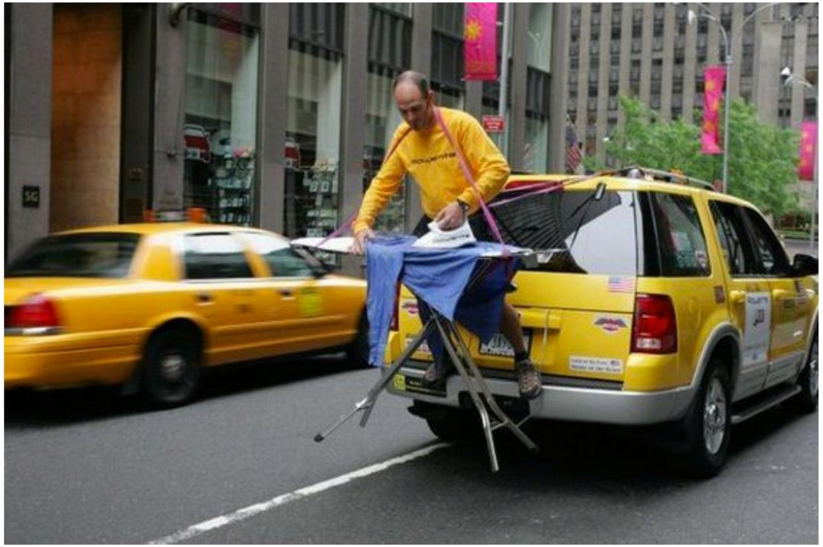
Machine Learning for **Language** and Vision

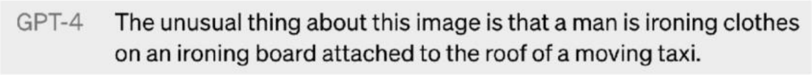
Seminar SS23
Kickoff Meeting, Apr 18, 2023

Xudong Hong, Ruitao Feng

Saarland University

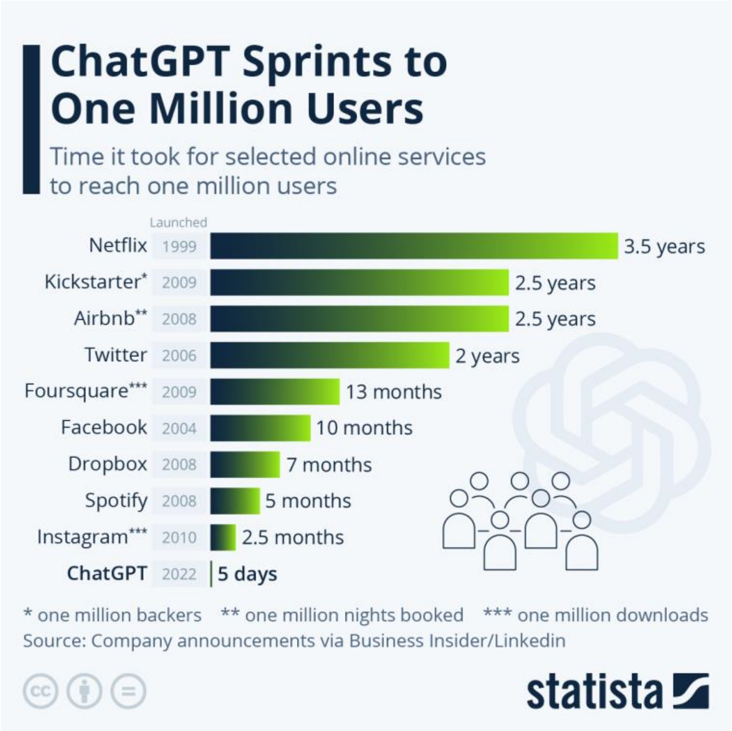
GPT-4





GPT-4 The unusual thing about this image is that a man is ironing clothes on an ironing board attached to the roof of a moving taxi.

The Stats



ChatGPT Sprints to One Million Users

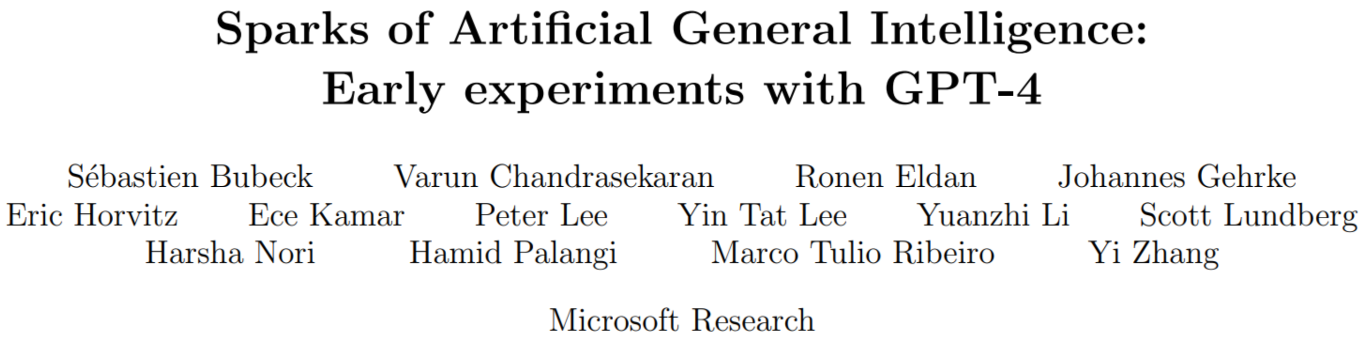
Time it took for selected online services to reach one million users

* one million backers ** one million nights booked *** one million downloads

Source: Company announcements via Business Insider/Linkedin

statista

Artificial General Intelligence?



Sparks of Artificial General Intelligence: Early experiments with GPT-4

Sébastien Bubeck Varun Chandrasekaran Ronen Eldan Johannes Gehrke
Eric Horvitz Ece Kamar Peter Lee Yin Tat Lee Yuanzhi Li Scott Lundberg
Harsha Nori Hamid Palangi Marco Tulio Ribeiro Yi Zhang

Microsoft Research

The Hype

PRO ROBOTS
GPT-5
WILL DESTROY
OUR WORLD

Be realistic

We are fully booked

https://huggingface.co/spaces/stabilityai/stable-diffusion

We are fully booked

SORRY WE'RE FULL

https://huggingface.co/spaces/stabilityai/stable-diffusion

Candidate Selection

SORRY
WE'RE FULL

- Being here does **NOT** guarantee your acceptance
- We need to further select 20 students (10 from CS and 10 from LST)
- For CS student, revise your answer to our survey form
- For LST student, remember to fill out the from by **Apr 21, Friday 23:00**
- We will finish the selection process by **Apr 23, Sunday**

Course overview

- Tue 12:15-13:45, 10 meetings (tentative, depends on our progress)

- Xudong Hong
  - xhong@coli.uni-saarland.de, C7 2, 2.02 by appointment.

- Ruitao Feng
  - fruitao@coli.uni-saarland.de, by appointment.

Course overview

- Assessment:
  - Presentation of a chosen topic (paper)
  - Prepare questions
  - Participation in discussion
  - Peer review of weekly presentations
  - Report (for the 7 CP version)

Prerequisites

- *How much do I need to know about neural networks/ NLP / CV to be able to follow the course?*

- 1. Students with practical experience with **NLP or CV** and want to
- deepen their theoretical understanding

- Or

- 2. Students who are familiar with neural NLP or CV (classes like NNIA,
- related seminars) and would like to gain practical experience

The most important prerequisite is you are really interested in the topic and you want to participate

How familiar is "familiar"?

- Do you know what the following mean:
  - Seq2Seq
  - Self-attention
  - Transformer

If you're unclear about some, that's okay!
I will talk about them in the next meeting.

Plan for today

1. **Introduction**
2. Reading List
3. Organization issues
4. Get to know each other

What is Language and Vision (L&V)?

Language-vision learning is an interdisciplinary research area that involves the development of sophisticated algorithms and models that can comprehend and generate complex, context-rich information from both visual and linguistic inputs.

Visual
Auditory
Read/Write
Kinesthetic

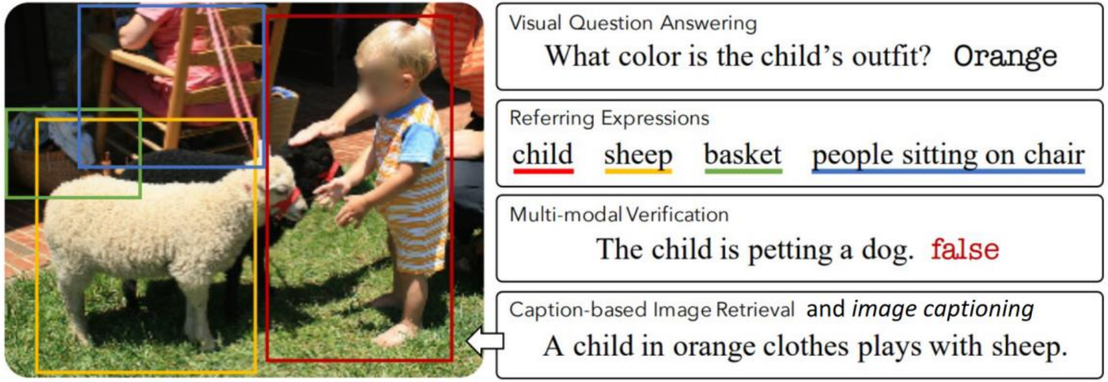
Why L&V Research?

- This is how we interact with and learn from the world
  - Vision is a large portion of how humans perceive
  - Language is a large portion of how humans communicate
  - A smart Ai system should be able to perform well on both
  - One step towards AGI

Why L&V Research?

- Multimodal learning has applications in various fields, including visual storytelling and human-computer interaction

Popular tasks:



Visual Question Answering

What color is the child's outfit? Orange

Referring Expressions

child sheep basket people sitting on chair

Multi-modal Verification

The child is petting a dog. false

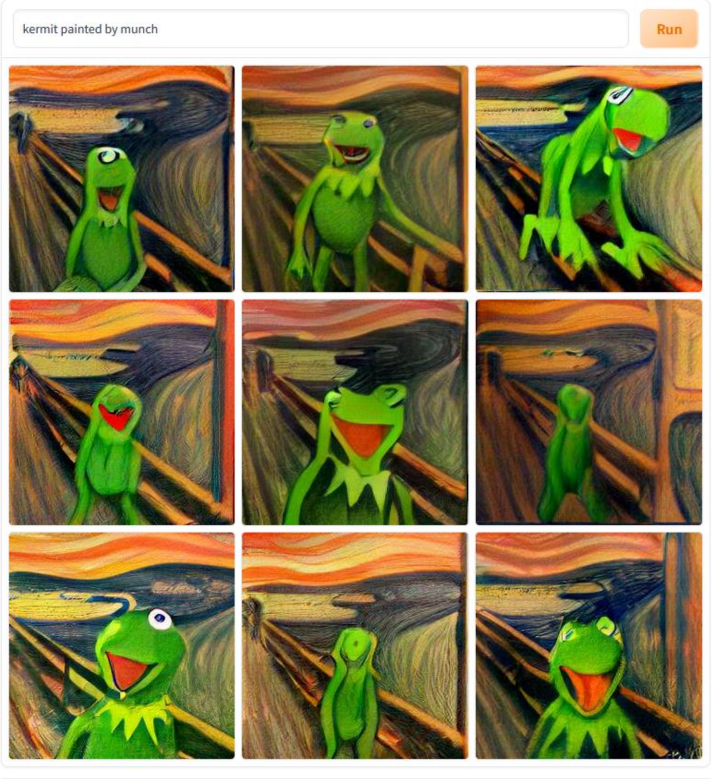
Caption-based Image Retrieval and *image captioning*

A child in orange clothes plays with sheep.

Why L&V Research?

It is just fun!

- OpenAI DALL-E 2
  - creates realistic images from a given text prompt


kermit painted by munch

Common VL Tasks- Image Captioning

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

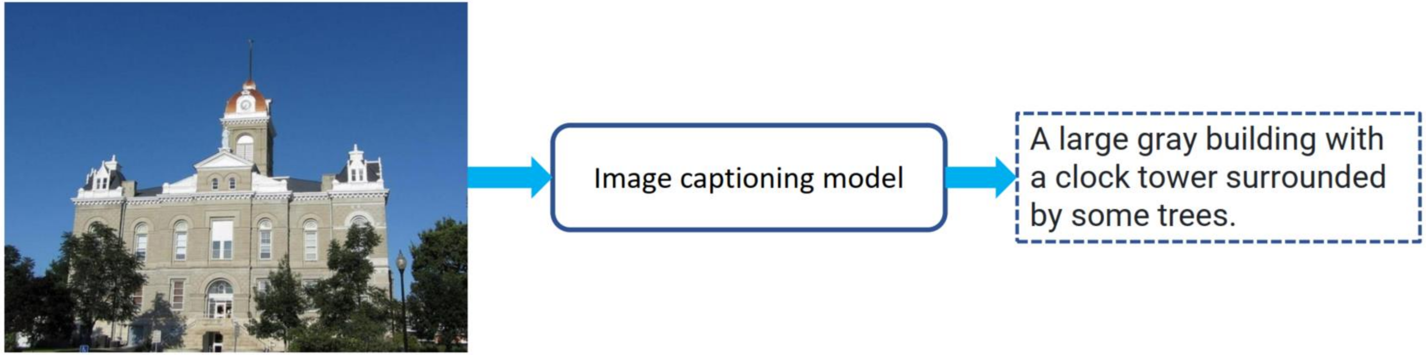


Image captioning model

A large gray building with a clock tower surrounded by some trees.

Common VL Tasks- Visual Question Answering

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

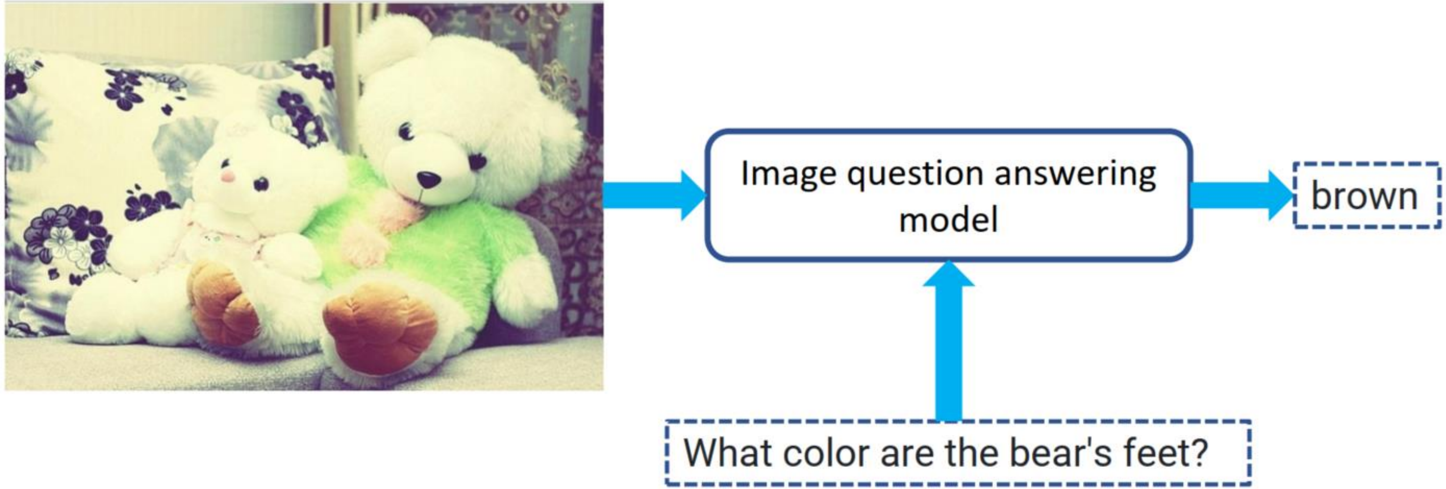
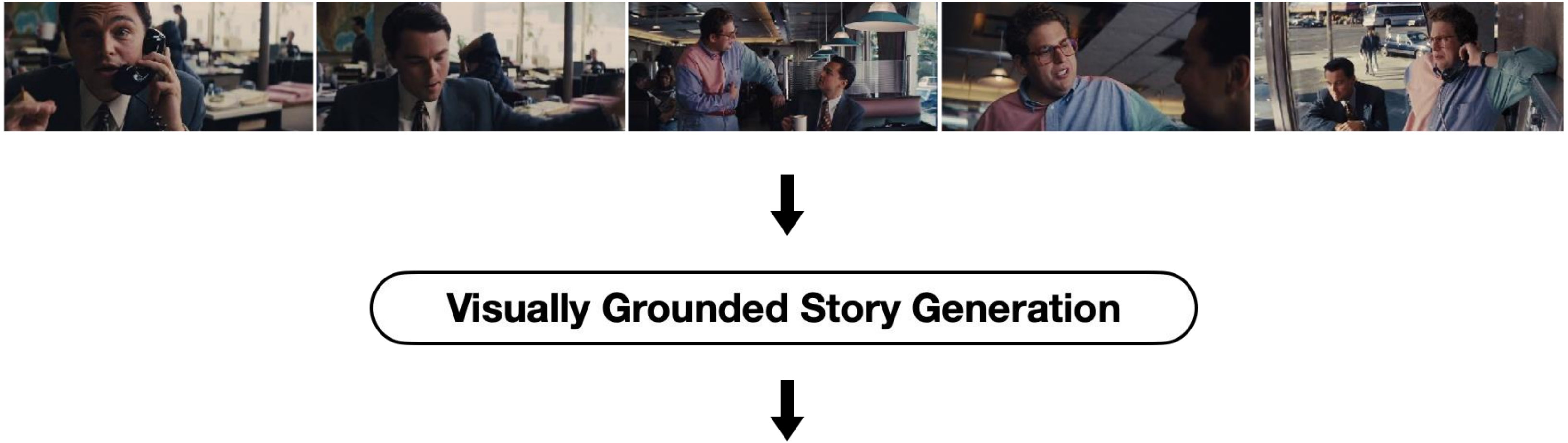


Image question answering model
brown
What color are the bear's feet?

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.

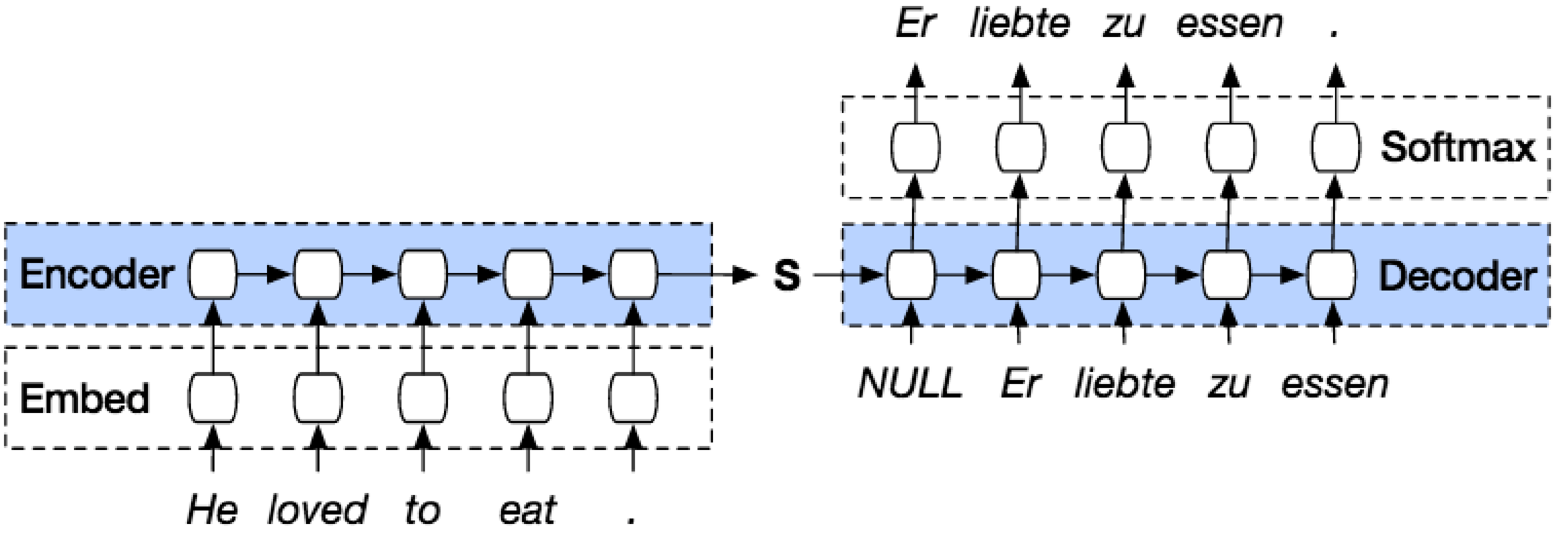
Jack put the phone down after an unsuccessful deal and decided to go get a coffee at the nearby coffee.

At the coffee shop, he started talking to the waiter Will about the unfortunate call.

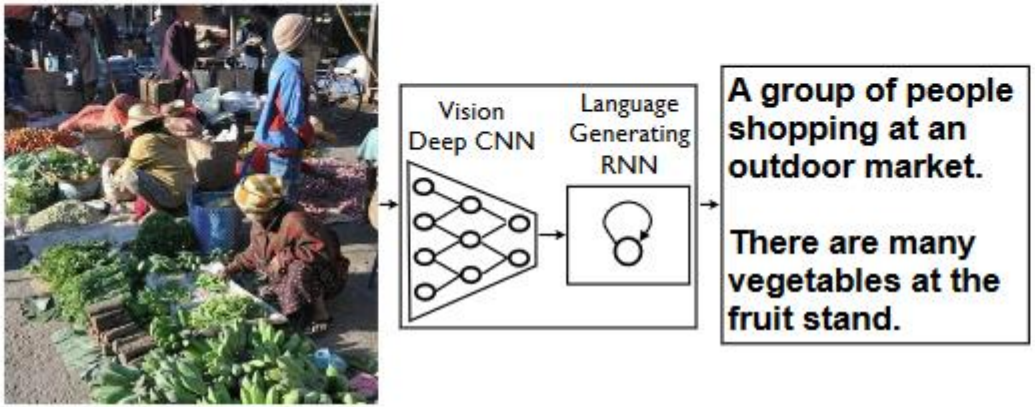
Will told him he would convince the client to accept the deal if he could work for Jack.

Will then called the client and successfully struck the deal.…

Seq2seq Models



Seq2seq Models


Vision
Deep CNN
Language
Generating
RNN
A group of people shopping at an outdoor market.
There are many vegetables at the fruit stand.

Vinyals, O., Toshev, A., Bengio, S. and Erhan, D., 2015. Show and tell: A neural image caption generator. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 3156-3164).

Plan for today

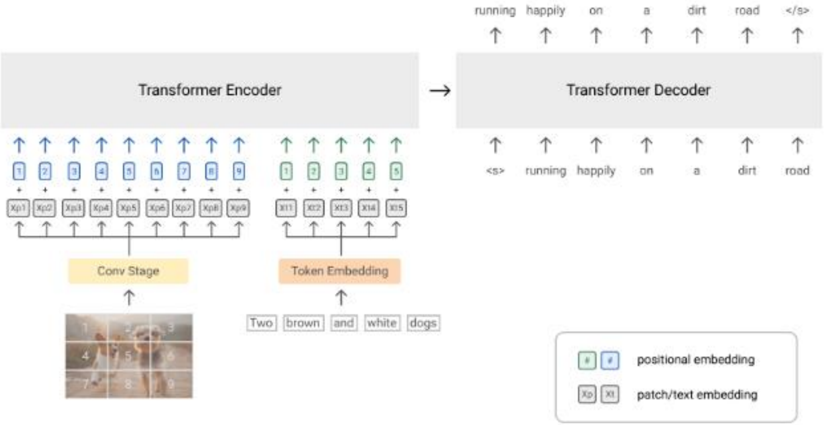
1. Introduction
2. **Reading List**
3. Organization issues
4. Get to know each other

Topics

- 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

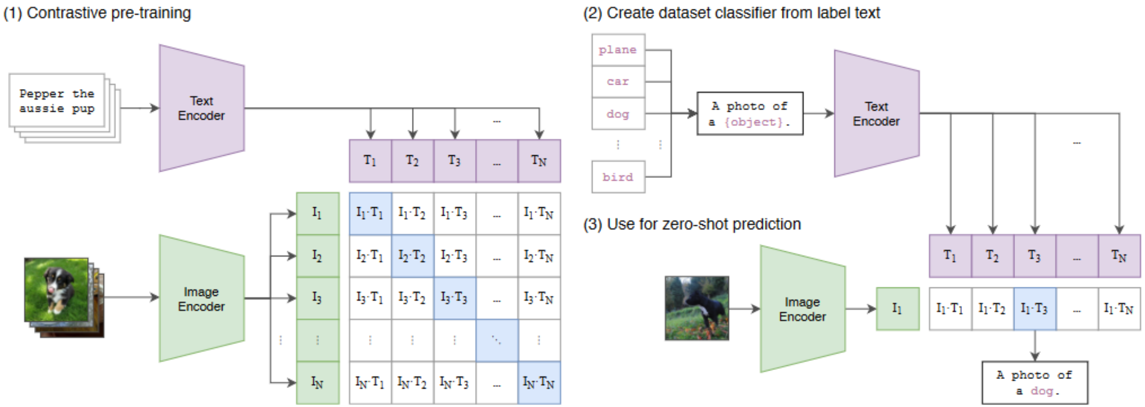
Seq2seq Pre-training - Image

- 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).



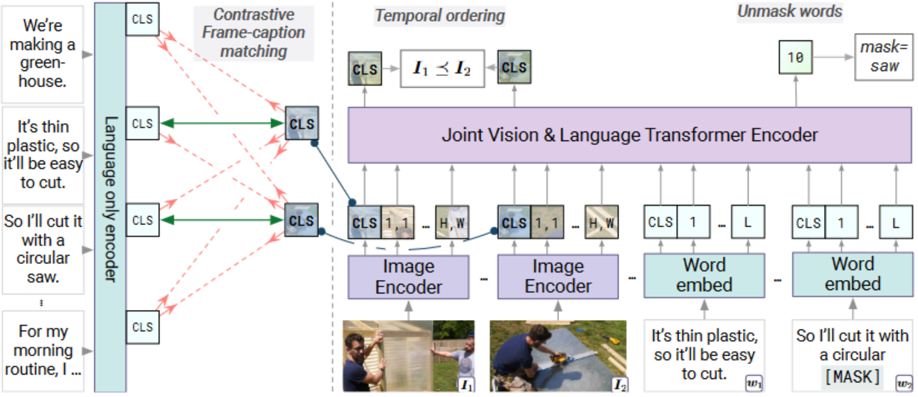
Contrastive Pre-training - Image

- 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.
- 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.


(1) Contrastive pre-training
Pepper the aussie pup
Text Encoder
Image Encoder
T1 T2 T3 ... TN
I1 I2 I3 ⋮ IN
I1·T1 I1·T2 I1·T3 ... I1·TN
I2·T1 I2·T2 I2·T3 ... I2·TN
I3·T1 I3·T2 I3·T3 ... I3·TN
⋮ ⋮ ⋮ ⋱ ⋮
IN·T1 IN·T2 IN·T3 ... IN·TN
(2) Create dataset classifier from label text
plane
car
dog
⋮
bird
A photo of a {object}.
Text Encoder
(3) Use for zero-shot prediction
Image Encoder
I1
T1 T2 T3 ... TN
I1·T1 I1·T2 I1·T3 ... I1·TN
A photo of a dog.

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.



We're making a green-house.
It's thin plastic, so it'll be easy to cut.
So I'll cut it with a circular saw.
For my morning routine, I ...
Language only encoder
CLS
Contrastive Frame-caption matching
Temporal ordering
$I_1 \preceq I_2$
Unmask words
10
mask= saw
Joint Vision & Language Transformer Encoder
1,1
H,W
1
L
Image Encoder
Word embed
I_1
I_2
It's thin plastic, so it'll be easy to cut.
w_1
So I'll cut it with a circular [MASK]
w_2

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.

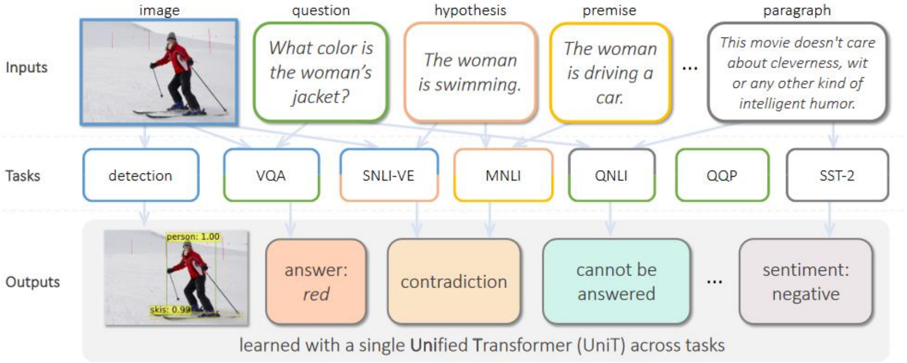
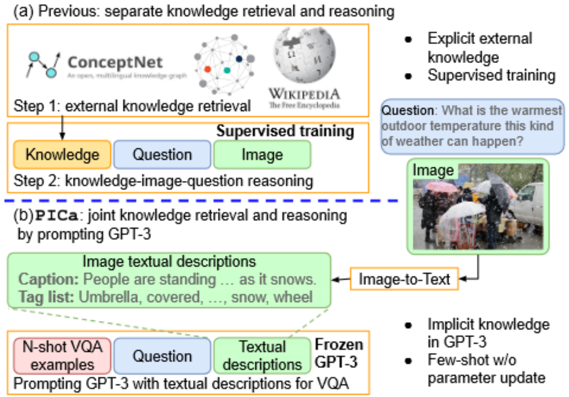


image
question
hypothesis
premise
paragraph
Inputs
What color is the woman's jacket?
The woman is swimming.
The woman is driving a car.
...
This movie doesn't care about cleverness, wit or any other kind of intelligent humor.
Tasks
detection
VQA
SNLI-VE
MNLI
QNLI
QQP
SST-2
Outputs
person: 1.00
skis: 0.99
answer: red
contradiction
cannot be answered
...
sentiment: negative
learned with a single Unified Transformer (UniT) across tasks

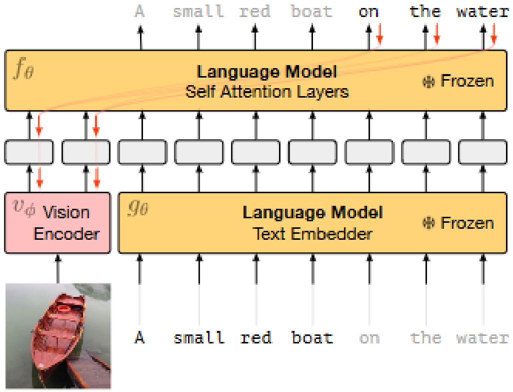
Parameter Efficiency - Prompting

- 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.


(a) Previous: separate knowledge retrieval and reasoning
ConceptNet
WIKIPEDIA
Step 1: external knowledge retrieval
Supervised training
Knowledge
Question
Image
Step 2: knowledge-image-question reasoning
Explicit external knowledge
Supervised training
Question: What is the warmest outdoor temperature this kind of weather can happen?
Image
(b) PICa: joint knowledge retrieval and reasoning by prompting GPT-3
Image textual descriptions
Caption: People are standing … as it snows.
Tag list: Umbrella, covered, …, snow, wheel
Image-to-Text
N-shot VQA examples
Question
Textual descriptions
Frozen GPT-3
Prompting GPT-3 with textual descriptions for VQA
Implicit knowledge in GPT-3
Few-shot w/o parameter update

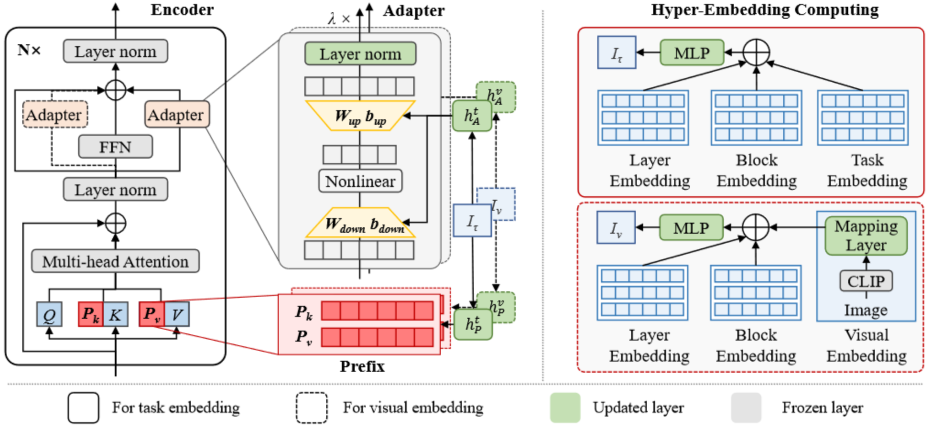
Parameter Efficiency - Prompt Tuning

- 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).



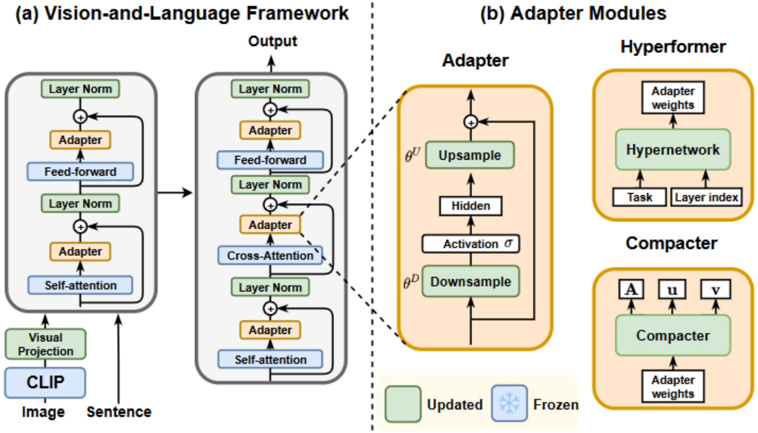
Parameter Efficiency - Prefix-Tuning

- 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).



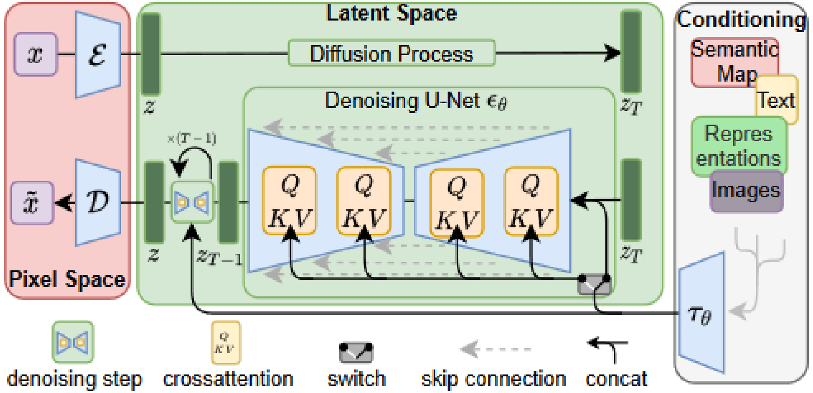
Parameter Efficiency - Adapters

- Sung, Y.L., Cho, J. and Bansal, M., 2022. Vl-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*.



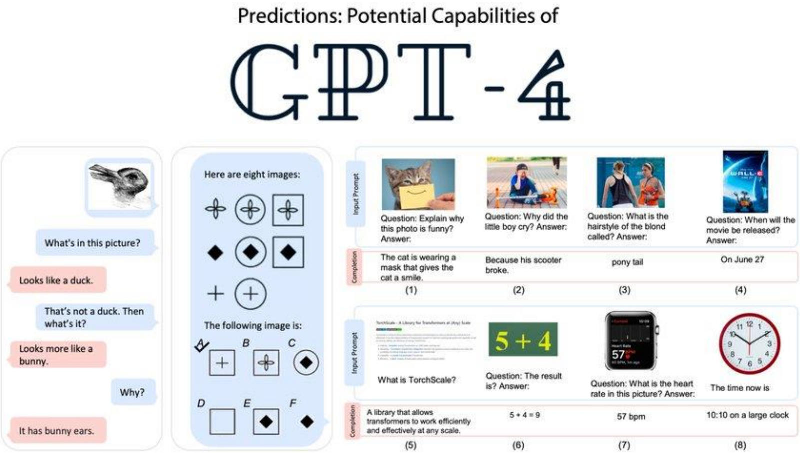
Generative Model - Text-to-Image

- 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.



Predictions: Potential Capabilities of

GPT-4

What's in this picture?

Looks like a duck.

That's not a duck. Then what's it?

Looks more like a bunny.

Why?

It has bunny ears.

Here are eight images:

The following image is:

Question: Explain why this photo is funny? Answer:

The cat is wearing a mask that gives the cat a smile.

Question: Why did the little boy cry? Answer:

Because his scooter broke.

Question: What is the hairstyle of the blond called? Answer:

pony tail

Question: When will the movie be released? Answer:

On June 27

What is TorchScale?

A library that allows transformers to work efficiently and effectively at any scale.

Question: The result is? Answer:

5 + 4 = 9

Question: What is the heart rate in this picture? Answer:

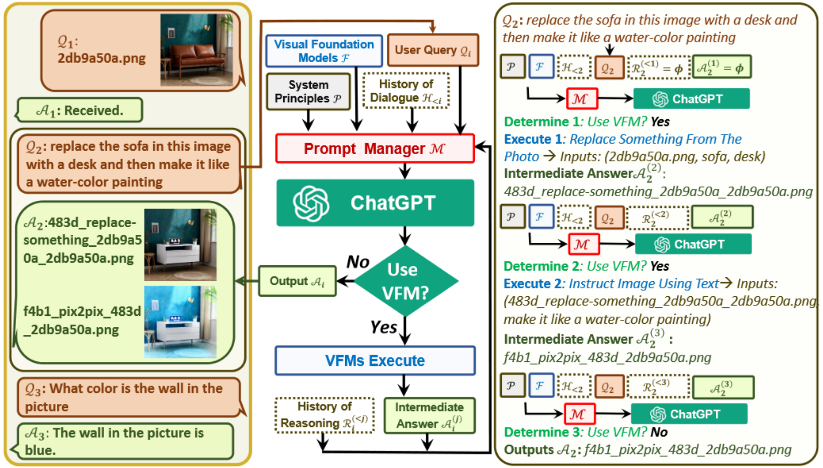
57 bpm

The time now is

10:10 on a large clock

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.*


Visual Foundation Models $\mathcal{F}$
User Query Q_i
System Principles $\mathcal{P}$
History of Dialogue $\mathcal{H}_{<i}$
Prompt Manager $\mathcal{M}$
ChatGPT
Use VFM?
No
Yes
Output $\mathcal{A}_i$
VFMs Execute
History of Reasoning $\mathcal{R}_i^{(<j)}$
Intermediate Answer $\mathcal{A}_i^{(j)}$
$\mathcal{P}$
$\mathcal{F}$
$\mathcal{H}_{<2}$
Q_2
$\mathcal{R}_2^{(<1)} = \phi$
$\mathcal{A}_2^{(1)} = \phi$
$\mathcal{M}$
ChatGPT
$\mathcal{R}_2^{(<2)}$
$\mathcal{A}_2^{(2)}$
$\mathcal{R}_2^{(<3)}$
$\mathcal{A}_2^{(3)}$

Q_1: 2db9a50a.png

$\mathcal{A}_1$: Received.

Q_2: replace the sofa in this image with a desk and then make it like a water-color painting

$\mathcal{A}_2$:483d_replace-something_2db9a50a_2db9a50a.png

f4b1_pix2pix_483d_2db9a50a.png

Q_3: What color is the wall in the picture

$\mathcal{A}_3$: The wall in the picture is blue.

Q_2: replace the sofa in this image with a desk and then make it like a water-color painting

Determine 1: *Use VFM?* ***Yes***

Execute 1: *Replace Something From The Photo* → *Inputs: (2db9a50a.png, sofa, desk)*

Intermediate Answer $\mathcal{A}_2^{(2)}$:

483d_replace-something_2db9a50a_2db9a50a.png

Determine 2: *Use VFM?* ***Yes***

Execute 2: *Instruct Image Using Text* → *Inputs: (483d_replace-something_2db9a50a_2db9a50a.png, make it like a water-color painting)*

Intermediate Answer $\mathcal{A}_2^{(3)}$:

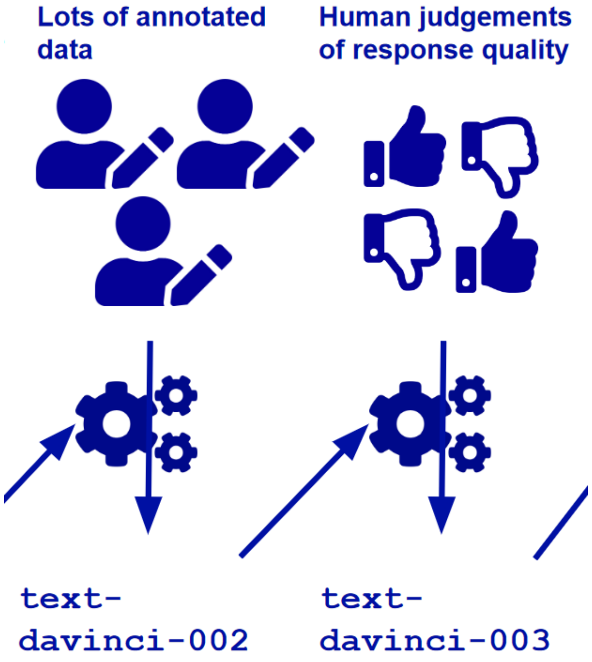
f4b1_pix2pix_483d_2db9a50a.png

Determine 3: *Use VFM?* ***No***

Outputs $\mathcal{A}_2$: *f4b1_pix2pix_483d_2db9a50a.png*

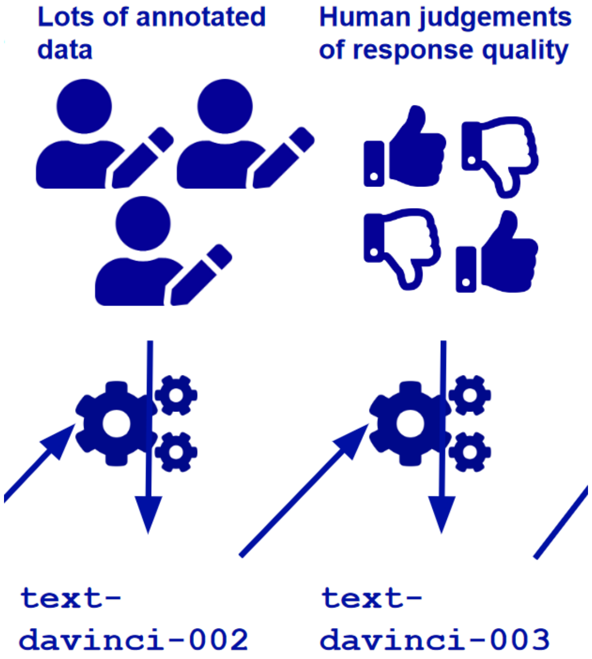
Reinforcement Learning

- 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).



Plan for today

1. Introduction
2. Reading List
3. **Organization issues**
4. Get to know each other

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.)

Quick guide for a good presentation

Have a slide that answer these questions:

- Why is this work important?
  i.e. What is the general problem ?
  or what is the main question ?
- What is the purpose of this work?
  i.e. How does this work contribute to the sub-problem of the general problem?
  or what specific questions are this work trying to answer?

Quick guide for a good presentation

- Problem 1, Problem 2, Problem 3
  Experiment 1, Experiment 2, Experiment 3,
  Results 1, Results 2, Results 3

Bad !!

  - Problem 1, Experiment 1, Results 1,
    Problem 2, Experiment 2, Results 2,
    Results 1, Results 2, Results 3

Better !!

- "But I want to compare the experiments"
  - Just show Results 1 again alongside Results 2

Quick guide for a good presentation

Have a slide that answer these questions:

- Why is this work important?
  i.e. What is the general problem ?
  or what is the main question ?
- What is the purpose of this work?
  i.e. How does this work contribute to the sub-problem of the general problem?
  or what specific questions are this work trying to answer?

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.

Plan for today

1. Introduction
2. Reading List
3. Organization issues
4. **Get to know each other**

It’ s your time!

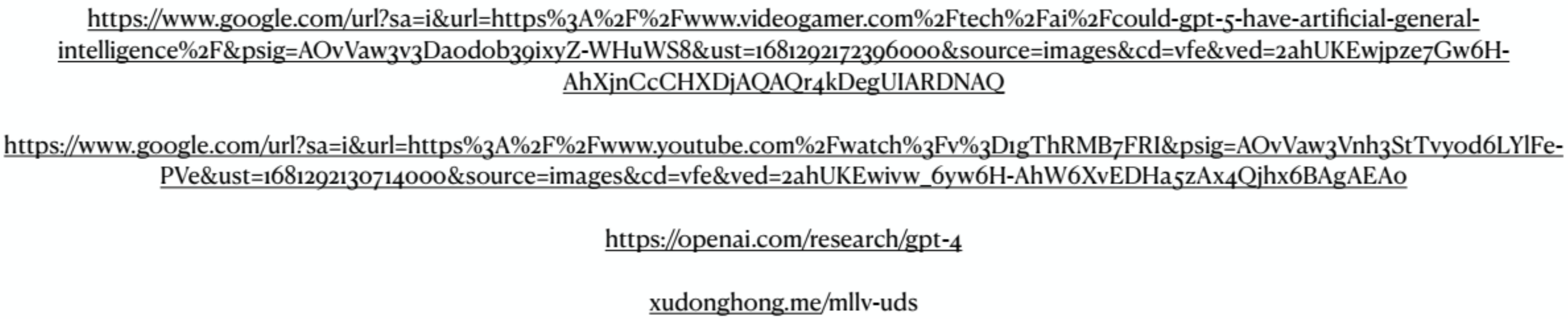
Time to get to know each other

1. Your name and what you prefer to be called
2. Year in BSc/MSc
3. Native language(s)
4. Prior degree and areas of interest/specialization
5. Have you taken a related course before?

shutterstock.com · 1078542572

References

- Yung, 2019, Discourse Relations: cognition, resources, NLP
- Iza, 2021, Controllable Text Generation



https://www.google.com/url?sa=i&url=https%3A%2F%2Fwww.videogamer.com%2Ftech%2Fai%2Fcould-gpt-5-have-artificial-general-intelligence%2F&psig=AOvVaw3v3Daodob39ixyZ-WHuWS8&ust=1681292172396000&source=images&cd=vfe&ved=2ahUKEwjpze7Gw6H-AhXjnCcCHXDjAQAQr4kDegUIARDNAQ

https://www.google.com/url?sa=i&url=https%3A%2F%2Fwww.youtube.com%2Fwatch%3Fv%3D1gThRMB7FRI&psig=AOvVaw3Vnh3StTvyod6LYlFe-PVe&ust=1681292130714000&source=images&cd=vfe&ved=2ahUKEwivw_6yw6H-AhW6XvEDHa5zAx4Qjhx6BAgAEAo

https://openai.com/research/gpt-4

xudonghong.me/mllv-uds