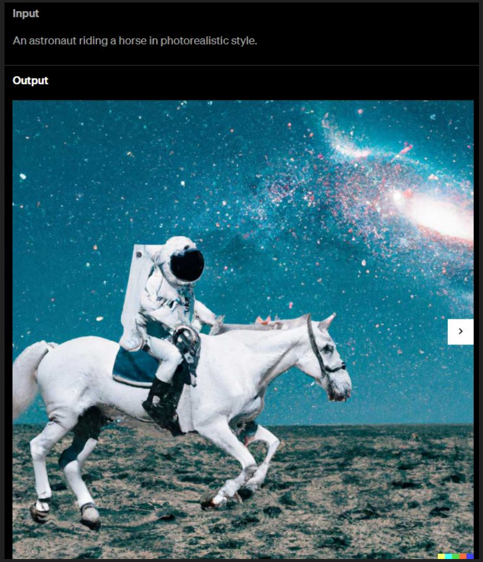
3. Multimodal Learning | Tasks

- Generation tasks
  - Visual Commonsense Reasoning (VCR)



Lee, J.; Kim, I. Vision–Language–Knowledge Co-Embedding for Visual Commonsense Reasoning. Sensors 2021, 21, 2911.

Visual Generation (VG)



https://openai.com/product/dall-e-2

3. Multimodal Learning | Architectures

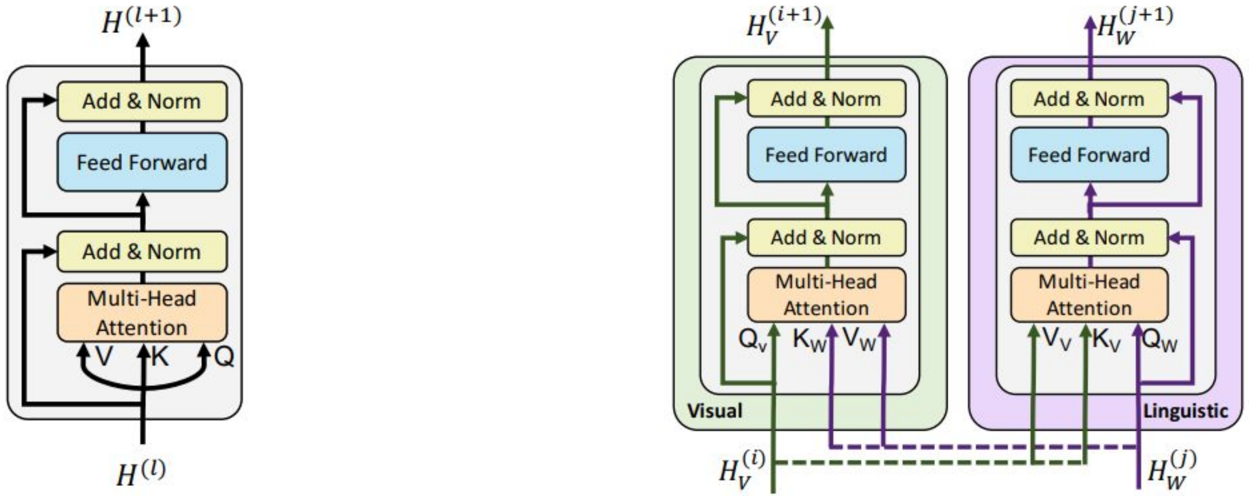
- BERT-like Architectures
  - Two-stream models
  - Single-stream models
- Generative models
- Contrastive Learning models

3. Multimodal Learning | Architectures

- BERT-like Architectures
  - Big Zoo!
    - VisualBERT ViLBERT VL-BERT LXMERT
    - Pixel-BERT ImageBERT VD-BERT UNITER
  - Process texts and images with a transformer-like architecture
  - Pretrained on huge datasets
  - Fine-tuned on downstream tasks

3. Multimodal Learning | Architectures

- BERT-like Architectures
  - Two-stream models
    - Example: ViLBERT
    - A separate transformer for each modality
    - A “co-attention” module is added



(a) Standard encoder transformer block

(b) Our co-attention transformer layer

Figure 2: We introduce a novel co-attention mechanism based on the transformer architecture. By exchanging key-value pairs in multi-headed attention, this structure enables vision-attended language features to be incorporated into visual representations (and vice versa).

Lu, J., Batra, D., Parikh, D., & Lee, S. (2019). ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks.

3. Multimodal Learning | Architectures

- BERT-like Architectures
  - Single-stream models
    - Examples: VisualBERT
    - Encodes both modalities within the same module

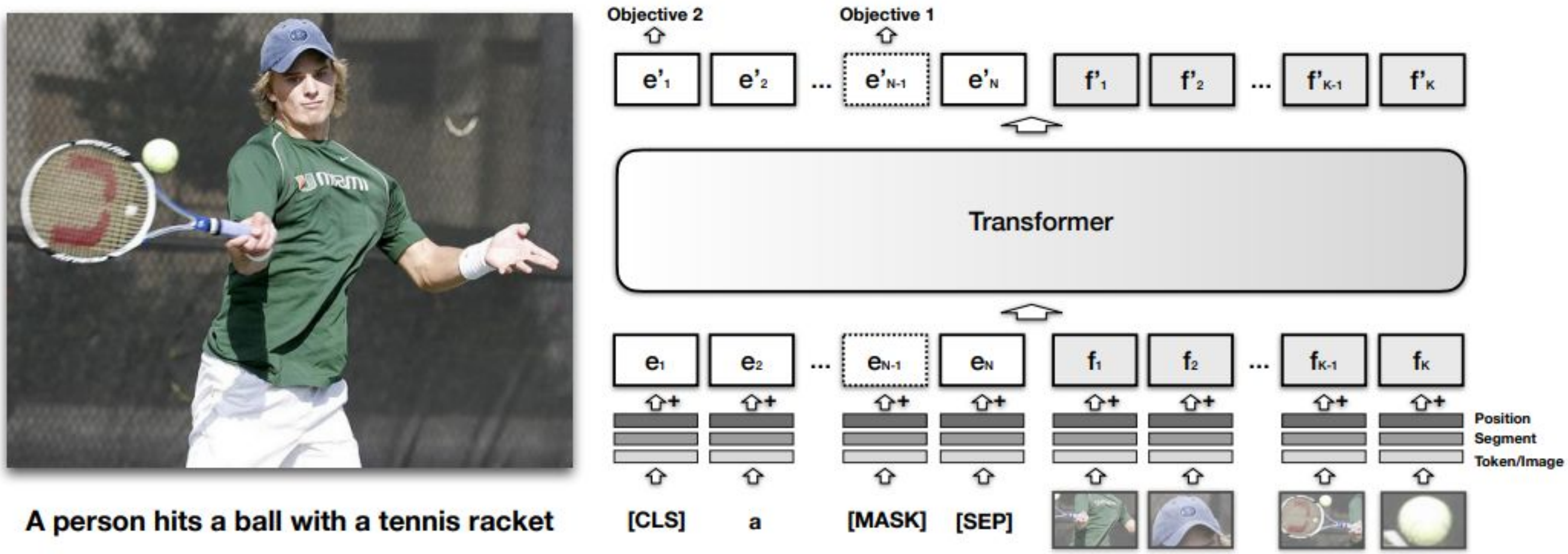
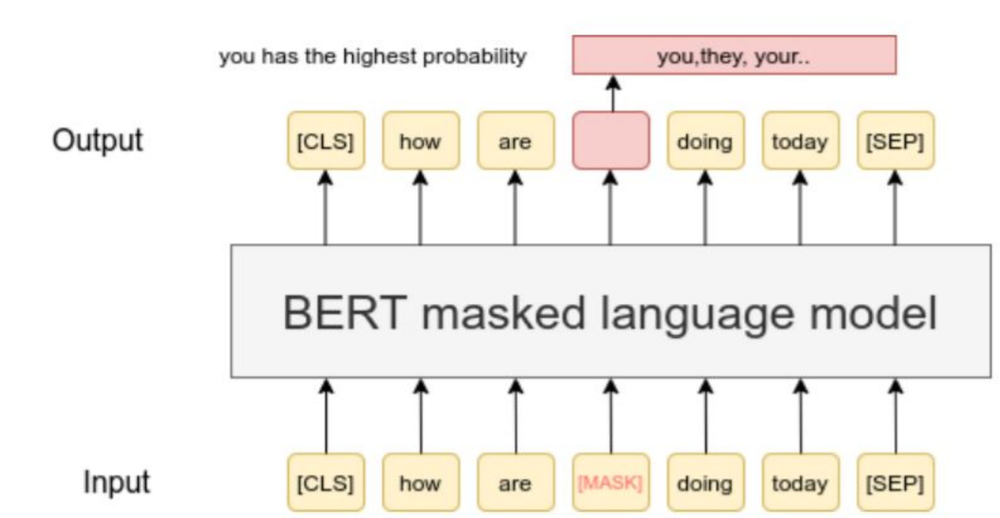


Figure 2: The architecture of VisualBERT. Image regions and language are combined with a Transformer to allow the self-attention to discover implicit alignments between language and vision. It is pre-trained with a masked language modeling (Objective 1), and sentence-image prediction task (Objective 2), on caption data and then fine-tuned for different tasks. See §3.3 for more details.

Li, L. H., Yatskar, M., Yin, D., Hsieh, C., & Chang, K. (2019). VisualBERT: A Simple and Performant Baseline for Vision and Language. ArXiv. /abs/1908.03557

3. Multimodal Learning | Architectures

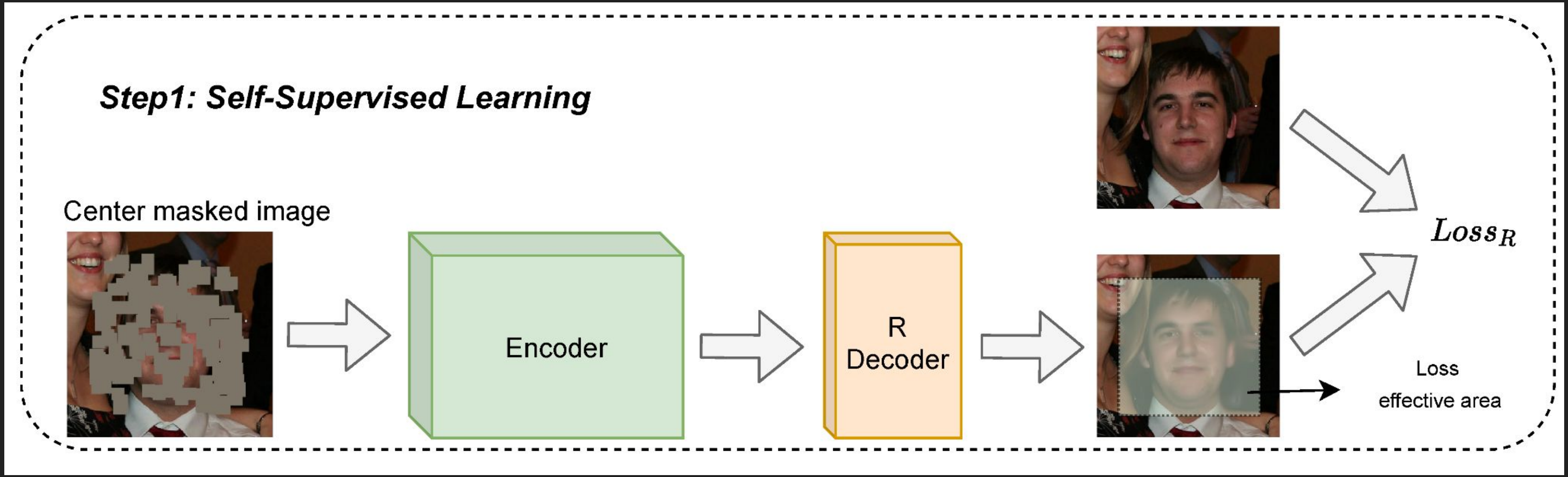
- BERT-like Architectures - Pre-training strategies
  - Masked Language Modeling



https://www.sbert.net/examples/unsupervised_learning/MLM/README.html

3. Multimodal Learning | Architectures

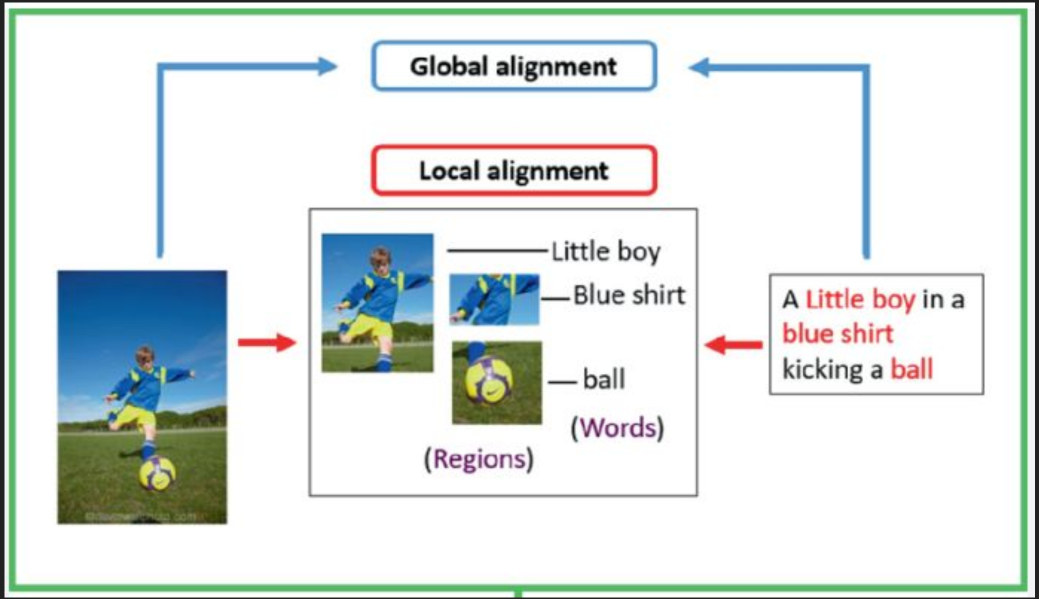
- BERT-like Architectures - Pre-training strategies
  - Masked Region Modeling



Li, Z., Cao, L., Wang, H., & Xu, L. (2022). A Masked Self-Supervised Pretraining Method for Face Parsing. Mathematics, 10(12), 2002. https://doi.org/10.3390/math10122002

3. Multimodal Learning | Architectures

- BERT-like Architectures - Pre-training strategies
  - Image-Text Alignment
  - Word-Region Alignment



Abdullah, T., Rangarajan, L. (2021). Image-Text Matching: Methods and Challenges. Networks and Systems, vol 204. Springer, Singapore.

3. Multimodal Learning | Architectures

- Generative models
  - DALL-E
    - Text-to-Image
    - Closed-source
    - VQ-VAE (Vector Quantized Variational AutoEncoders) + BART
  - GLIDE
    - Diffusion Models

3. Multimodal Learning | Architectures

- Contrastive Learning models
  - Visual-Semantic Embeddings
    - CLIP
    - ALIGN
    - Florence

Take-home Message

- Multimodal Learning is a step towards Generalization
- Unsupervised & Self-supervised pre-training has suppressed supervised approaches. *“Yann LeCun 2022”*
- Transformers Architectures made it possible for Language and Vision to be learnt in a multi-modal fashion effectively.
- 3 Main Design Choices:
    1. **Alignment** (Separate Spaces) Vs. **Fusion** (One Space)
    2. **Encoder/Decoders types** (Transformers / Diffusion Autoencoders / dVAE / etc…)
    3. **Pre-training strategy** and **Learning Objective**

References

- *Akkus, C., Chu, L., Djakovic, V., Koch, P., Loss, G., Marquardt, C., Moldovan, M., Sauter, N., Schneider, M., Schulte, R., Urbanczyk, K., Goschenhofer, J., Heumann, C., Hvingelby, R., Schalk, D., & Aßenmacher, M. (2023).* ***Multimodal Deep Learning****. ArXiv. /abs/2301.04856*

- *Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., & Sutskever, I. (2021).* ***Learning Transferable Visual Models From Natural Language Supervision.*** *ArXiv. /abs/2103.00020*

- *Lu, J., Batra, D., Parikh, D., & Lee, S. (2019).* ***ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks.*** *ArXiv. /abs/1908.02265*

Thanks!

Now dive deeper into the CLIP model and its applications… with Jan.

Multimodal Embeddings: CLIP

Omar & Jan

Image space is vast

- https://youtu.be/Dt2WYkqZfbs (Steve Brunton): 20x20 1 bit image is larger than the Universe: Shader Demo: https://www.shadertoy.com/view/Dty3Ww
- real world: 224x224 images with 256x3 pixels: 1.2 * 10^1372
- Every image can be described with a text description (any one disagrees?)
- images can be graphics, drawings, photos. Those include the natural world and aren’t inherently human.

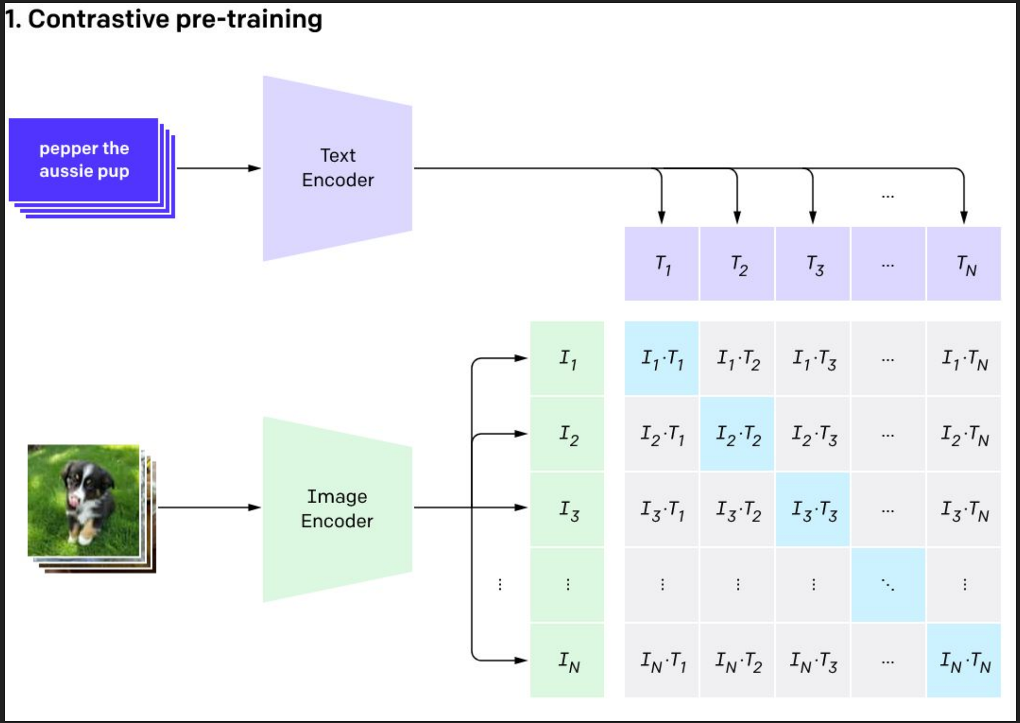
Language is vast too

- if you give a million monkeys a million typewriters they would eventually come up with the complete works of Shakespeare: Empirical Evidence: https://web.archive.org/web/20090318143423/http://www.vivaria.net/experiments/notes/publication/NOTES_EN.pdf
- Who polices the Police? The (Police)+ .
- What do you build, if your neighbor has missiles? (anti)^n (missiles)^(n+1).
- context length (76?) ^ vocab size 49407 ~ 2.3* 10^92925
- Can every piece of text be conveyed by an image? (I think no)
- Language is completely human!

The idea behind CLIP: Contrastive pairs

https://openai.com/research/clip

- contrastive learning (2 weeks)
- N-Pair with batch size 32,768
- text model has ~ 63M
- vision model has ~ 87M
- Cosine similarity -> dot product
- approach scales

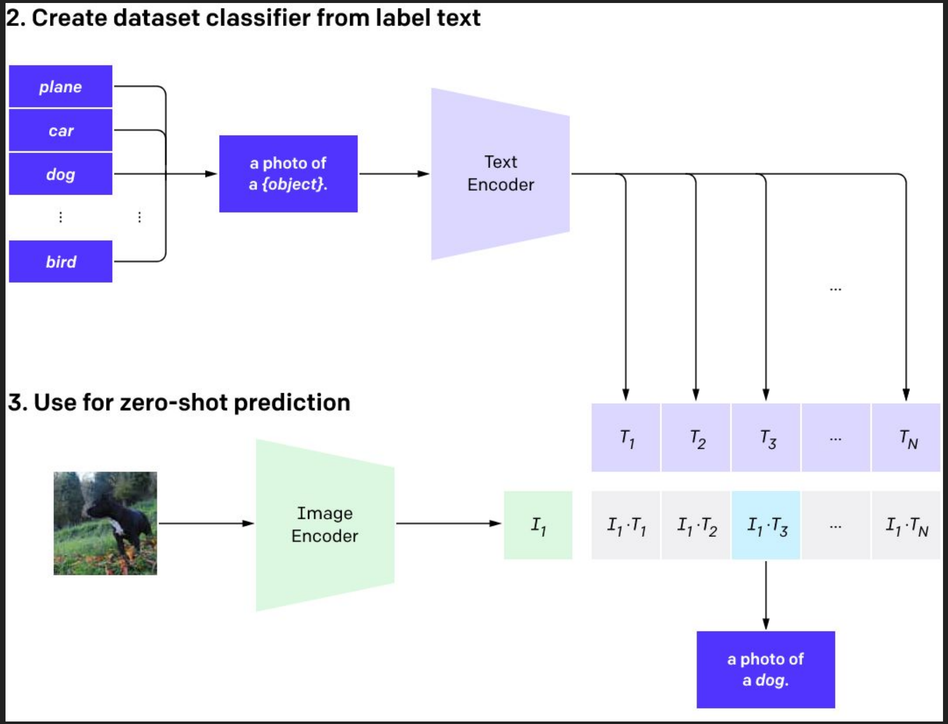


The Dataset

- from the actual CLIP paper https://arxiv.org/pdf/2103.00020.pdf: **400M** pairs from image search, words from wikipedia 100+
- LAION**5B**: exploration -> https://rom1504.github.io/clip-retrieval/
- https://laion.ai/blog/laion-5b/ :"The images are under their copyright."
- Various filters for watermark, aesthetic score, NSFW, language, duplicates

Main task for CLIP: zero-shot image classification

- supervised: distinct labels, one domain
- CLIP “natural language supervision”
- 1 model for any domain
- Construct your classifier labels with prompts
- Is robust



Research example: visual Word-Sense-Disambiguation

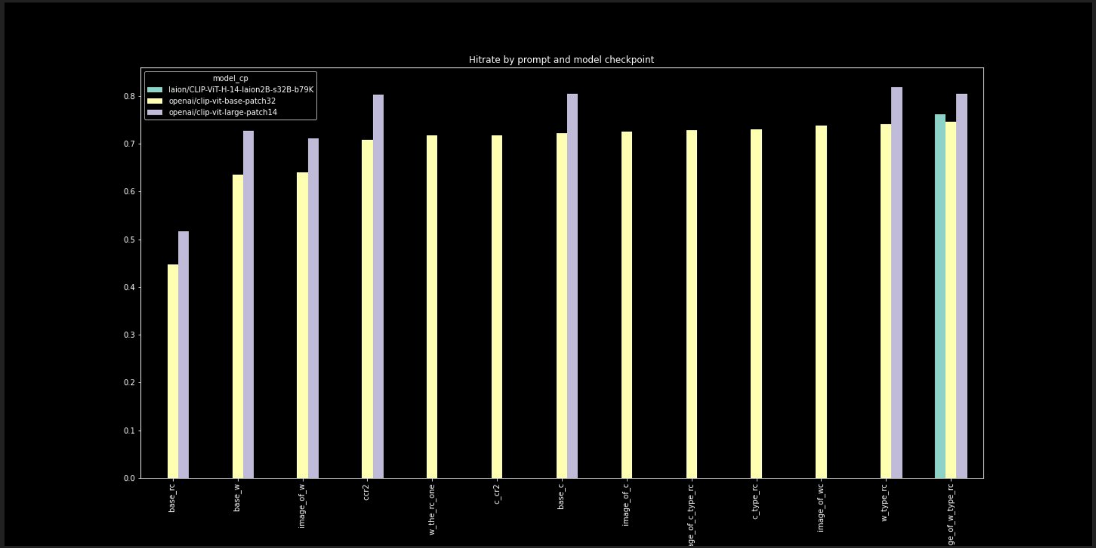
- shared task: https://raganato.github.io/vwsd/
- prompt engineering, ensemble, semantic modelling?
- my results:

  larger models do better

  ccr2 (repeat words?)

  nearly 80% on train set

  50% on test (baseline)



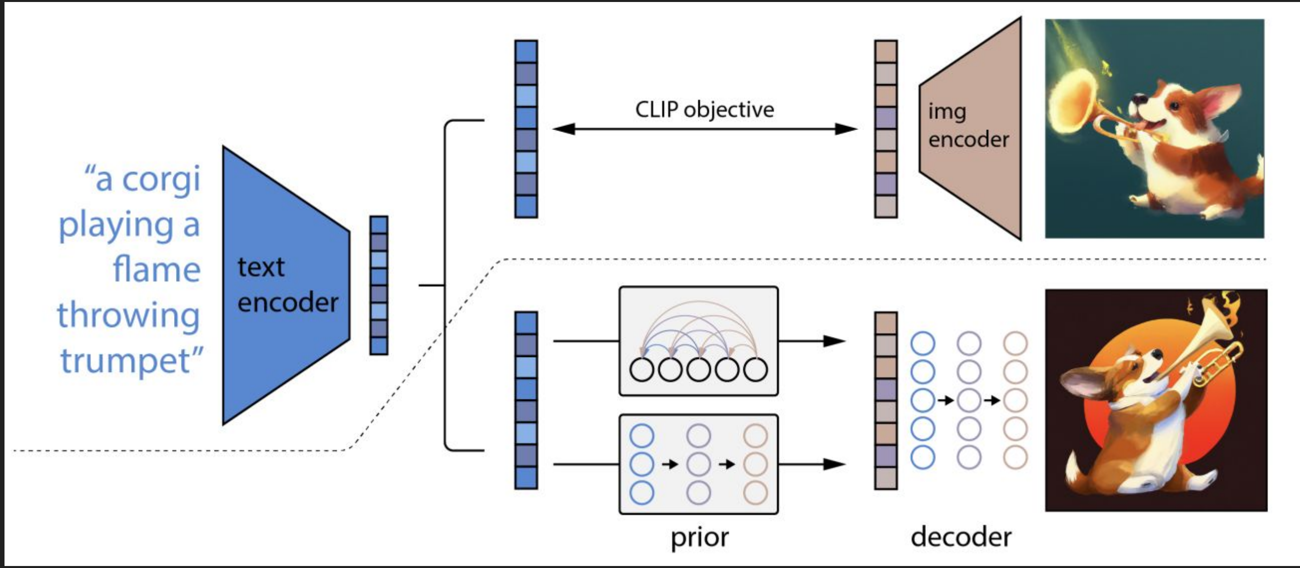
TNX, qs?

BONUS SLIDES!

A. Text to image generation
B. Image Saliency maps
C. Shader Match? (my own research, WIP)

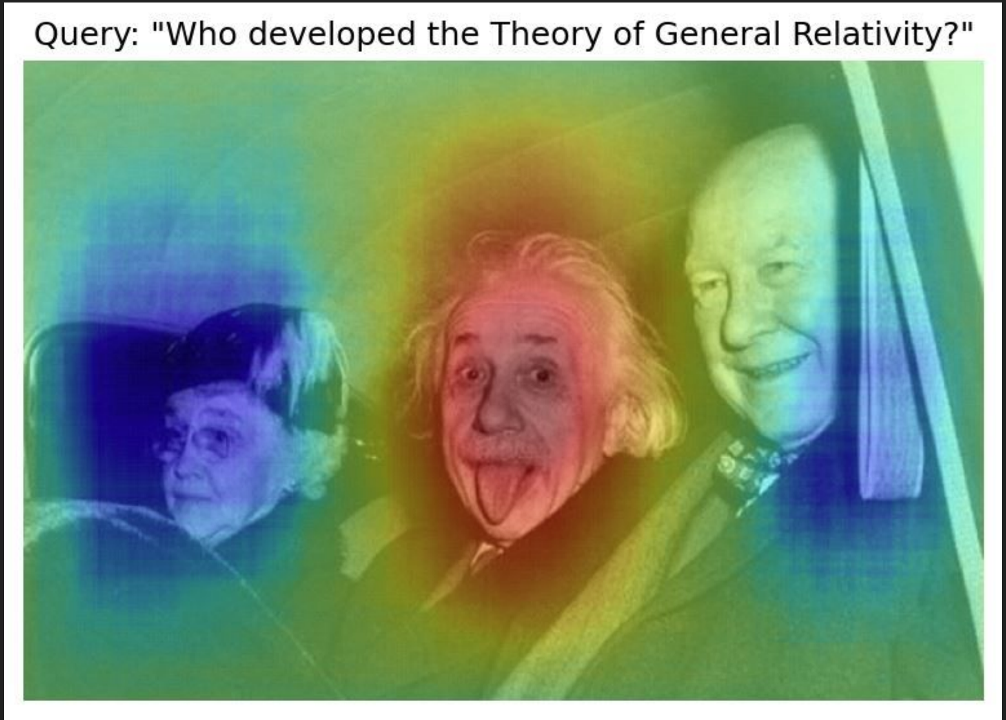
BONUS SLIDE: A - text2img

- Precursors: ImageGPT, Dall-E
- We need to decode the embedding into pixels for an image: Diffusion
- “unCLIP”
- StableDiffusion
- Midjourney
- ControlNet
- text2nerf/3D
- text2video



BONUS SLIDE B - Image Saliency Maps

- Research background might be astronomy?
- Random patches/crops
- Any ideas what to use this for?
- There is some way to do this with cross attention, to see which part of an image attend to each word (not directly clip related)



BONUS SLIDE C - ShaderEval task X?

- Term paper turned hobby project (maybe Bachelor Thesis)
- My own dataset of title, description, code, frame sequences(any length)
- “text2img” via shader code, kinda
- Anyone interested for work with me on this?

BYE!