Machine Learning for **Language** and Vision

Seminar SS23
Introduction, Apr 26, 2023

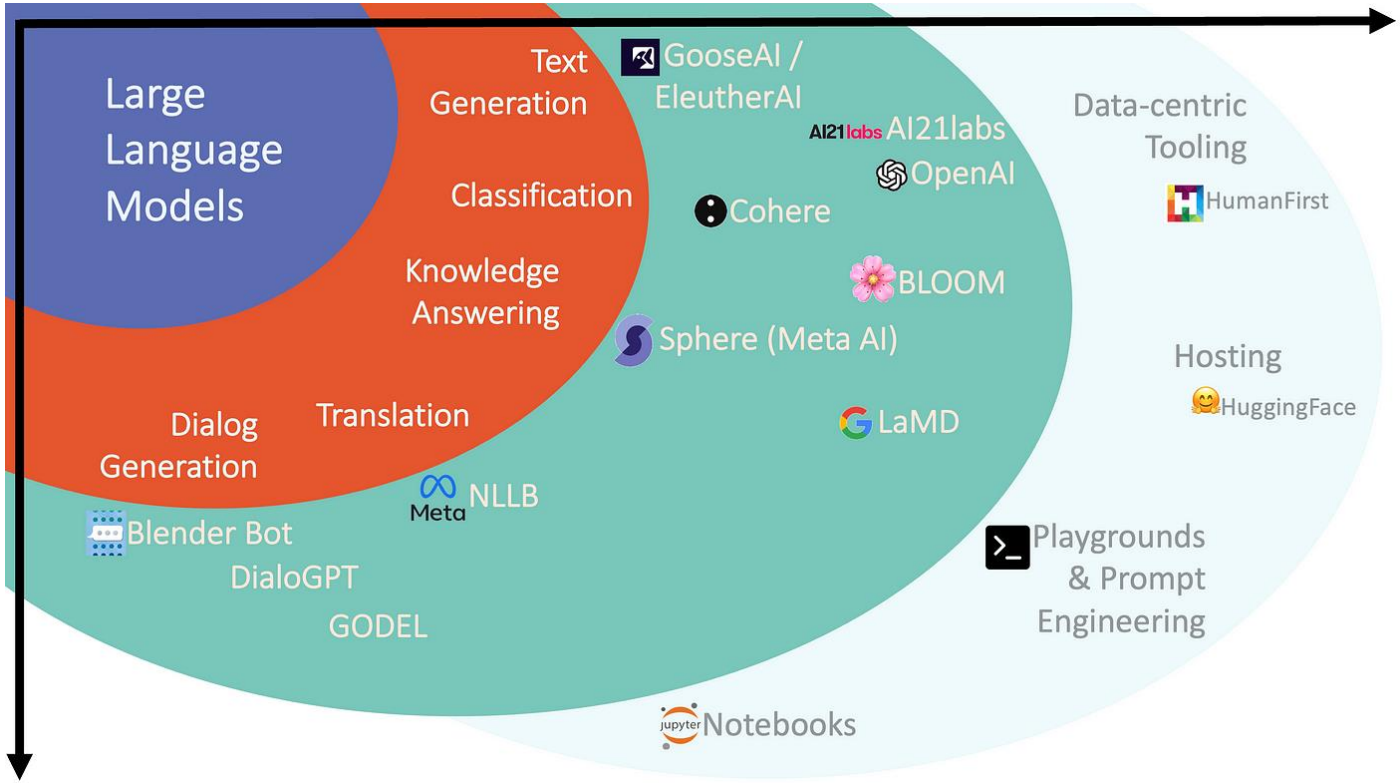
Xudong Hong, Ruitao Feng

Saarland University

Plan for today

1. Introduction
2. Paper Assignment

Large Language Models



Large Language Models

Can large language models understand meaning?

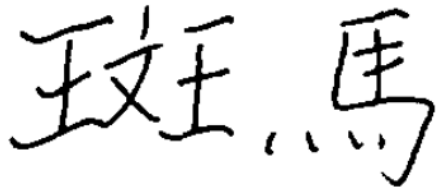
Large Language Models

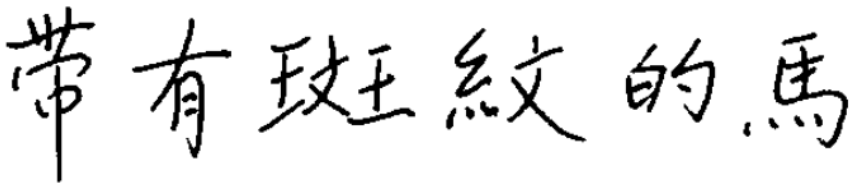
Can large language models understand meaning?

斑馬　　帶有斑紋的馬

Large Language Models

Can large language models understand meaning?





Symbol grounding problem (Harnad et al., 1990)

- E.g. To learn Chinese with a Chinese/Chinese dictionary

Grounded Language

Two friends traveling through rough country who were suddenly confronted by a bear. One saved himself by scrambling up a tree while the other who couldn't climb asked for help but got nothing. So he threw himself on the ground and pretended to be dead. The animal came close and sniffed him over but then left. Then the man in the tree asked what the bear had been saying to him. His friend said "he told me never to trust someone who deserts you in need."

The Bear and the Travelers — Aesop,

painting from Milo Winter 1919

Grounded Language

The Bear and the Travelers — Aesop,

painting from Milo Winter 1919

Grounded Language

friend?

The Bear and the Travelers — Aesop,

painting from Milo Winter 1919

Visual Perception

One Look Is Worth A Thousand Words--

One look at our line of Republic, Firestone, Miller and United States tires can tell you more than a hundred personal letters or advertisements.

WE WILL PROVE THEIR VALUE BEFORE YOU INVEST ONE DOLLAR IN THEM.

Ever consider buying Supplies from a catalog?

What's the use! Call and see what you are buying. One look at our display of automobile and motorcycle accessories will convince you of the fact.

THAT WE HAVE EVERYTHING FOR THE AUTO

Piqua Auto Supply House

133 N. Main St.—Piqua, O.

wiki

Neural Network Image Encoder

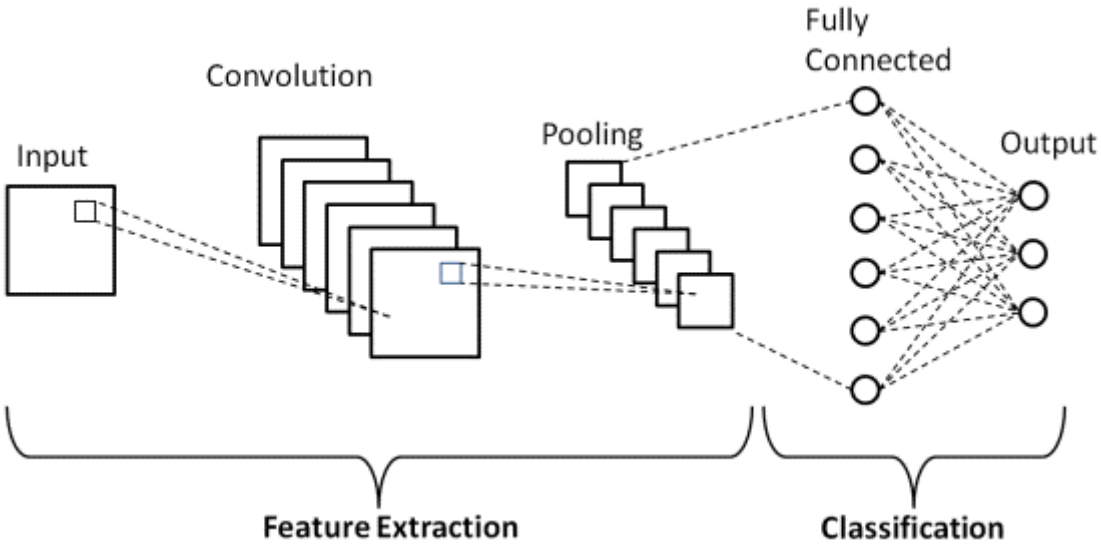


Image Encoder

Sparse feature

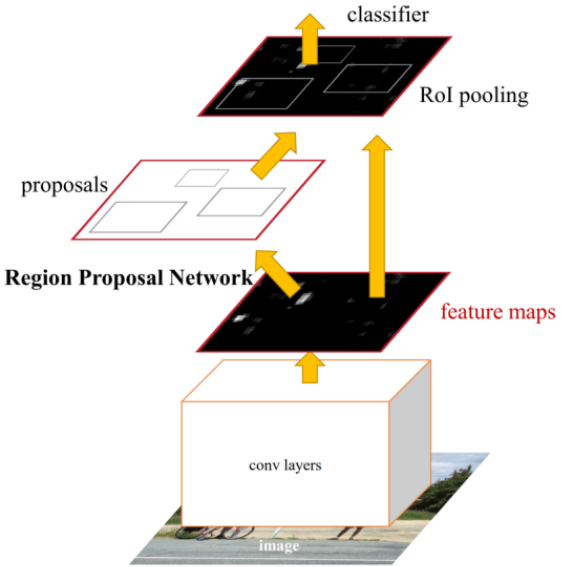
- Object detector

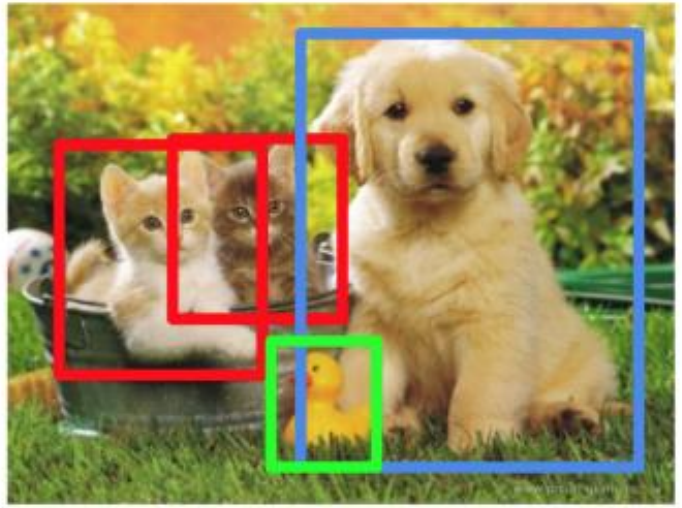
Dense feature

- Convolutional neural network (CNN), Vision transformer (ViT) …

Sparse Features - Object detector

- Faster RCNN



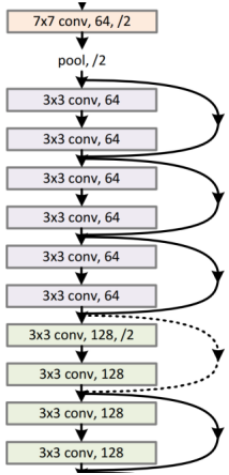


CAT, DOG, DUCK

- *Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks*, 2015
- *Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering*, 2018

Dense feature - CNN

- Convolutional neural network
  - Resnet50, Resnet101
    - Pre-trained on ImageNet



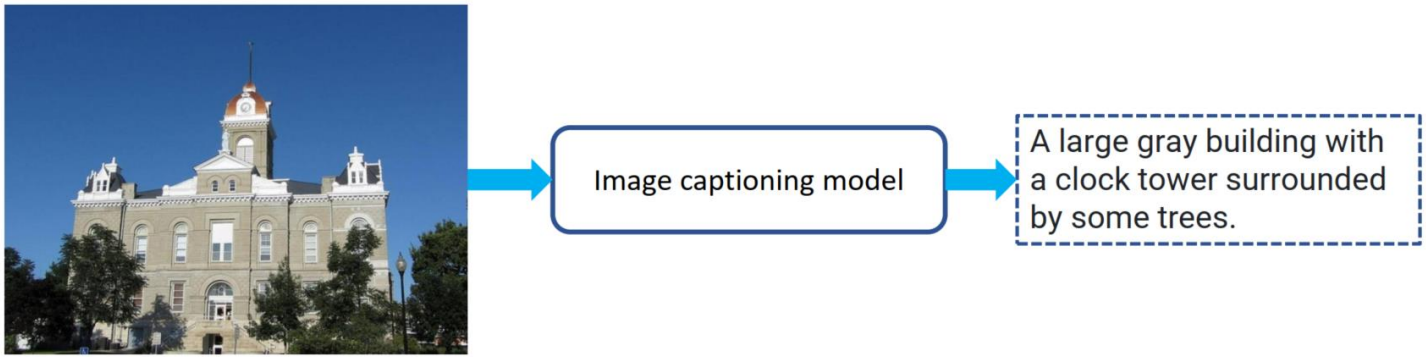
- *Deep Residual Learning for Image Recognition*, 2015

Practical Applications

- Image Captioning
- Visual Question Answering
- Visual Storytelling

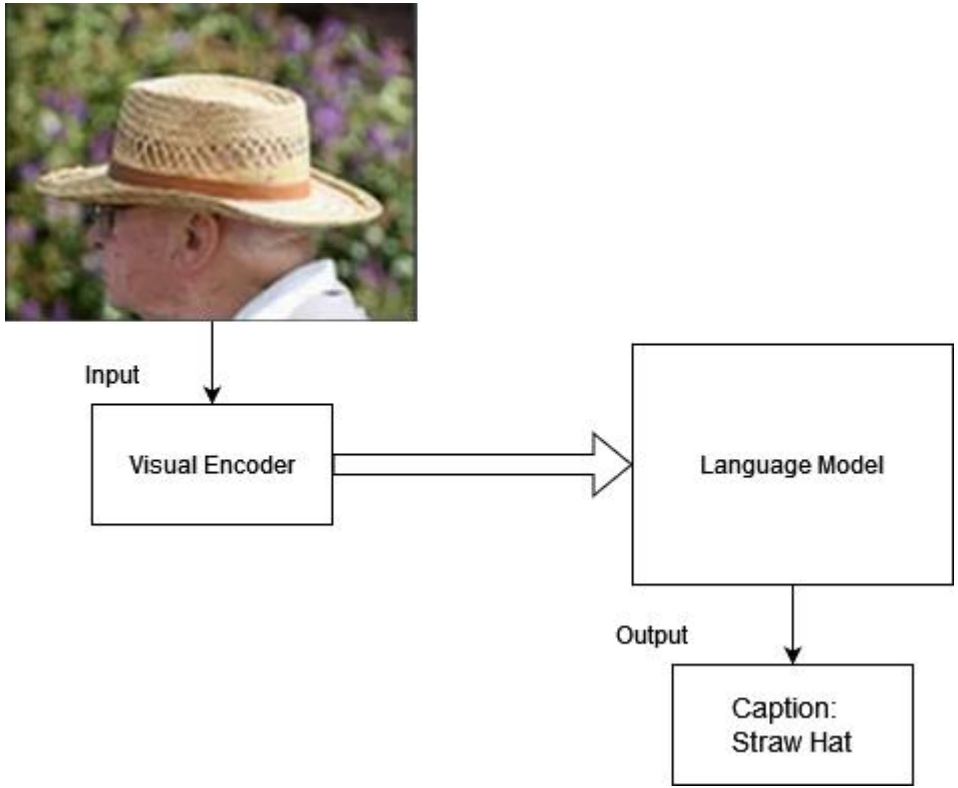
Common VL Tasks- Image Captioning

- Language generation
- Difficult automatic evaluation (BLEU, CIDEr, Rogue)



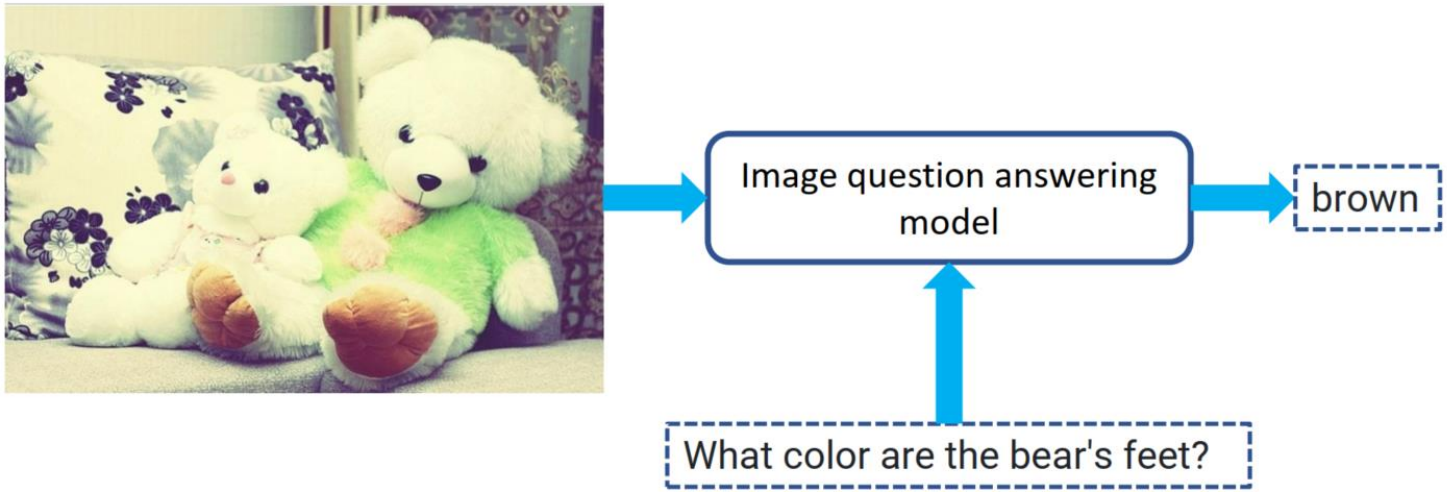
Common VL Tasks- Image Captioning

- Input: image
- Output: caption



Common VL Tasks- Visual Question Answering

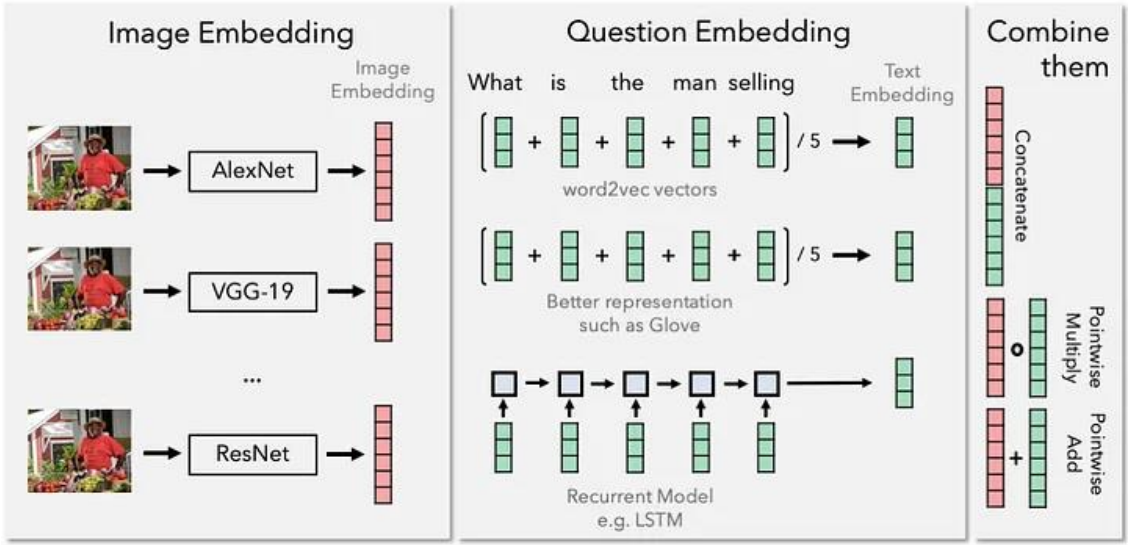
- Elicit specific information from images
- Relatively easier evaluation (accuracy using string matching)



Making the V in VQA Matter: Elevating the Role of Image Understanding in Visual Question Answering, 2016

Common VL Tasks- Visual Question Answering

- Input: image and caption
- Output: answer



Making the V in VQA Matter: Elevating the Role of Image Understanding in Visual Question Answering, 2016

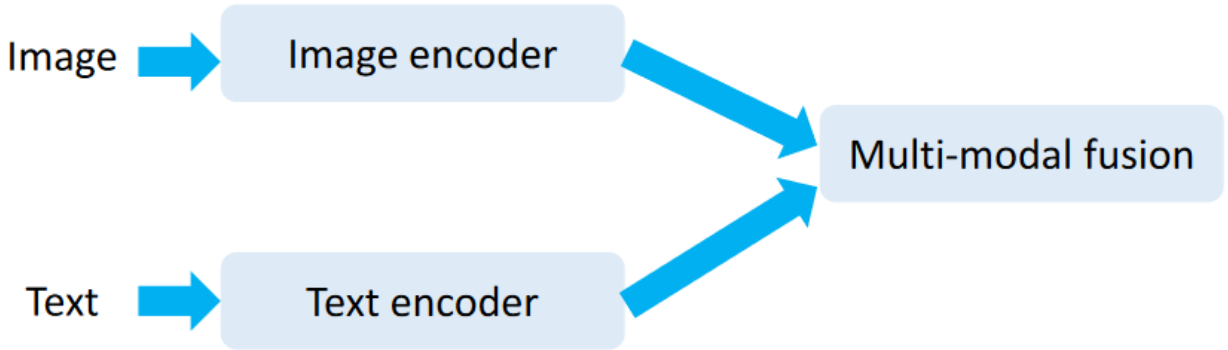
Common VL Tasks- Visual Storytelling

- Creative Language generation

Visually Grounded Story Generation

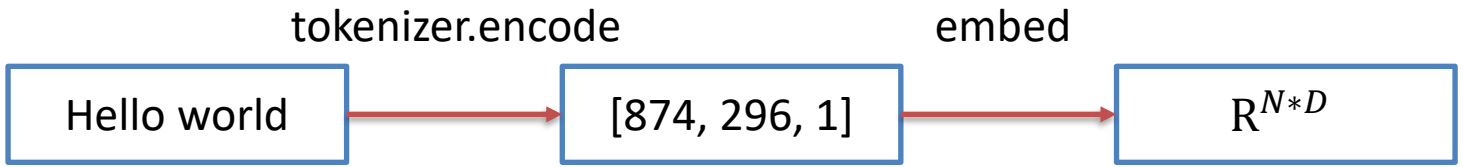
Jack was on a call with a client, getting stressed over a business deal that wasn't going well.
Jack put the phone down after an unsuccessful deal and decided to go get a coffee at the nearby coffee.
At the coffee shop, he started talking to the waiter Will about the unfortunate call.
Will told him he would convince the client to accept the deal if he could work for Jack.
Will then called the client and successfully struck the deal.…

Basic Network architecture - Pretraining Multimodal model

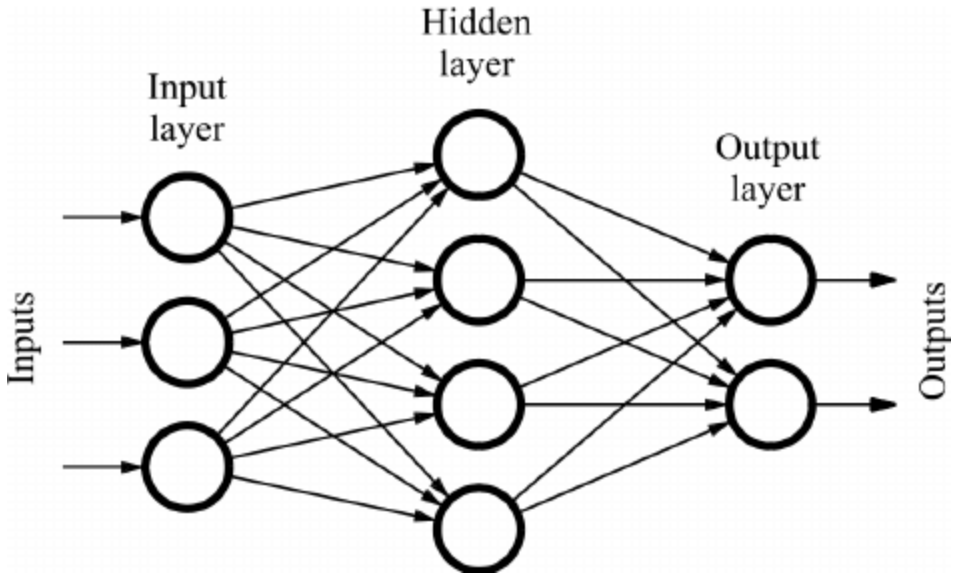


Text Encoder - Embedding

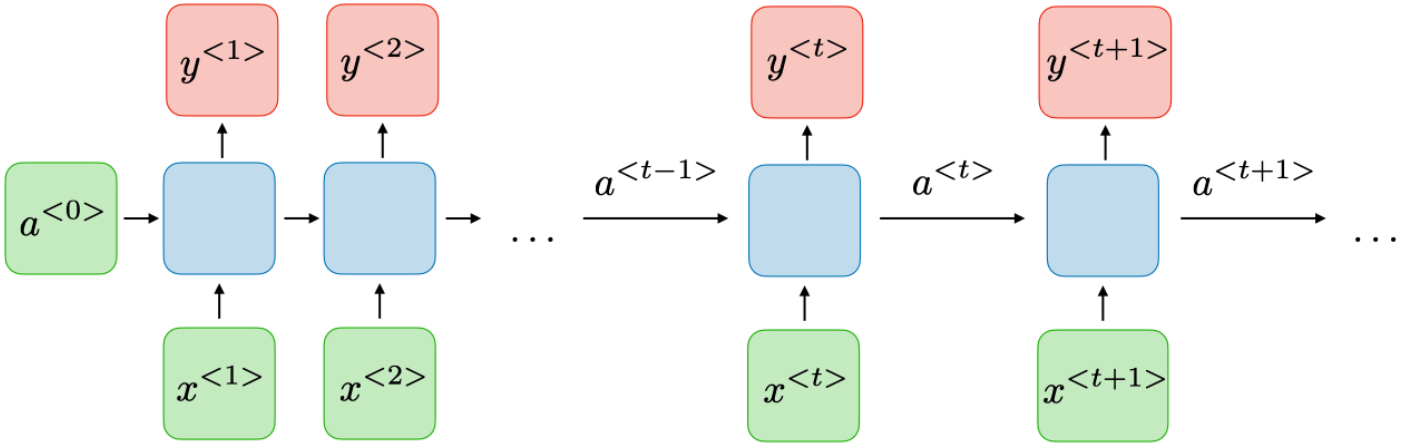
- Tokenize
  - Map the input string to the index of its tokens
  - Input: String
  - Output: $x_i \in \{0, 1, \cdots, T-1\}, i \in \{0, 1, \cdots, N\}$
    - N: number of subtokens
    - T: vocabulary size
- Embedding
  - Map the input string to the index of its tokens
  - Input: $x_i \in \{0, 1, \cdots, T-1\}$
  - Output: $y_i \in R^D, i \in \{0, 1, \cdots, N\}$
    - D: embedding dimension
    - N: number of subtokens



Text Encoder – Feed-Foward Network



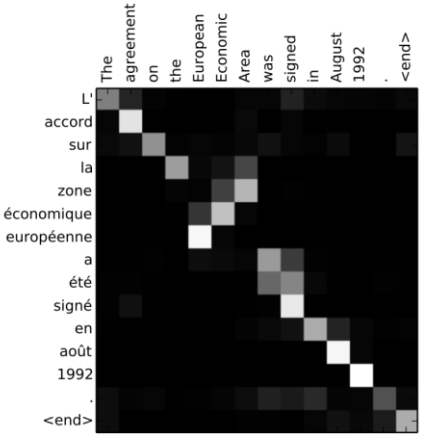
Text Encoder - RNN



Attention Mechanisms

Motivation

- Retrieval of a value v for a given query q based on a key k
- Applications on image captioning (Xu et al., 2015), machine translation (Bahdanau et al., 2015)



General form:

$$Attention(q, k, v) = \sum sim(q, k_i)$$

- additive: $sim(q, k) = w_q^T \cdot q + w_k^T \cdot k$
- dot: $sim(q, k) = q^T \cdot k$
- self: scaled dot attention when $q = k$

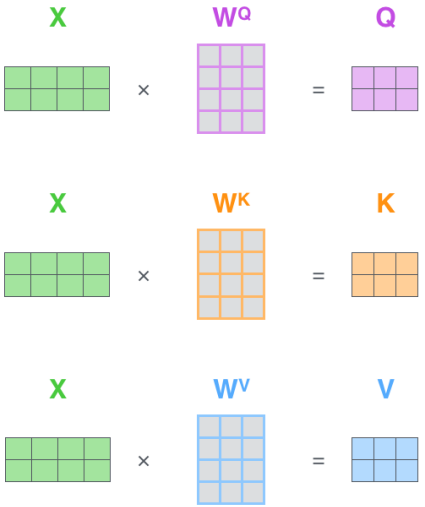
Benefits

- $O(1)$ operations to draw global dependencies between two positions
- Easy to interpret

Self-attention Mechanism

Matrix form

- Each row in X is a symbol embedding in a sequence



Benefits

- Can compute in parallel
- No additional parameters

Normalisation

- Softmax over a sequence
- $Attention(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = softmax(\mathbf{Q} \cdot \mathbf{K}^T)\mathbf{V}$

Layer Type	Complexity per Layer	Sequential Operations	Maximum Path Length
Self-Attention	$O(n^2 \cdot d)$	$O(1)$	$O(1)$
Recurrent	$O(n \cdot d^2)$	$O(n)$	$O(n)$
Convolutional	$O(k \cdot n \cdot d^2)$	$O(1)$	$O(log_k(n))$
Self-Attention (restricted)	$O(r \cdot n \cdot d)$	$O(1)$	$O(n/r)$

Transformer

Paper [1]

Vaswani, Ashish, et al.

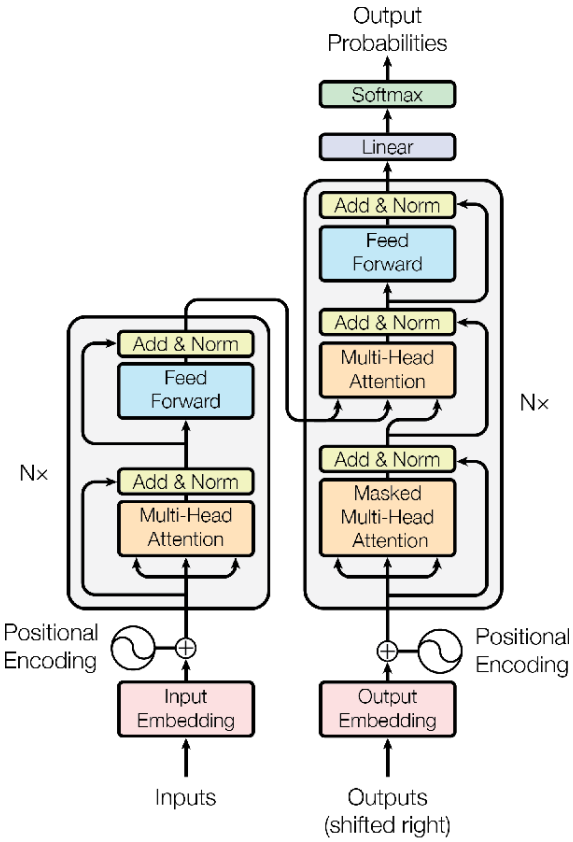
“Attention is all you need.", NeuriPS 2017

Architecture

- Left: encoder
- Right: decoder

Key Components

- Positional Encoding
- Stacked self-attention
- Point-wise, fully connected layers
- Residual connections, layer norm.

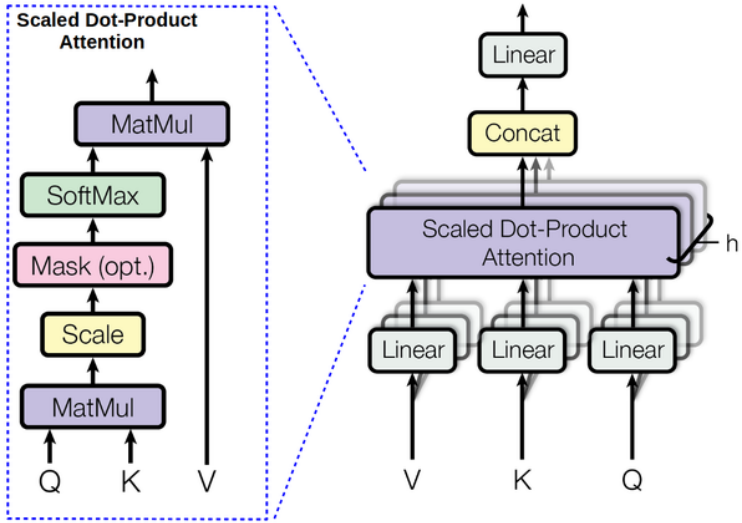


Multi-Head Attention

Scaled Dot-Product Attention

$$Attention(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = softmax(\frac{\mathbf{Q}, \mathbf{K}^T}{\sqrt{d_k}})\mathbf{V}$$

- Dot products grow large in magnitude, pushing the softmax function into regions where it has extremely small gradients
- scaled by $\sqrt{d_k}$



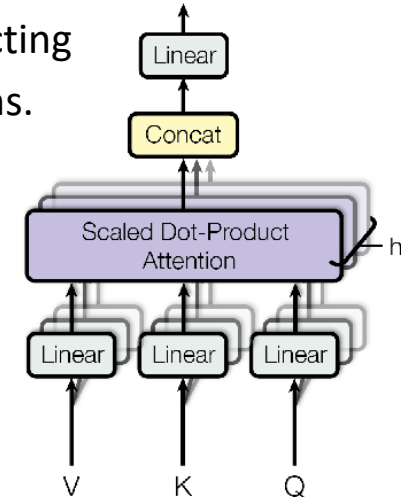
Multi-Head Attention

Multi-Head Attention

$$MultiHead(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = Concat(head_1, head_2, \ldots, head_h)W^O$$

where $head_1 = Attention(QW_i^Q, KW_i^K, VW_i^V)$

- Create different attention heads by linearly projecting Q, K, and V h-times to d_k, d_k and d_v dimensions.
- different heads capture different groups of global dependencies
- $d_k = d_v = d_{model}/h = 64$



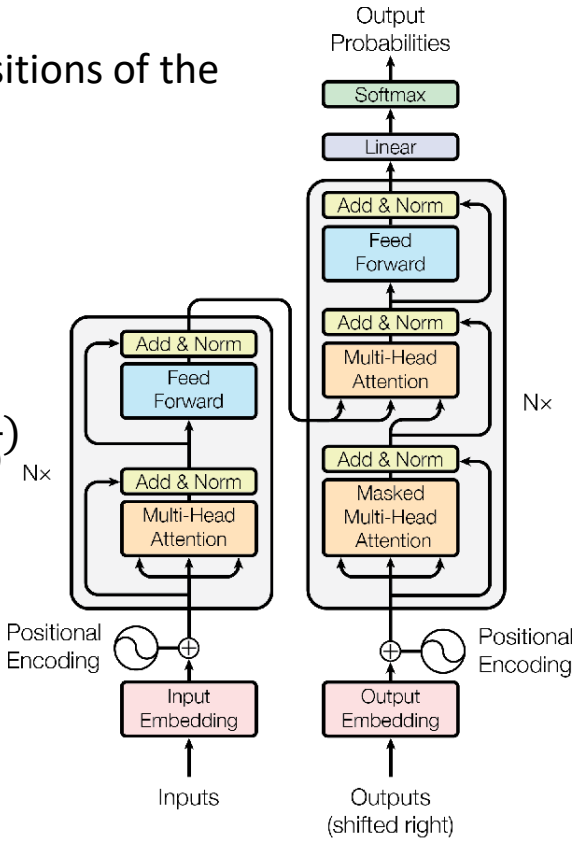
Positional Encoding

Motivation

- Need to inject information about the positions of the tokens in the sequence

Fomula

- $PE(pos, 2i) = sin(\frac{pos}{100000^{(2i/d_{model})}})$
- $PE(pos, 2i+1) = cos(\frac{pos}{100000^{(2i/d_{model})}})$
- i is the index of dimension
- pos is the position of the symbol



Experiment and Result

Sequence to sequence learning

- Machine Translation
- Data: WMT 2014 English-German, English-French

Result

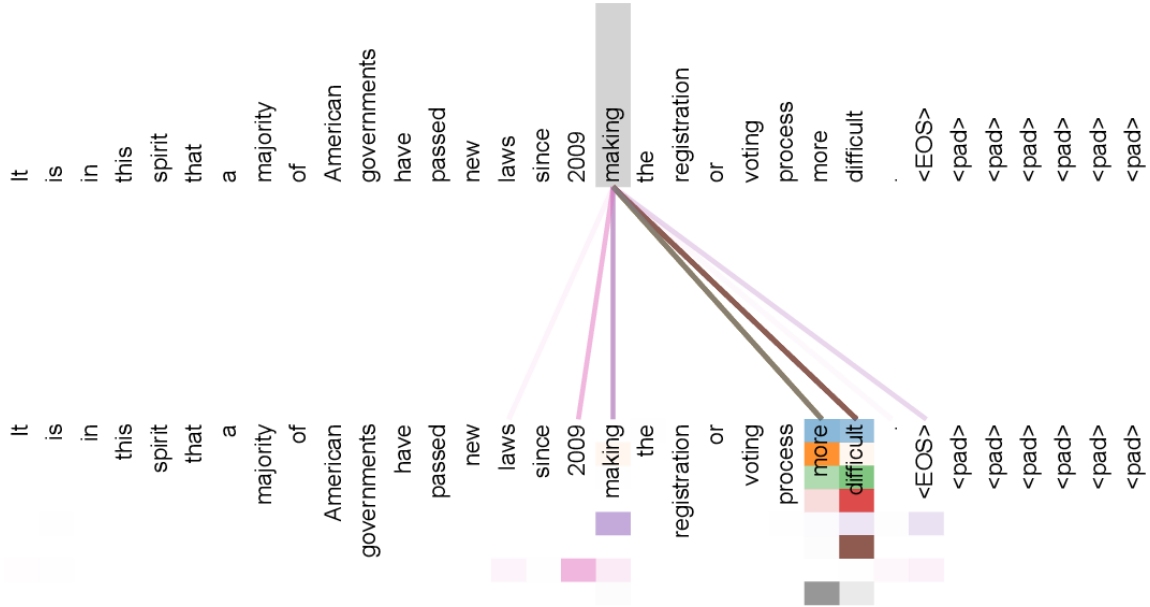
- Metric: BLEU

Model	BLEU		Training Cost (FLOPs)	
	EN-DE	EN-FR	EN-DE	EN-FR
ByteNet [18]	23.75			
Deep-Att + PosUnk [39]		39.2		$1.0 \cdot 10^{20}$
GNMT + RL [38]	24.6	39.92	$2.3 \cdot 10^{19}$	$1.4 \cdot 10^{20}$
ConvS2S [9]	25.16	40.46	$9.6 \cdot 10^{18}$	$1.5 \cdot 10^{20}$
MoE [32]	26.03	40.56	$2.0 \cdot 10^{19}$	$1.2 \cdot 10^{20}$
Deep-Att + PosUnk Ensemble [39]		40.4		$8.0 \cdot 10^{20}$
GNMT + RL Ensemble [38]	26.30	41.16	$1.8 \cdot 10^{20}$	$1.1 \cdot 10^{21}$
ConvS2S Ensemble [9]	26.36	**41.29**	$7.7 \cdot 10^{19}$	$1.2 \cdot 10^{21}$
Transformer (base model)	27.3	38.1	**$3.3 \cdot 10^{18}$**	
Transformer (big)	**28.4**	**41.8**	$2.3 \cdot 10^{19}$	

Experiment and Result

Visualisation

- 8 attention heads
- Syntactic dependency: *make … more difficult*



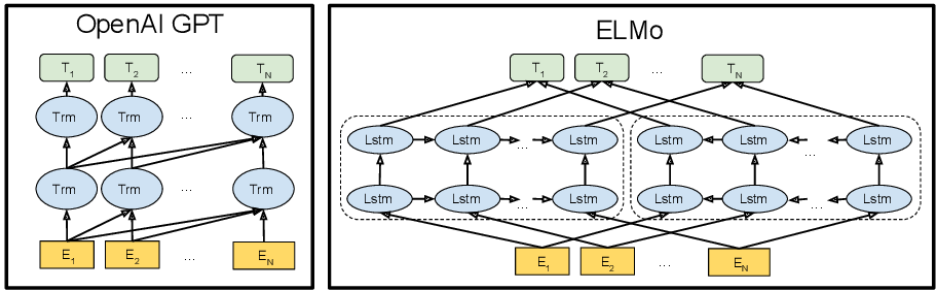
Representation Learning

Word embedding

- word2vec / Glove

Contextual Representations

- Semi-Supervised Sequence Learning, 2015
- ELMo: Deep Contextual Word Embeddings (Peters et al., 2018)
  - pre-train Bi-LSTM and concat hidden weights as embedding
- GPT: Improving Language Understanding by Generative Pre-Training, (Peters et al., 2018)
  - pre-train Transformer on auto-regressive language modeling



Masked Language Modeling

Auto-regressive

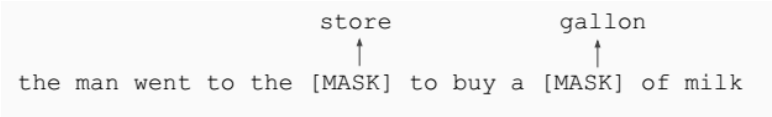
- Only train from one end to another end
- Can’t have context from both side for target

Paper [2]

- Devlin, Jacob, et al.

"Bert: Pre-training of deep bidirectional transformers for language understanding.", NAACL 2019

Masked LM: Mask out k% of the input words, and then predict the masked words (k=15)



- Too little masking: Too expensive to train
- Too much masking: Not enough context

Masked Language Modeling

Problem

- Mask token never seen at fine-tuning

Solution:

- 80% of the time, replace with [MASK]
  went to the store → went to the [MASK]
- 10% of the time, replace random word
  went to the store → went to the running
- 10% of the time, keep same
  went to the store → went to the store

Next Sentence Prediction

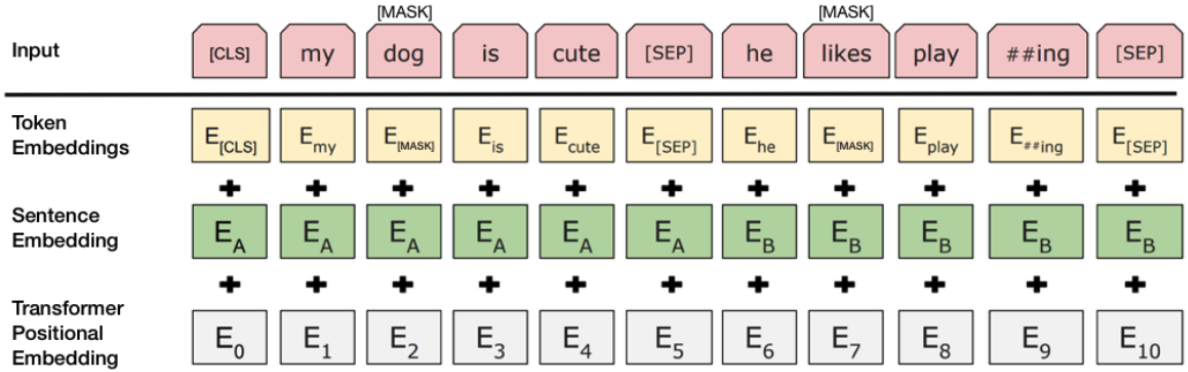
Additional objective function

To learn relationships between sentences, predict whether Sentence B is actual sentence that proceeds Sentence A, or a random sentence

Sentence A = `The man went to the store.`
Sentence B = `He bought a gallon of milk.`
Label = `IsNextSentence`

Sentence A = `The man went to the store.`
Sentence B = `Penguins are flightless.`
Label = `NotNextSentence`

Embeddings



Multimodal Fusion - Input Concatenation

- Self-attention
- modality-unaware

Figure 2: Model architecture for pre-training. The input comprises of image input, sentence input, and three special tokens ([CLS], [SEP], [STOP]). The image is processed as N Region of Interests (RoIs) and region features are extracted according to Eq. 1. The sentence is tokenized and masked with [MASK] tokens for the later masked language modeling task. Our Unified Encoder-Decoder consists of 12 layers of Transformer blocks, each having a masked self-attention layer and feed-forward module, where the self-attention mask controls what input context the prediction conditions on. We implemented two self-attention masks depending on whether the objective is bidirectional or seq2seq. Better viewed in color.

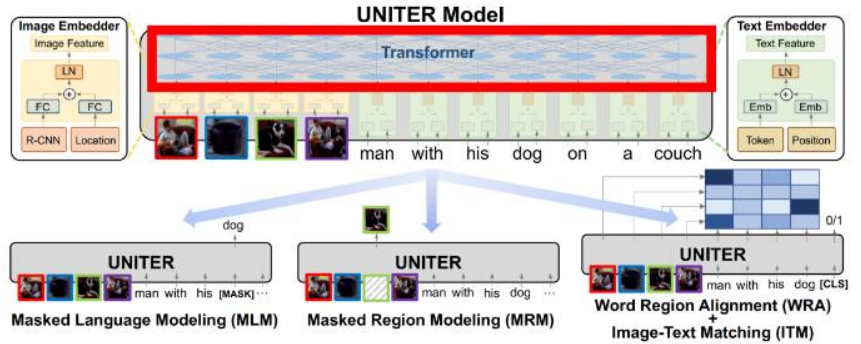


Fig. 1: Overview of the proposed UNITER model (best viewed in color), consisting of an Image Embedder, a Text Embedder and a multi-layer Transformer, learned through four pre-training tasks

Multimodal Fusion - Cross-attention

- self-attention
- modality-aware

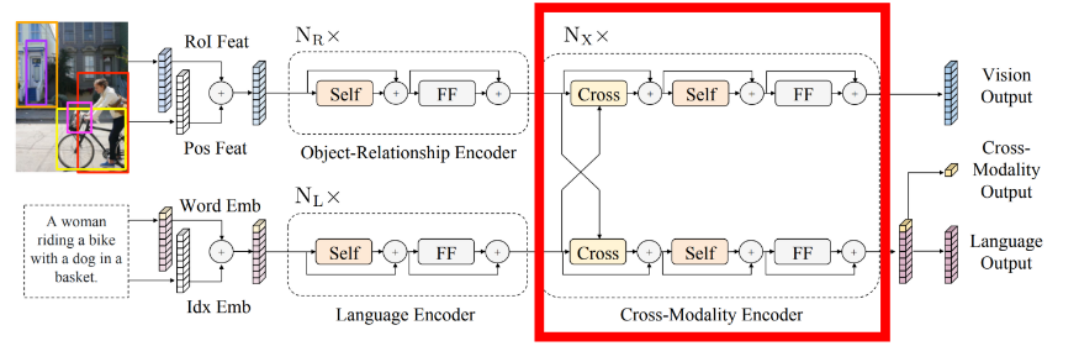


Figure 1: The LXMERT model for learning vision-and-language cross-modality representations. 'Self' and 'Cross' are abbreviations for self-attention sub-layers and cross-attention sub-layers, respectively. 'FF' denotes a feed-forward sub-layer.

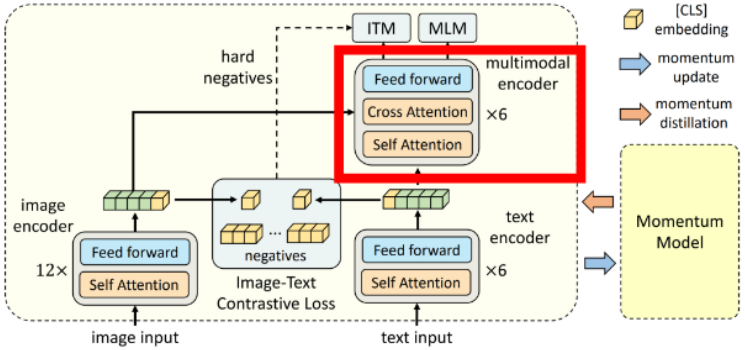
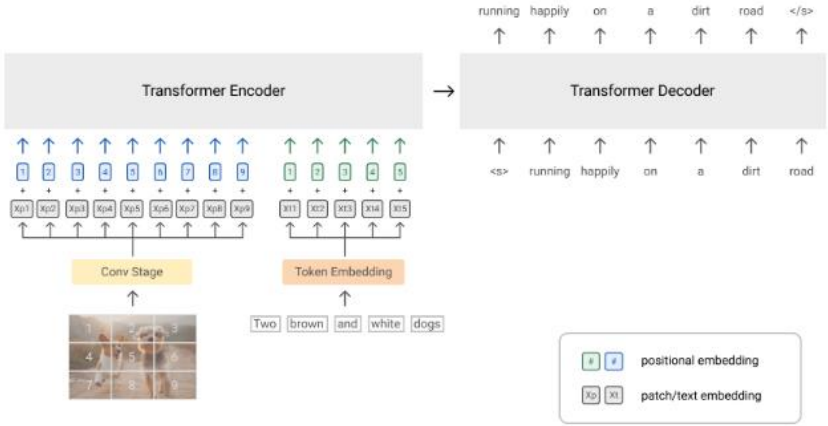


Figure 1: **Illustration of ALBEF.** It consists of an image encoder, a text encoder, and a multimodal encoder. We propose an image-text contrastive loss to align the unimodal representations of an image-text pair before fusion. An image-text matching loss (using in-batch hard negatives mined through contrastive similarity) and a masked-language-modeling loss are applied to learn multimodal interactions between image and text. In order to improve learning with noisy data, we generate pseudo-targets using the momentum model (a moving-average version of the base model) as additional supervision during training.

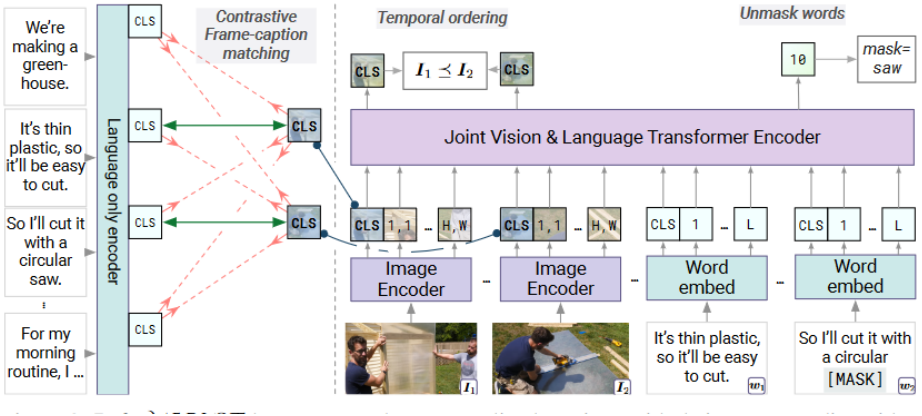
Seq2seq Pre-training - Image

- Wang, Z., Yu, J., Yu, A.W., Dai, Z., Tsvetkov, Y. and Cao, Y., SimVLM: Simple Visual Language Model Pretraining with Weak Supervision. In *International Conference on Learning Representations.*
- Singh, A., Hu, R., Goswami, V., Couairon, G., Galuba, W., Rohrbach, M. and Kiela, D., 2022. Flava: A foundational language and vision alignment model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 15638-15650).

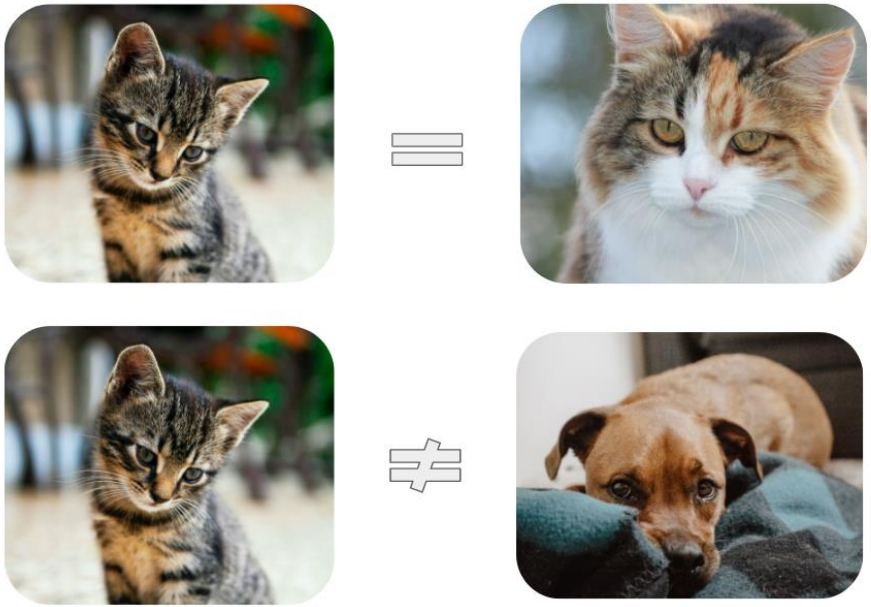


Pre-training - Video

- Zellers, R., Lu, J., Lu, X., Yu, Y., Zhao, Y., Salehi, M., Kusupati, A., Hessel, J., Farhadi, A. and Choi, Y., 2022. Merlot reserve: Neural script knowledge through vision and language and sound. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 16375-16387).
- Zellers, R., Lu, X., Hessel, J., Yu, Y., Park, J.S., Cao, J., Farhadi, A. and Choi, Y., 2021. Merlot: Multimodal neural script knowledge models. *Advances in Neural Information Processing Systems*, *34*, pp.23634-23651.

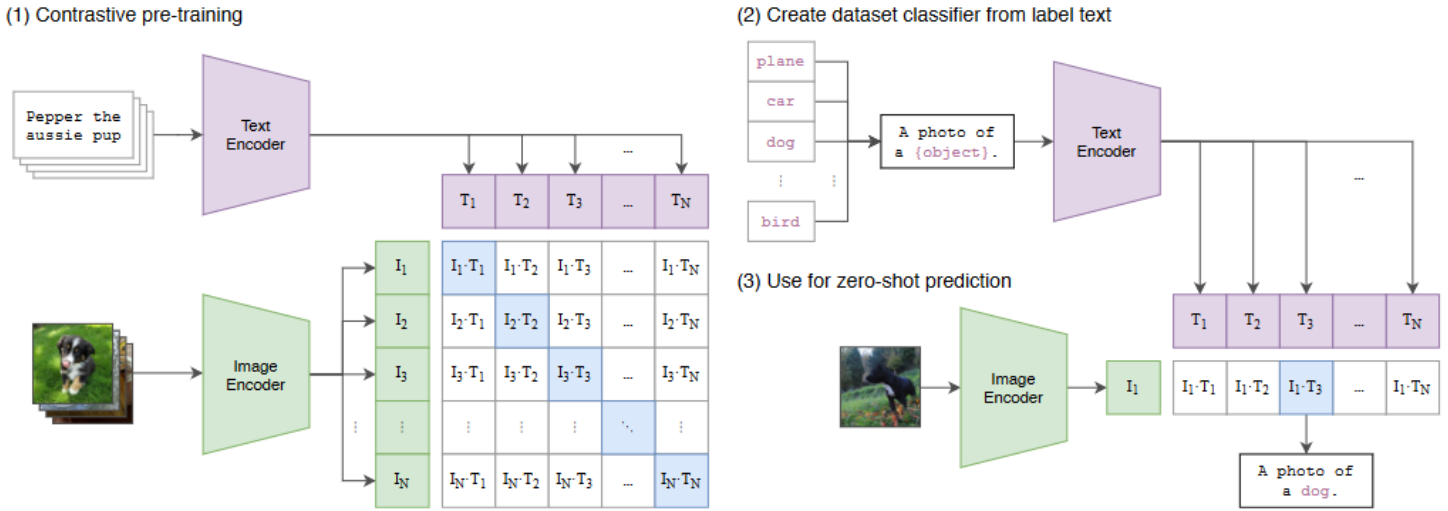


Contrastive



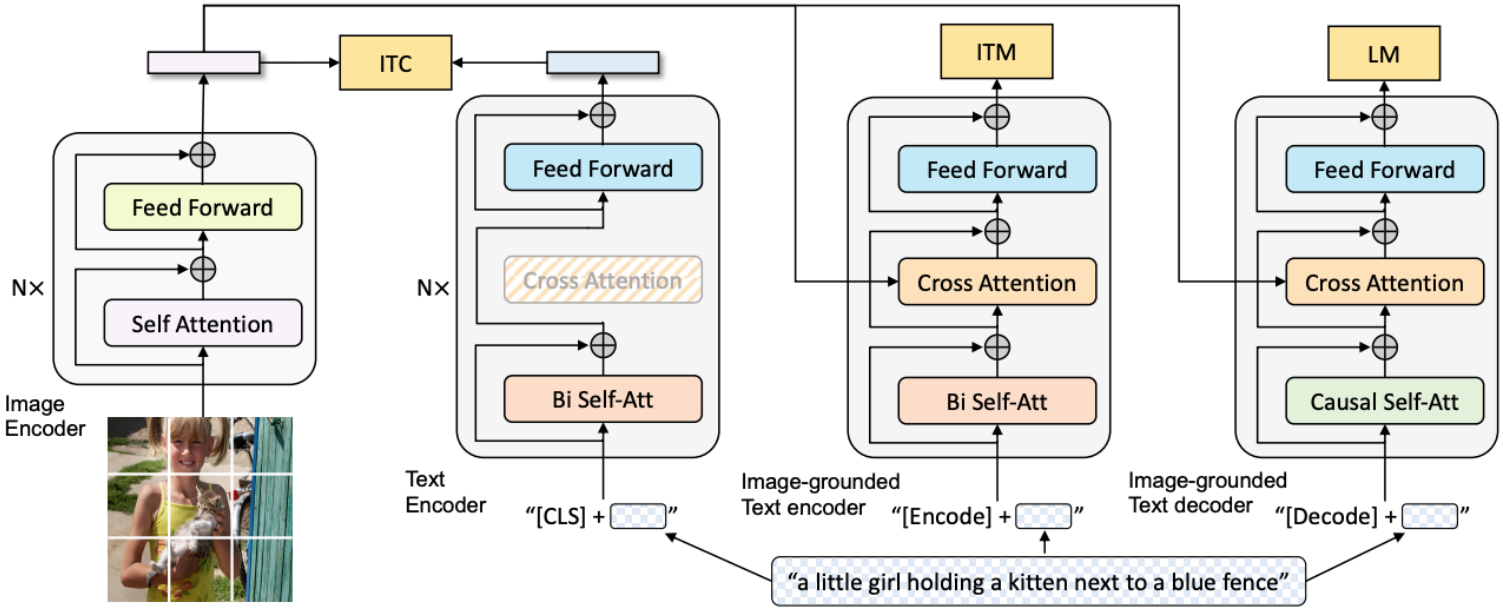
Contrastive Pre-training - Image

- Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J. and Krueger, G., 2021, July. Learning transferable visual models from natural language supervision. In International conference on machine learning (pp. 8748-8763). PMLR.



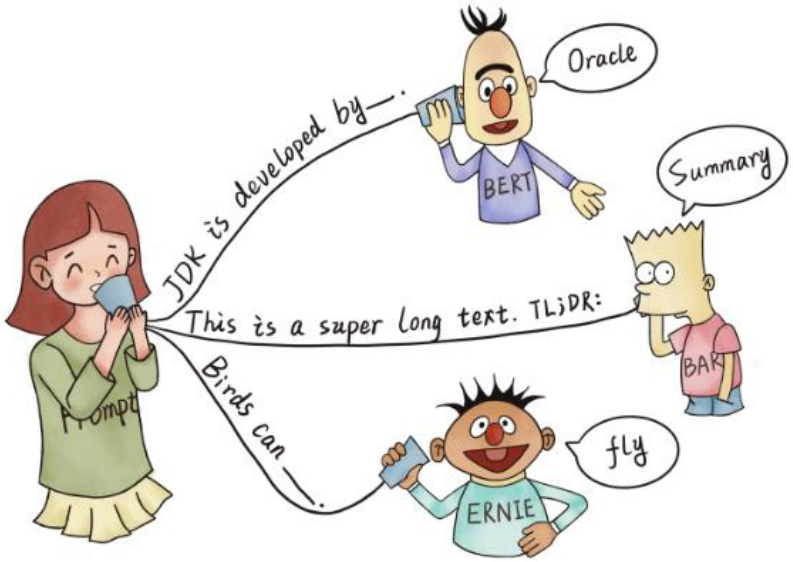
Contrastive Pre-training - Image

- Li, J., Li, D., Xiong, C. and Hoi, S., 2022, June. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning* (pp. 12888-12900). PMLR.



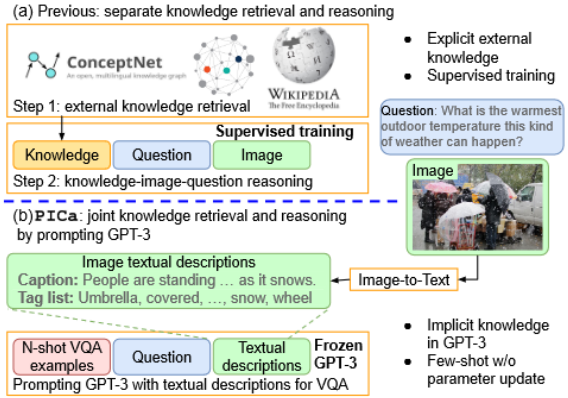
Parameter Efficiency - Prompting

- Prompting
- Prompt Tuning
- Prefix-Tuning



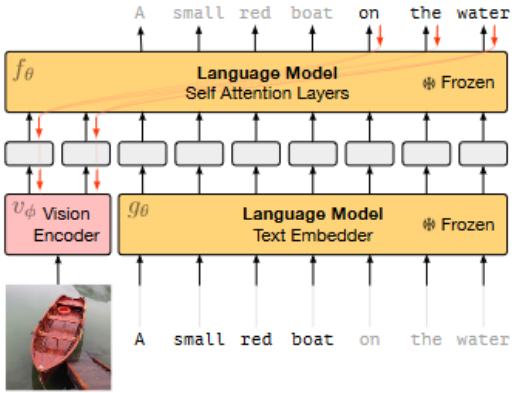
Parameter Efficiency - Prompting

- Yang, Z., Gan, Z., Wang, J., Hu, X., Lu, Y., Liu, Z. and Wang, L., 2022, June. An empirical study of gpt-3 for few-shot knowledge-based vqa. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 36, No. 3, pp. 3081-3089).
- Zeng, A., Wong, A., Welker, S., Choromanski, K., Tombari, F., Purohit, A., Ryoo, M., Sindhwani, V., Lee, J., Vanhoucke, V. and Florence, P., 2022. Socratic Models: Composing Zero-Shot Multimodal Reasoning with Language. ICLR 2023. *arXiv e-prints*, pp.arXiv-2204.



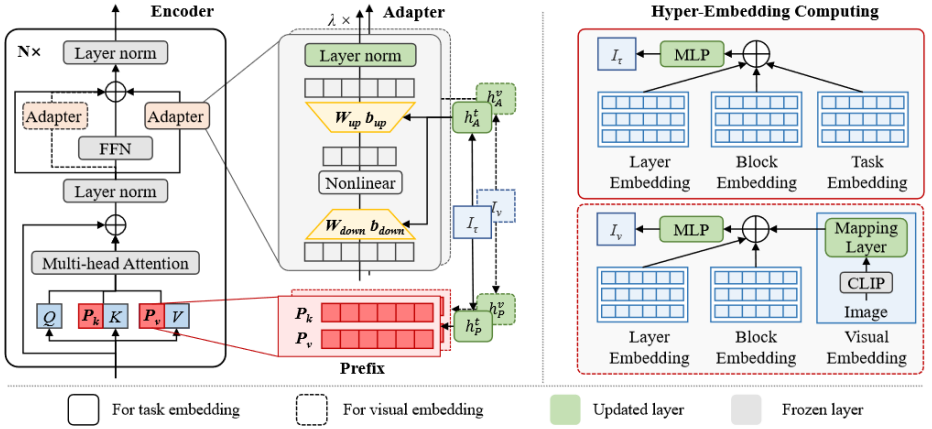
Parameter Efficiency - Prompt Tuning

- Tsimpoukelli, M., Menick, J.L., Cabi, S., Eslami, S.M., Vinyals, O. and Hill, F., 2021. Multimodal few-shot learning with frozen language models. Advances in Neural Information Processing Systems, 34, pp.200-212.
- Yu, Y., Chung, J., Yun, H., Kim, J. and Kim, G., 2021. Transitional adaptation of pretrained models for visual storytelling. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 12658-12668).



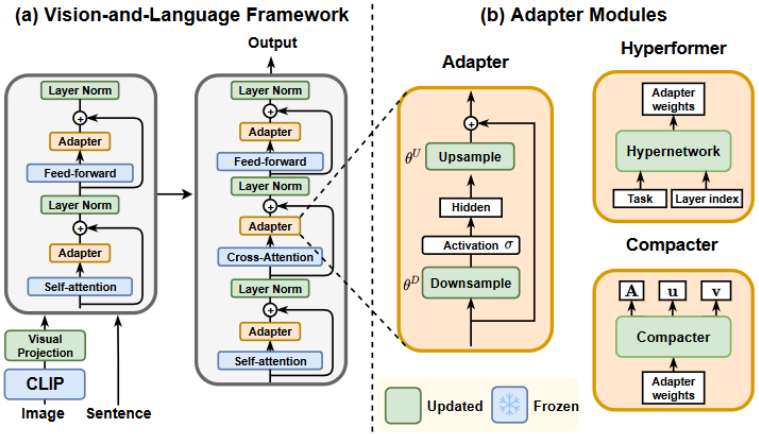
Parameter Efficiency - Prefix-Tuning

- Zhang, Z., Guo, W., Meng, X., Wang, Y., Wang, Y., Jiang, X., Liu, Q. and Yang, Z., 2022. Hyperpelt: Unified parameter-efficient language model tuning for both language and vision-and-language tasks. arXiv preprint arXiv:2203.03878.
- Jia, M., Tang, L., Chen, B.C., Cardie, C., Belongie, S., Hariharan, B. and Lim, S.N., 2022, October. Visual Prompt Tuning. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXIII* (pp. 709-727).



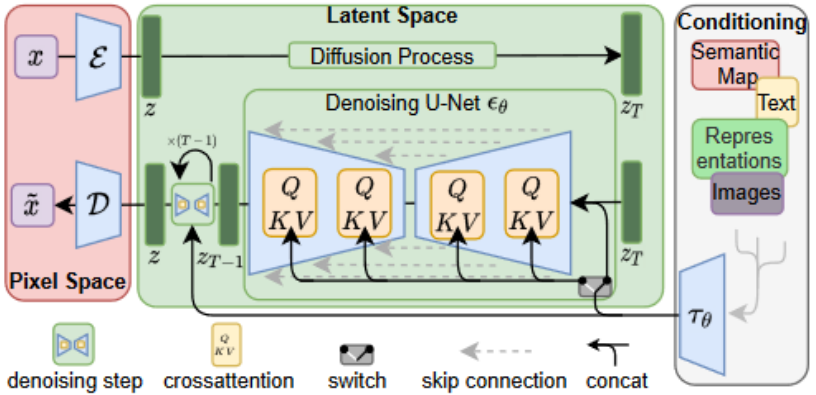
Parameter Efficiency - Adapters

- Sung, Y.L., Cho, J. and Bansal, M., 2022. VI-adapter: Parameter-efficient transfer learning for vision-and-language tasks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 5227-5237).
- Sung, Y.L., Cho, J. and Bansal, M., 2022. LST: Ladder Side-Tuning for Parameter and Memory Efficient Transfer Learning. In *Advances in Neural Information Processing Systems 2022*.



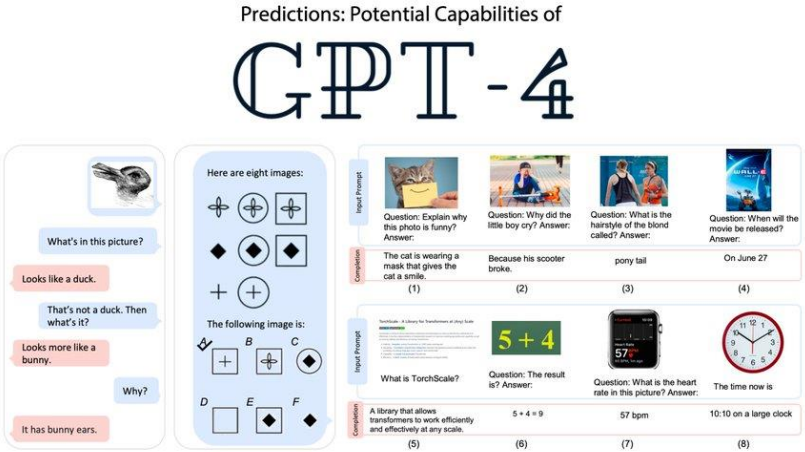
Generative Model - Text-to-Image

- Rombach, R., Blattmann, A., Lorenz, D., Esser, P. and Ommer, B., 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 10684-10695).



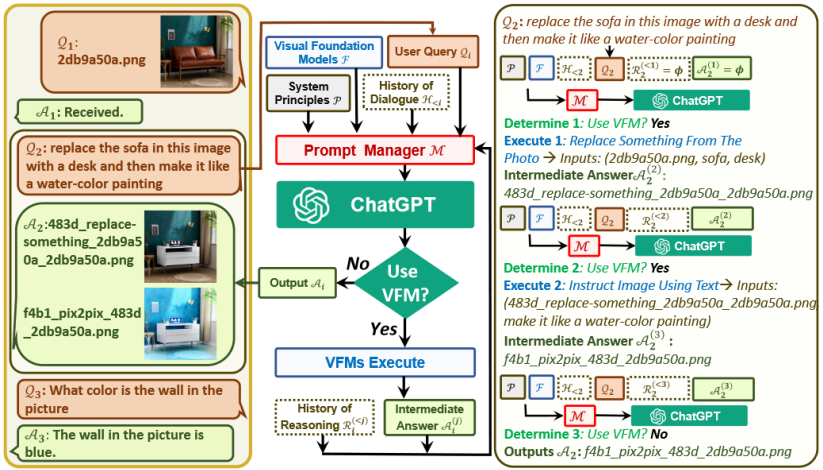
Generative Model - GPT

- OpenAI, 2023. GPT-4. Available at: https://openai.com/research/gpt-4. March 14, 2023.
  - (Optional) OpenAI (2023). GPT-4 Technical Report. ArXiv, abs/2303.08774.



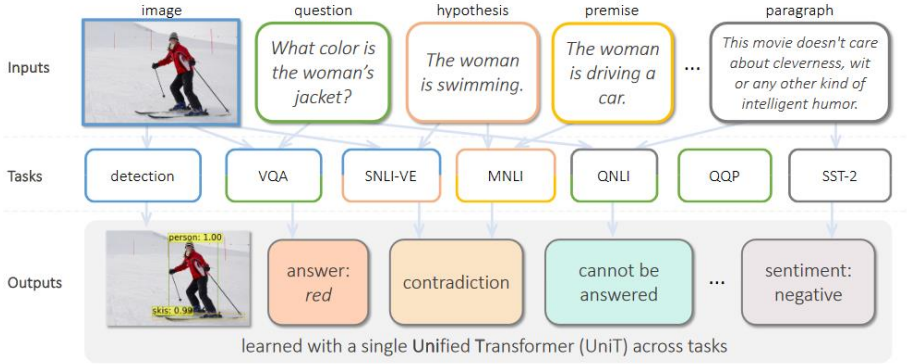
Generative Model - GPT

- Wu, C., Yin, S., Qi, W., Wang, X., Tang, Z. and Duan, N., 2023. Visual chatgpt: Talking, drawing and editing with visual foundation models. *arXiv preprint arXiv:2303.04671.*



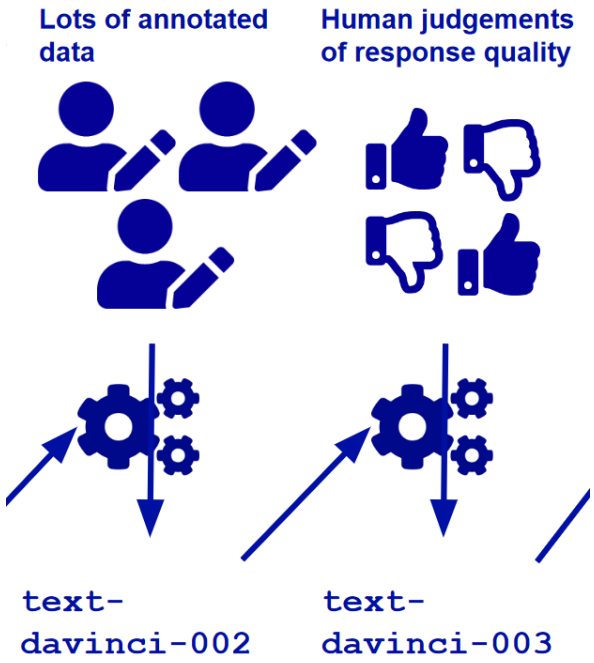
Multitask Learning

- Cho, J., Lei, J., Tan, H. & Bansal, M.. (2021). Unifying Vision-and-Language Tasks via Text Generation. *Proceedings of the 38th International Conference on Machine Learning*, in *Proceedings of Machine Learning Research* 139:1931-1942
- Hu, R. and Singh, A., 2021. Unit: Multimodal multitask learning with a unified transformer. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 1439-1449).
- Wang, P., Yang, A., Men, R., Lin, J., Bai, S., Li, Z., Ma, J., Zhou, C., Zhou, J. and Yang, H., 2022, June. Ofa: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In *International Conference on Machine Learning* (pp. 23318-23340). PMLR.



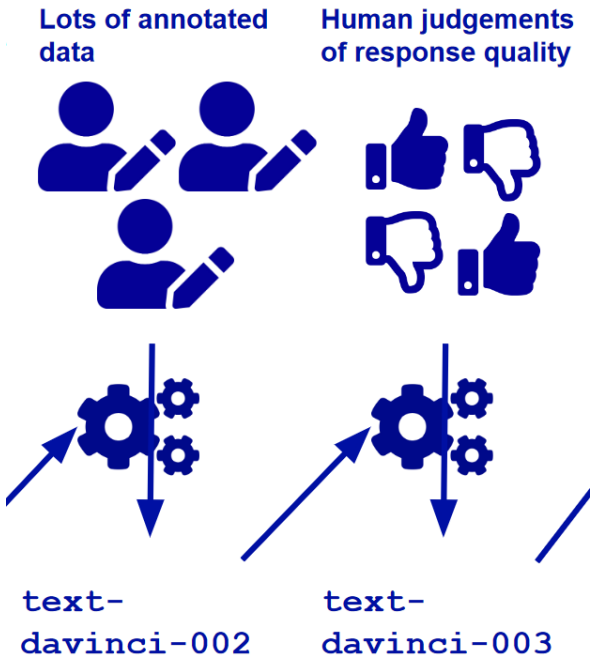
Reinforcement Learning

- Why ChatGPT works?



Reinforcement Learning

- Wang, X., Chen, W., Wang, Y.F. and Wang, W.Y., 2018, July. No Metrics Are Perfect: Adversarial Reward Learning for Visual Storytelling. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 899-909).
- Hu, J., Cheng, Y., Gan, Z., Liu, J., Gao, J. and Neubig, G., 2020, April. What makes a good story? designing composite rewards for visual storytelling. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 34, No. 05, pp. 7969-7976).



Paper Assignment

- Pre-training – Image
  - Contrastive
  - Seq2Seq Pre-training – Image
- Pre-training – Video
- Multitask Learning
- Parameter Efficiency
  - Prompting
  - Prompt Tuning
  - Prefix-Tuning
  - Adapters
- Recent Generative Models
  - Text-to-Image
  - Diffusion
  - GPT
- Reinforcement Learning

Peer review

- Helpful feedback for the presenters
- Single-blind review
- Don’ t worry. Your grades will not be affected by others’ review on you. (On the contrary, the reviews will be graded.)

Repeat:
Practical arrangement

- For presenter:
  - Make an appointment with me and show me your slides by the **Wednesday** before your talk.
  - Build a demo with Google Colab
- For everyone else except the presenter:
  - Read the paper to be presented, and post questions on MS Teams by **Friday** before the talk.
  - Submit peer review of the presentation by **Friday** after the talk.