

Multimodal NLP

AthNLP 2024

Raquel Fernández

Institute for Logic, Language and Computation
University of Amsterdam



About me

- Barcelona ➤ London ➤ Potsdam ➤ Stanford ➤ Amsterdam
- Background in computational linguistic/NLP and cognitive science
- General interest in language use for communication (“dialogue”) – language shaped by perception and social interaction
- Some topics me and my group are currently working on: using LMs for questions about human cognition, uncertainty, visual grounding, evaluation, ...

Multimodal models

Multimodal models process information from two or more *modalities* (i.e., means to convey information): text, speech, images, video, smells, sounds, actions, code,

I will focus on the interplay between **language & vision**, from an NLP perspective

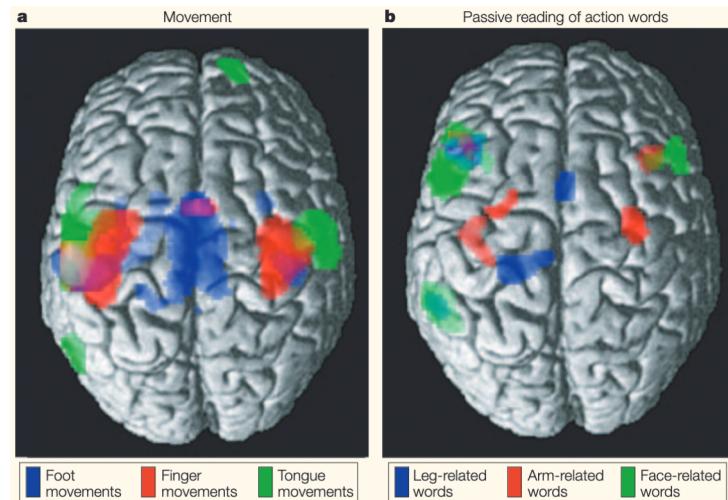
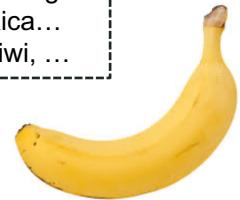
Why multimodality?

Human conceptual representations are multimodal

The **symbol grounding problem**: how do symbols (like words) acquire meaning? Arguably, extra-linguistic information is necessary to connect words to the world.

Embodied cognition: Evidence that acting with our bodies and processing related action words activate similar brain regions.

... eating a **banana** after exercising...
... ripe **bananas** from Costa Rica...
... a fruit salad with **banana**, kiwi, ...



Harnad (1990). The Symbol Grounding Problem, *Physica*.

Barsalou et al. (2003). Grounding conceptual knowledge in modality-specific systems. *Trends in Cognitive Sciences*, 7(2):84–91.
Pulvermüller. (2005). Brain mechanisms linking language and action. *Nature Reviews Neuroscience*, 6(7), 576–582.

Why multimodality?

Not only conceptual representations, but
human perception is multimodal too.

The McGurk effect: what we see may
overwrite what we hear...

Listen with your eyes closed, then open. What
do you hear: /ba-ba/ or /ta-ta/ ?



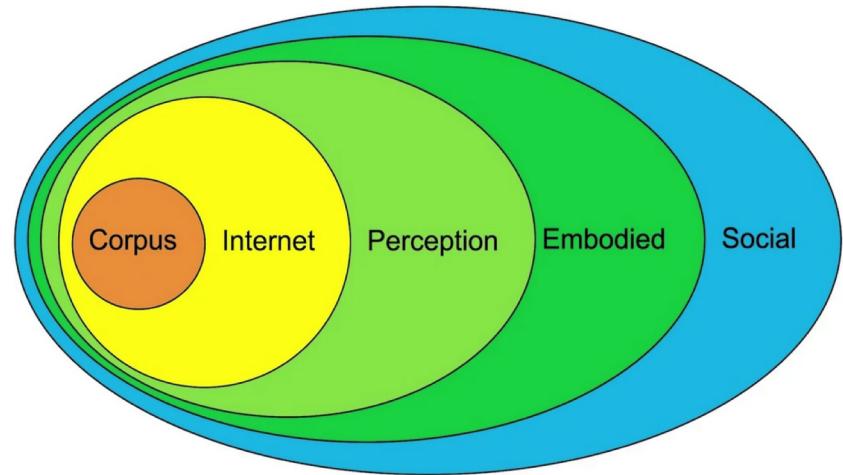
<https://auditoryneuroscience.com/McGurkEffect>
<https://www.youtube.com/watch?v=2k8fHR9jKVM>

Why multimodality?

Language is tight to experience and human experience is multimodal.

You can't fully learn language:

- Just listening to the radio or reading books → *lack of perception*
- Just watching television → *lack of embodiment and action*
- Without communicating with others → *lack of social interaction*



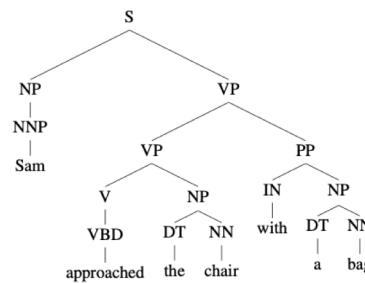
Bisk et al. Experience Grounds Language. EMNLP 2020.

Why multimodality?

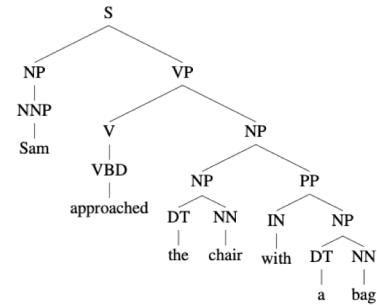
From a more practical point of view,
multimodality reduces ambiguity.

Plus many useful applications are
multimodal!

- Improving accessibility for different types of user
- Robot assistants
- ...



(a) First interpretation



(b) Second interpretation



Berzak et al. (2015) Do you see what I mean? Visual resolution of linguistic ambiguities. EMNLP.

The Plan

Part 1

Two types of approaches:
task specific & general purpose

- Datasets
- Modelling techniques

Part 2

- Ethical issues re. training data
 - Evaluation
 - New directions
-

General trend....

- ... from task-specific approaches...
 - Image captioning, visual question answering, etc
- ... to general purpose (task-agnostic) models.
 - foundation multimodal models

As in text-based in NLP.

Task-specific approaches

- Datasets and architectures designed to tackle specific tasks
 - Supervised learning
-

Many types of tasks

- Image-conditioned text generation
 - Image captioning, video description, visual storytelling
- Classification
 - Visual question answering
 - Visually grounded reasoning
- Ranking and alignment
 - Image retrieval
 - Referring expression localization

All these tasks involve combining information from language and vision.

and more and more and more (visual entailment, image generation, ...)

Representing visual data

You have already seen how to learn text representations.

How do we represent information from other modalities, in particular vision?

- In very early approaches, **symbolic features** were used to represent objects or scenes, *without any vision*



```
<ENTITY ID="54" IMAGE="Eilenberg.jpg" TYPE="target">
  <ATTRIBUTE NAME="hasBeard" TYPE="boolean" VALUE="1"/>
  <ATTRIBUTE NAME="hasTie" TYPE="boolean" VALUE="0"/>
  <ATTRIBUTE NAME="type" TYPE="literal" VALUE="person"/>
  <ATTRIBUTE NAME="hasHair" TYPE="boolean" VALUE="0"/>
  <ATTRIBUTE NAME="hasGlasses" TYPE="boolean" VALUE="0"/>
  <ATTRIBUTE NAME="hasSuit" TYPE="boolean" VALUE="0"/>
  <ATTRIBUTE NAME="age" TYPE="literal" VALUE="old"/>
  <ATTRIBUTE NAME="hairColour" TYPE="literal" VALUE="light"/>
  <ATTRIBUTE NAME="orientation" TYPE="literal" VALUE="left"/>
  <ATTRIBUTE NAME="hasShirt" TYPE="boolean" VALUE="1"/>
</ENTITY>
```

The person with a light colored beard

early referring expression generation approaches focused on this mapping

Representing visual data

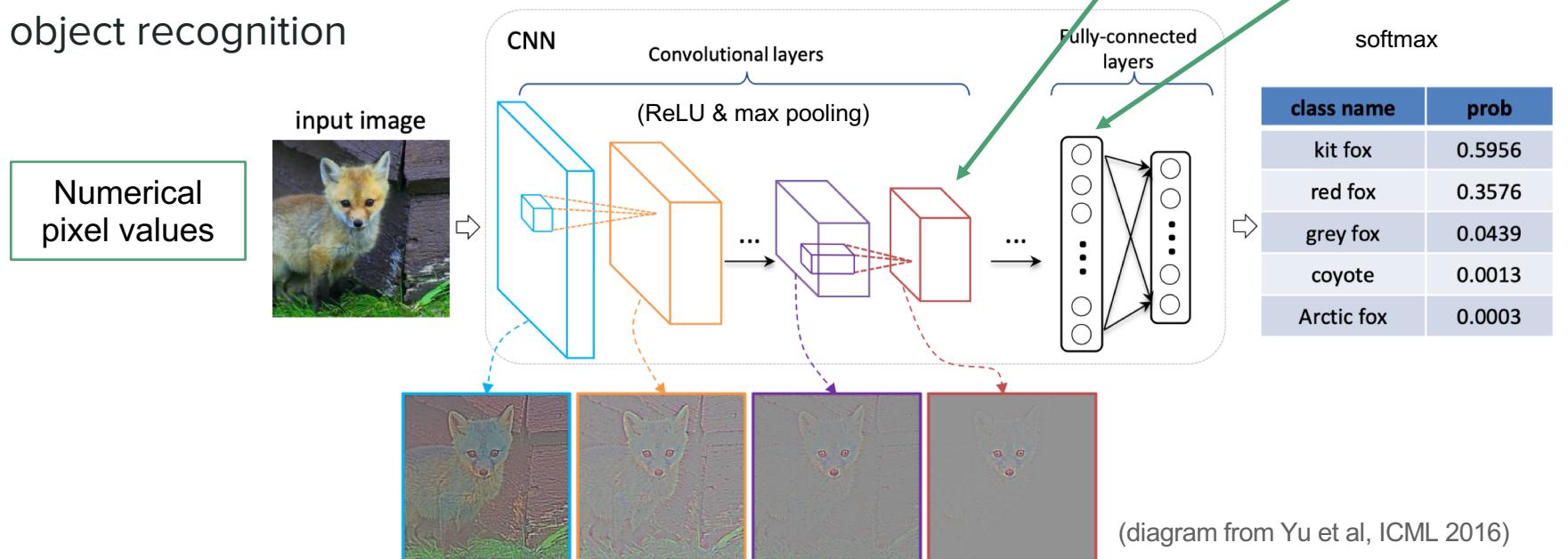
You have already seen how to learn text representations.

How do we represent information from other modalities, in particular vision?

- In very early approaches, **symbolic features** were used to represent objects or scenes, *without any vision*
- As computer vision methods started to be further developed, the focus shifted towards **automatically learning to represent visual information**

CNNs

Earliest work in the neural-network era used features from CNNs pre-trained on object recognition



Zisserman & Simonyan, 2015, Very Deep Convolutional Networks for Large-Scale Image Recognition, ICLR.

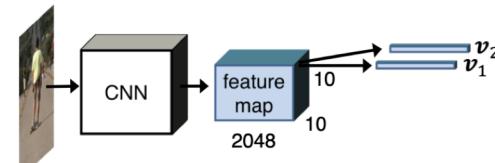
CNNs: Features for regions of interest

From general spatial information to regions corresponding to objects/entities

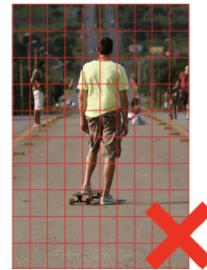
R-CNN region-based feature vectors:

- Trained on the Visual Genome Dataset for object recognition.
- The Region Proposal Network suggests the location of *regions of interest*.

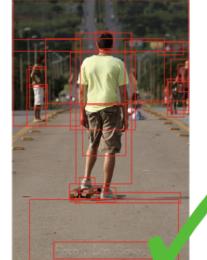
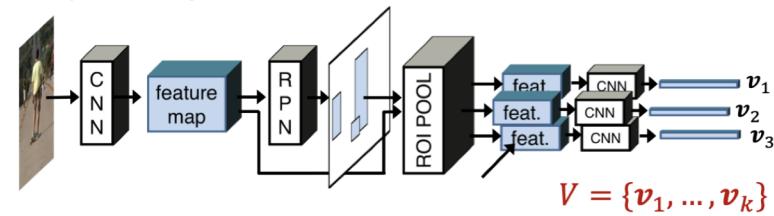
Spatial output of a CNN



$$V = \{v_1, \dots, v_{100}\}$$



Object regions with R-CNN



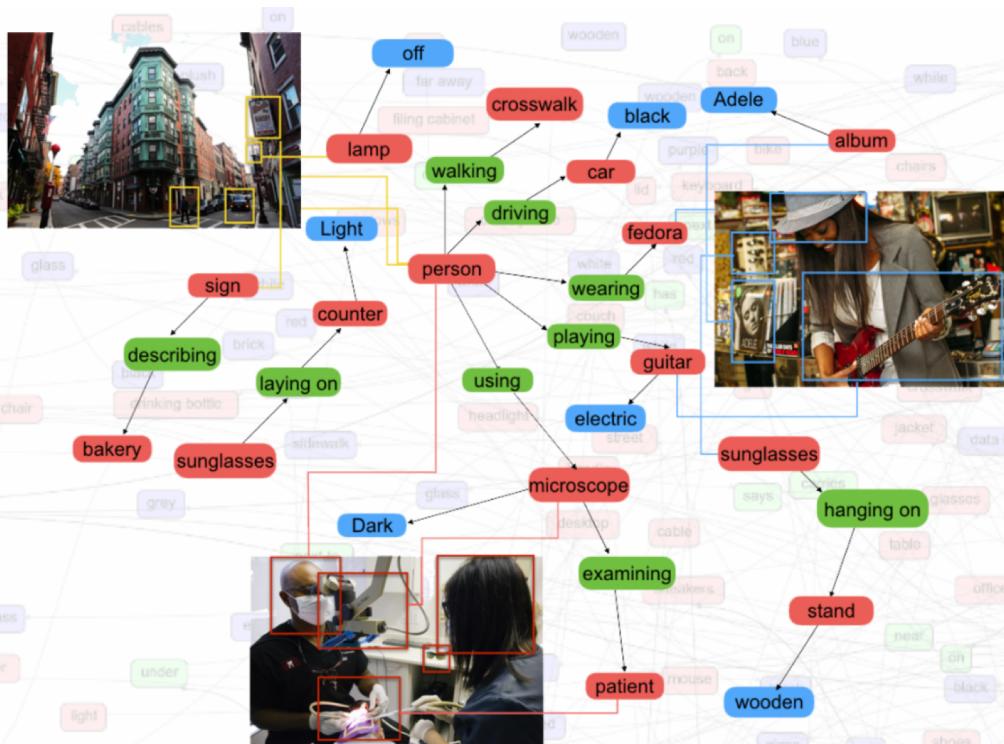
Ren et al. (2015). Faster R-CNN: Towards real-time object detection with region proposal networks. *NeurIPS*.

Anderson et al. 2018. Bottom-Up and Top-Down Attention for Image Captioning and VQA. *CVPR*.

Diagram from poster: https://panderson.me/images/cvpr18_UpDown_poster.pdf

Visual Genome

<https://homes.cs.washington.edu/~ranjay/visualgenome/>



Visual Genome is a dataset, a knowledge base, an ongoing effort to connect structured image concepts to language.

108,077 Images

5.4 Million Region Descriptions

1.7 Million Visual Question Answers

3.8 Million Object Instances

2.8 Million Attributes

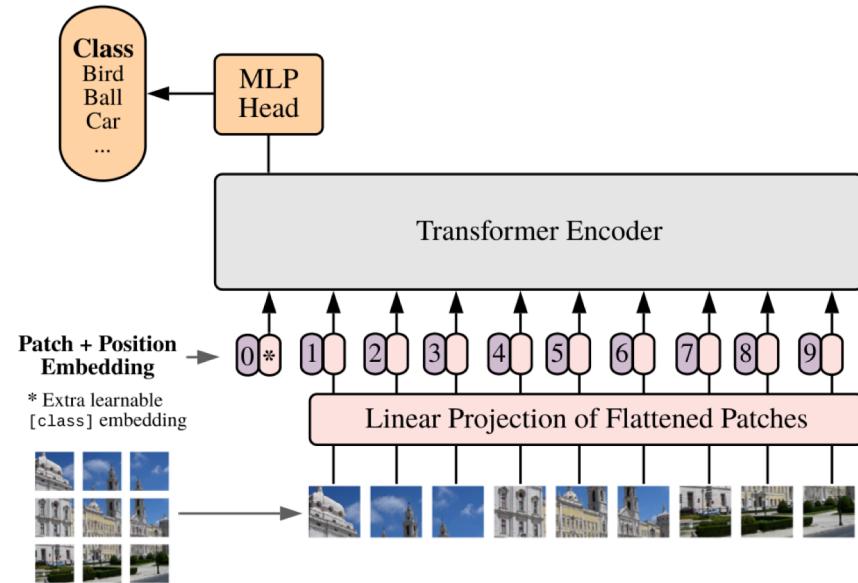
2.3 Million Relationships

Everything Mapped to Wordnet Synsets

Vision Transformer (ViT)

- Split image into patches
 - This transforms the image into “tokens” like text, and makes the process more efficient
- Embed each patch (flattening)
- Add positional embeddings
- Encode using Transformer blocks
- Possibly pretrain on image classification

Better spatial and contextual information than CNNs



Back to our many types of tasks

- Image-conditioned text generation
 - Image captioning, video description, visual storytelling
- Classification
 - Visual question answering
 - Visually grounded reasoning
- Ranking and alignment
 - Image retrieval
 - Referring expression localization

All these tasks involve combining information from language and vision.

Visual data representation ready thanks to deep learning revolution in CV

Datasets to learn from

and more and more and more (visual entailment, image generation, ...)

COCO: Common Objects in Context

Highly influential dataset

Multiple human-authored captions,
with object segmentation.

some sheep walking in the middle of a road
a herd of sheep with green markings walking down the road
a herd of sheep walking down a street next to a lush green grass covered hillside.
sheared sheep on roadway taken from vehicle, with green hillside in background.
a flock of freshly sheered sheep in the road.



What is COCO?



COCO is a large-scale object detection, segmentation, and captioning dataset.
COCO has several features:

- ✓ Object segmentation
- ✓ Recognition in context
- ✓ Superpixel stuff segmentation
- ✓ 330K images (>200K labeled)
- ✓ 1.5 million object instances
- ✓ 80 object categories
- ✓ 91 stuff categories
- ✓ 5 captions per image
- ✓ 250,000 people with keypoints

cocodataset.org

Lin et al. (2014), COCO: Common Objects in Context.
Chen et al. (2015), Microsoft COCO captions: Data collection and evaluation server.

Multi30K: Multilingual aligned image-sentence dataset

- English, German, French, Czech, Arabic, Japanese, Turkish, Ukrainian

A group of people are eating noodles.

Eine Gruppe von Leuten isst Nudeln.

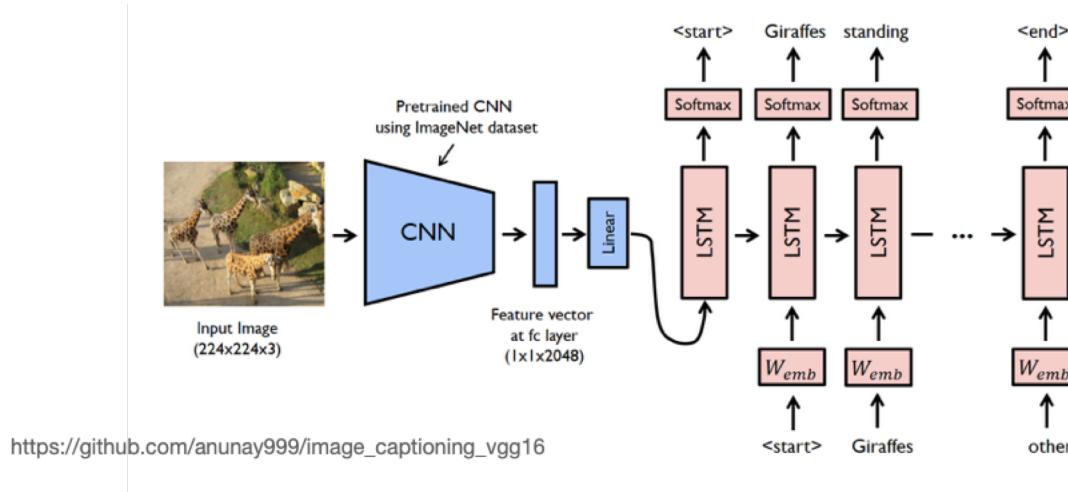
Un groupe de gens mangent des nouilles.

Skupina lidí jedí nudle.



Task-specific models: Image captioning

Encoder-decoder architecture: a language model conditioned on visual information



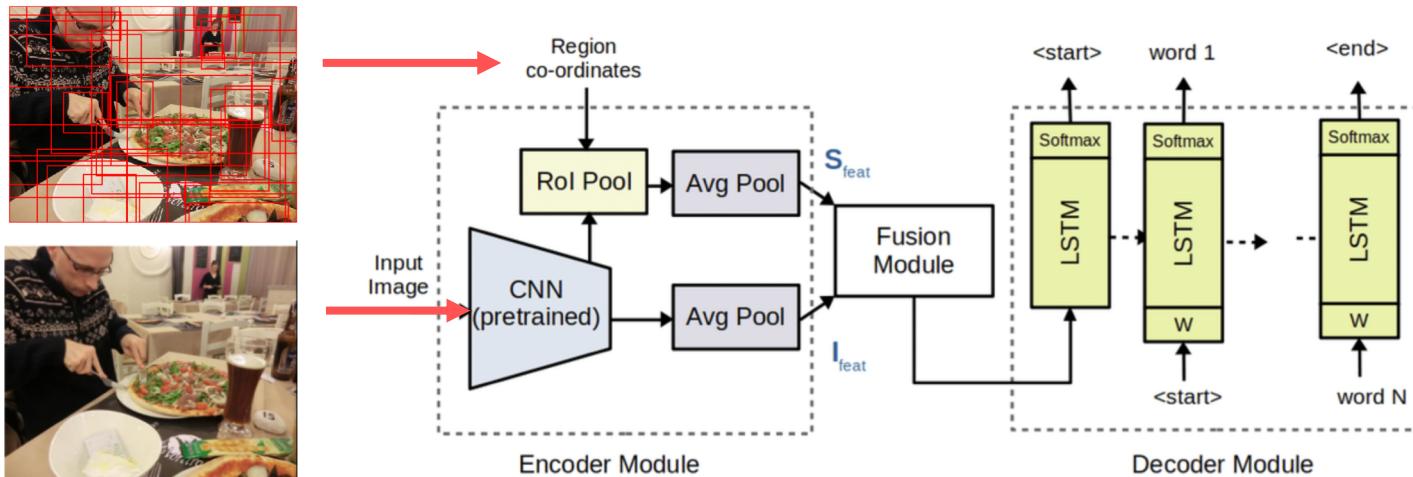
This basic architecture can be enriched in different ways...

Task-specific models: Image captioning

Encoder-decoder architecture: a language model conditioned on visual information

This basic architecture can be enriched in different ways...

- For example, by additionally using visual features for **regions of interest**, **attention** over these features, etc.



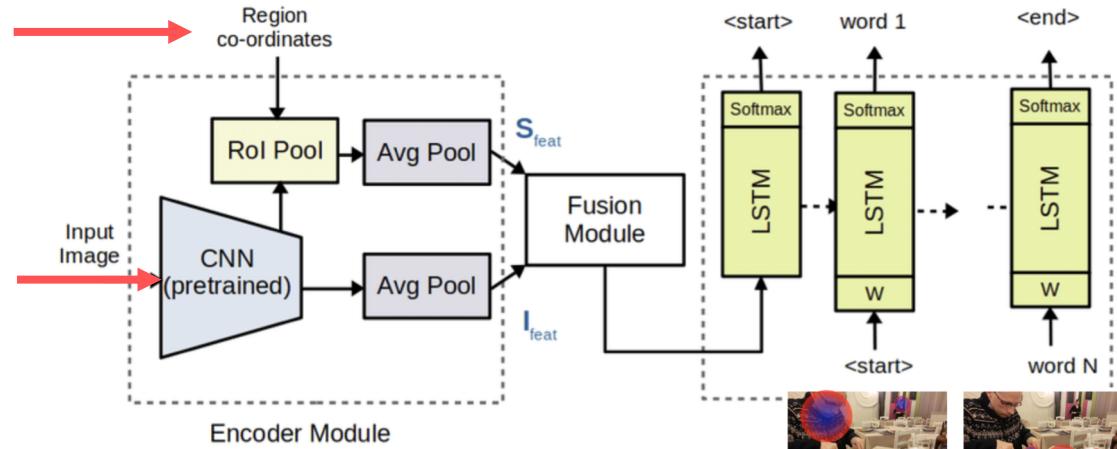
Task-specific models: Image captioning

Encoder-decoder architecture: a language model conditioned on visual information

This basic architecture can be enriched in different ways...

- Or by using information from **human gaze**, exploiting its sequential nature!

Static



sequential

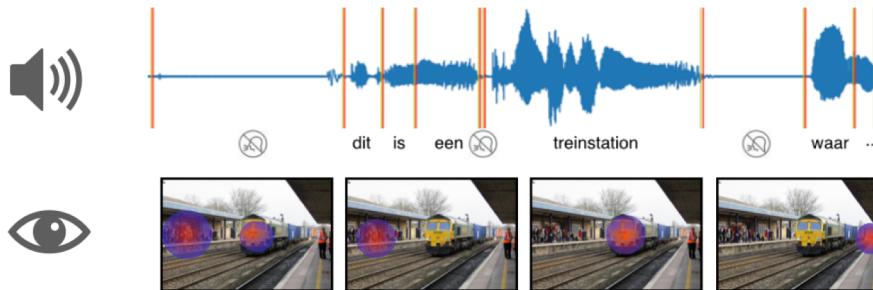
DIDEC: Dutch image description eye-tracking corpus



dit is een treinstation waarbij mensen op het perron aan het wachten zijn en waarbij net een goederentrein langsrijdt

(this is a train station where people are waiting on the platform and where a freight train is just passing by)

(DIDEC dataset; van Miltenburg et al. 2018)



Among other things, such a dataset allows us to investigate sequential cross-modal alignment

(van Miltenburg et al, 2018)

Generating Image Descriptions via Sequential Cross-Modal Alignment Guided by Human Gaze

Ece Takmaz¹, Sandro Pezzelle¹, Lisa Beinborn², Raquel Fernández¹

¹Institute for Logic, Language and Computation, University of Amsterdam

²Vrije Universiteit Amsterdam

(EMNLP 2020)

- Eye tracking coupled with language production as a guide to image description generation
- Adaptation of image captioning model by Anderson et al. (2018)
- This leads to more specific and human-like descriptions



specificity

NO-G	een vrouw die in de keuken staat... (<i>a woman who is standing in the kitchen...</i>)
2SEQ	een vrouw in een keuken met donuts (<i>a woman in the kitchen with donuts</i>)

R: een bakkerij met een rek met een heleboel donuts
(*a bakery with a rack with a lot of donuts*)



disfluency

een foto van een straat met een aantal vogels (<i>a photo of a street with a number of birds</i>)
uh uh uh uh met een aantal vogels (<i>uh uh uh uh with some birds</i>)

R: uh allemaal duiven
(*uh all [full of] pigeons*)

Visual Question Answering

- Answer questions about images
- Multimodal input: Image & Question
- Commonly tackled as a classification
- VQA dataset: around 600k image-question pairs

Who is wearing glasses?

man

woman



Is the umbrella upside down?

yes

no



Task-specific models: VQA

Specifically designed for and trained on the VQA task

Again, this basic architecture can be enhanced with R-CNN region features, attention, etc.

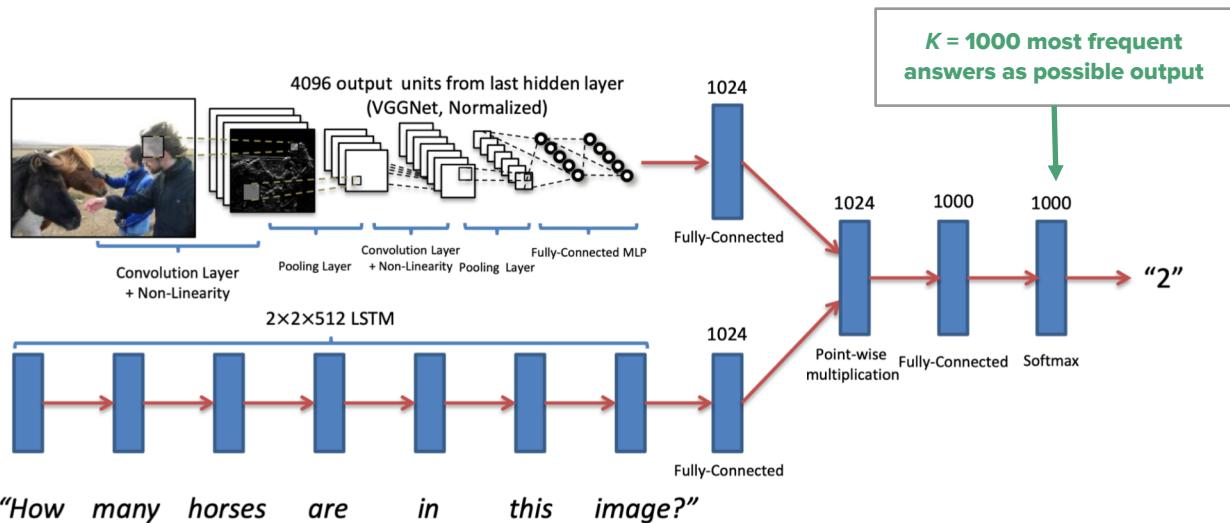


Fig. 8: Our best performing model (deeper LSTM Q + norm I). This model uses a two layer LSTM to encode the questions and the last hidden layer of VGGNet [48] to encode the images. The image features are then ℓ_2 normalized. Both the question and image features are transformed to a common space and fused via element-wise multiplication, which is then passed through a fully connected layer followed by a softmax layer to obtain a distribution over answers.

A few other tasks can be seen as extensions
of VQA or image description

VIST: visual storytelling

- 5 images from the same Flickr album (around 20k sequences in total)
- Crowdsourced stories: one sentence per image; several stories per image sequence

	1	2	3	4	5
					
Image description in isolation	A black frisbee is sitting on top of a roof.	A man playing soccer outside of a white house with a red door.	The boy is throwing a soccer ball by the red door.	A soccer ball is over a roof by a frisbee in a rain gutter.	Two balls and a frisbee are on top of a roof.
Sequence description (story)	A discus got stuck up on the roof.	Why not try getting it down with a soccer ball?	Up the soccer ball goes.	It didn't work so we tried a volley ball.	Now the discus, soccer ball, and volleyball are all stuck on the roof.

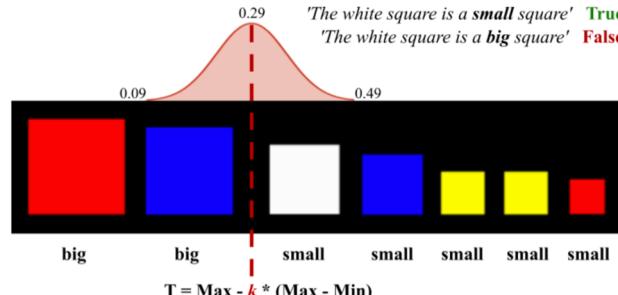
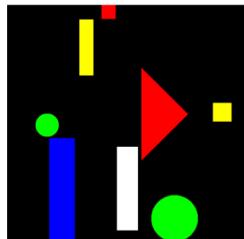
Visual reasoning

Binary classification task that requires reasoning over several images and a sentence:

- NLVR: synthetic images
- NLVR2: natural photographs

Or reasoning about context-dependent meanings

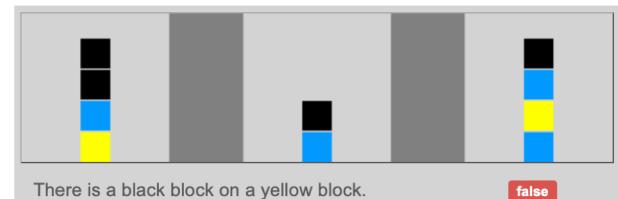
- MALeViC: modeling adjectives leveraging visual contexts



<https://github.com/sandropezzelle/malevic>

(Pezzelle & Fernández, EMNLP 2019)

<https://lil.nlp.cornell.edu/nlvr/>



The left image contains twice the number of dogs as the right image, and at least two dogs in total are standing.

(Suhr et al, ACL 2017; 2019)

Multi-turn interactions

Visual Dialogue <https://visuallialog.org>
(Das et al., CVPR 2017)

Image + Caption

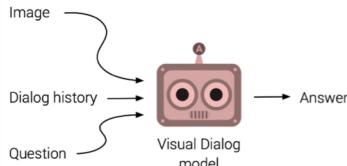


Human-Human dialog [4]

What are their genders?	1 man 1 woman
Are they both adults?	Yes
Do they wear goggles?	Looks like sunglasses
Do they have hats on?	Man does
Are there any other people?	No
What color is man's hat?	Black
Is it snowing now?	No
What is woman wearing?	Blue jacket and black pants
Are they smiling?	Yes
Do you see trees?	Yes

Tasks:

- Question answering



C: A dog with goggles is in a motorcycle side car.
Q: Is motorcycle moving or still?
A: It's parked
Q: What kind of dog is it?
A: Looks like beautiful pit bull mix
Q: What color is it?

- Image retrieval

GuessWhat

<https://github.com/GuessWhatGame/guesswhat>
(De Vries et al., CVPR 2017)



Questioner

- Is it a vase?
Is it partially visible?
Is it in the left corner?
Is it the turquoise and purple one?

Oracle

- Yes
No
No
Yes

Tasks:

- Asked informative questions
- Locate the target object, given image and dialogue history

Multi-turn interactions

PhotoBook dataset: more natural visually grounded dialogue

Participant A



Round 1 of 5

A: Hi

B: Hello

B: do you have a white cake on multi colored striped cloth?

A: I see a guy taking a picture. What about you?

B: is it of a cake with construction trucks on it?

A: Yeah. I don't see the cake you mentioned.

A: <common img_2>

Participant B



Repeated references to the same image



1 I see a guy taking a picture. What about you?

2 guy with camera

3 I have the guy with camera

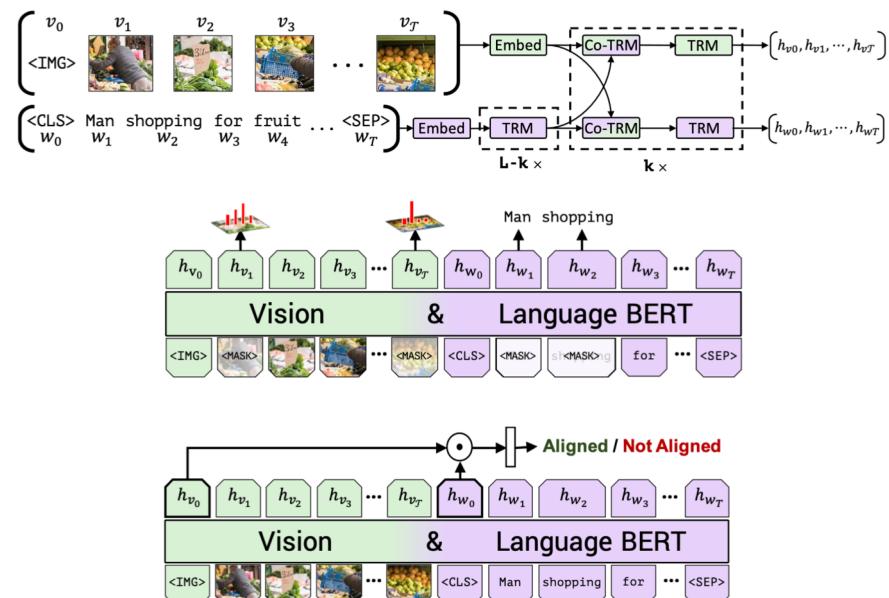
4 the last one is the camera guy

General purpose multimodal models

- Self-supervised learning
 - Multimodal encoders
 - Cross-modal alignment
 - Generative multimodal models
 - Most recently, datasets harvested from the Internet
-

ViLBERT

- Initialized from BERT
- Visual features extracted from 10-36 regions using Faster-RCNN
- Pretrained on Conceptual Captions
 - Masked Language Modelling
 - Masked Region Classification
 - Image-Text Matching

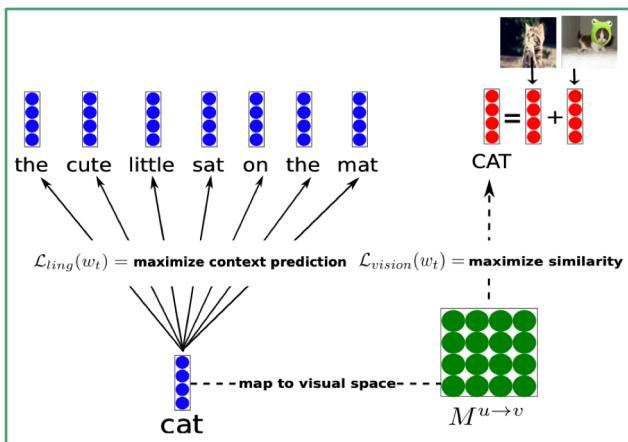


Lu et al. (2019). ViLBERT: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. NeurIPS.

Other multimodal encoders: LXMERT (Tan & Bansal, 2019), UNITER (Chen et al. 2019), VILT (Kim et al, 2021), etc.

Enriching representations of text LMs with visual grounding

Word2vec: word type embeddings



Lazaridou et al. (2015) Combining language and vision with a multimodal Skip-gram model, NAACL.

BERT: contextualized word embeddings

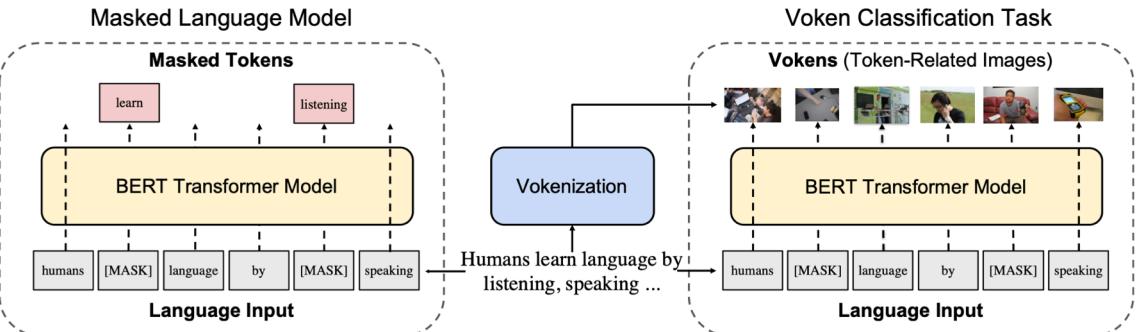


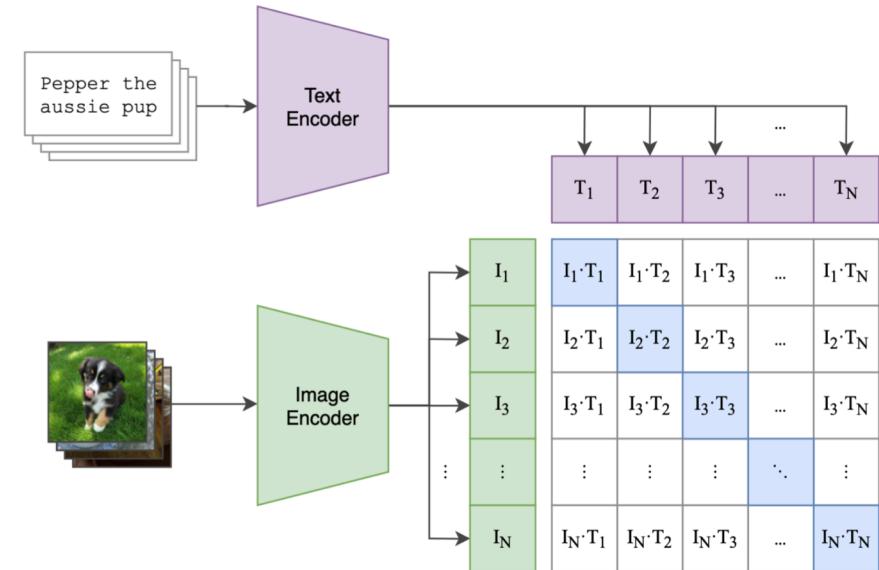
Figure 2: Illustration of the BERT transformer model trained with a visually-supervised language model with two objectives: masked language model (on the left) and voken classification (on the right). The first objective (used in original BERT pre-training) predicts the masked tokens as self-supervision while the second objective predicts the corresponding vokens (contextually generated by our vokenization process) as external visual supervision. Since the inputs are the same, we optimize the two objectives simultaneously and share the model weights.

Tan & Bansal (2020). Vokenization: Improving Language Understanding with Contextualized, Visual-Grounded Supervision. EMNLP.

CLIP: contrastive language-Image pretraining (OpenAI)

The backbone consists of two separate components:

- visual encoder: ViT or ResNet (CNN)
- language encoder: GPT
- Maximize the similarity of the embeddings of paired examples (I, T).
- Huge pretraining dataset of unclear provenance



Generative multimodal models

- A bunch of multimodal models exploiting pre-trained language-only and vision-only models
- They take as input a vision and language prompt, and generate text
 - LLM (e.g., LLAMA)
 - Vision encoder (e.g., CLIP's ViT / OpenClip)
 - Cross-attention layers over visual tokens, interleaved with the layers of a frozen LLM
 - Next token prediction objective
 - Can be used zero/few-shot
- Close models like Flamingo (by Deep Mind), with unknown training data
- Open models like IDEFICS, with public data



Large scale general-purpose datasets

Trend towards HUGE general purpose datasets used for model pretraining: Data scraped from the Internet – images aligned with alt-text. For example:

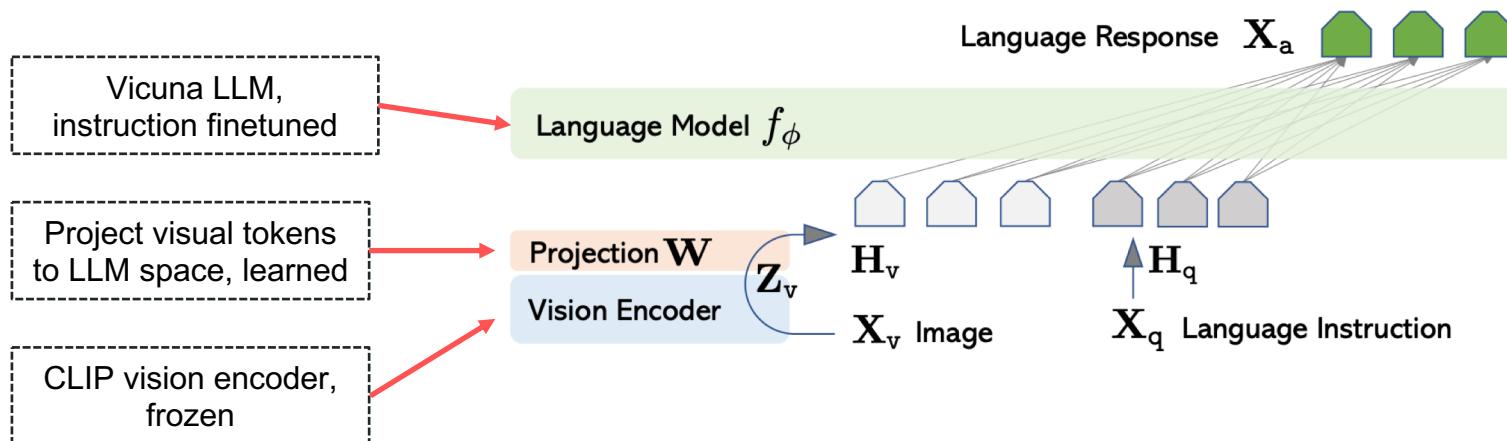
- Conceptual Captions: 3/12M images with (filtered) alt-text
- Public Multimodal Dataset: 70M pairs from existing datasets and other sources

COCO	Visual Genome	SBU captions	Localized narratives	WIT	RedCaps	CC12M	YFCC filtered
 A close up view of a pizza sitting on a table with a soda in the back.	 a lenovo laptop rebooting	 Front view of basket 13, from the sidewalk in front of the basket.	 The woman is touching a utensil in front of her on the grill stand.	 Typocerus balteatus, Subfamily: Flower Longhorns	 Deigdoh falls in india	 Jumping girl in a green summer dress stock illustration	 In the kitchen at the Muse Nissim de Camondo

- LAION-5B / LAION-400B: harvested from CommonCrawl, a dump of the Internet with more than 300TB of stuff.

LLaVA: Large Vision and Language Assistant

- **Instruction-tuning** extended to the multimodal space
 - Until then, applied only to text LLMs
- Architecture: CLIP visual encoder and Vicuna LLM language decoder
 - Both of them pretrained, then finetuned on **multimodal instruction following data**



- Data: Multimodal instruction following data generated by ChatGPT/GPT4

Example of instruction data for finetuning,
generated by text-only GPT models

Context type 1: Captions

A group of people standing outside of a black vehicle with various luggage.

Luggage surrounds a vehicle in an underground parking area

People try to fit all of their luggage in an SUV.

The sport utility vehicle is parked in the public garage, being packed for a trip

Some people with luggage near a van that is transporting it.



Context type 2: Boxes

person: [0.681, 0.242, 0.774, 0.694], backpack: [0.384, 0.696, 0.485, 0.914], suitcase: ...<omitted>

Response type 1: conversation

Question: What type of vehicle is featured in the image?

Answer: The image features a black sport utility vehicle (SUV) ...<omitted>

Response type 2: detailed description

The image is an underground parking area with a black sport utility vehicle (SUV) parked. There are three people in the scene, with one person standing closer to the left side of the vehicle, another person in the middle, and the third person on the right side. They are all working together to pack their luggage into the SUV for a trip. ...<omitted>

Response type 3: complex reasoning

Question: What challenges do these people face?

Answer: In the image, a group of people is standing outside a black SUV in a parking area, surrounded by various pieces of luggage, including suitcases and backpacks. They are facing the challenge of fitting all their luggage into the black SUV. There are multiple suitcases and backpacks to be packed, which suggests that the group has a significant amount of belongings ...<omitted>

Prompt used to generate the
instruction following data

```
messages = [ {"role": "system", "content": f"""You are an AI visual assistant, and you are seeing a single image. What you see are provided with five sentences, describing the same image you are looking at. Answer all questions as you are seeing the image. """}
```

```
Design a conversation between you and a person asking about this photo. The answers should be in a tone that a visual AI assistant is seeing the image and answering the question. Ask diverse questions and give corresponding answers.
```

```
Include questions asking about the visual content of the image, including the object types, counting the objects, object actions, object locations, relative positions between objects, etc. Only include questions that have definite answers:
```

- (1) one can see the content in the image that the question asks about and can answer confidently;
- (2) one can determine confidently from the image that it is not in the image. Do not ask any question that cannot be answered confidently.

```
Also include complex questions that are relevant to the content in the image, for example, asking about background knowledge of the objects in the image, asking to discuss about events happening in the image, etc. Again, do not ask about uncertain details. Provide detailed answers when answering complex questions. For example, give detailed examples or reasoning steps to make the content more convincing and well-organized. You can include multiple paragraphs if necessary."""}
```

```
]
for sample in fewshot_samples:
    messages.append({"role": "user", "content": sample['context']})
    messages.append({"role": "assistant", "content": sample['response']})
    messages.append({"role": "user", "content": '\n'.join(query)})
```

Table 1: One example to illustrate the instruction-following data. The top block shows the contexts such as captions and boxes used to prompt GPT, and the bottom block shows the three types of responses. Note that the visual image is not used to prompt GPT, we only show it here as a reference.



User

What are the meals that I can cook with these?



With the variety of food items stored in the refrigerator, you can prepare several meals. For example, you can create a fruit salad using the strawberries, blueberries, and carrots. Additionally, you can make a delicious smoothie using the fruits and vegetables, which would be a healthy and nutritious snack or breakfast option. The milk and eggs can be used to create various dishes, such as a milk-based smoothie, a baked casserole, or scrambled eggs. The juice can also be used to make a fresh fruit and juice mix or as a refreshing homemade popsicle. The possibilities are vast, and the meals you can create will depend on your preferences and dietary needs.

I would like to do the fruit salad. Show me the recipe.



Certainly! Here's a simple fruit salad recipe using the items in your refrigerator:

Ingredients:

- 2 cups fresh strawberries, sliced
- 1 cup fresh blueberries
- 1 large carrot, grated
- Juice of half a lemon
- A handful of chopped fresh parsley or mint
- A few ice cubes
- Optional: honey or maple syrup (adjust to taste)

Instructions:

1. In a large bowl, combine the sliced strawberries, blueberries, and carrot. Gently toss them together to mix.
2. Drizzle the juice of half a lemon over the fruit and vegetables to prevent them from oxidizing and add a refreshing citrus flavor.
3. Sprinkle the chopped fresh parsley or mint over the fruit and gently toss once more.
4. If desired, add a drizzle of honey or maple syrup for a touch of sweetness. Adjust the amount according to your taste preferences.
5. Cover the bowl and refrigerate the fruit salad for at least 30 minutes or up to a few hours. This allows the flavors to meld and creates a more refreshing salad.
6. Serve the fruit salad in cups or bowls, garnished with a sprinkle of fresh parsley or mint, if desired.

This fruit salad recipe is perfect for a healthy snack or side dish, and with the variety of ingredients available in your refrigerator, it should be easy to put together quickly.



User

Wiki-LLaVA: Hierarchical RAG for Multimodal LLMs

First approach to integrating an external knowledge source into multimodal generative LLMs

- Retrieval module with two steps:
 1. Retrieve documents via CLIP similarity of input image and document titles
 2. Retrieve relevant passages via embedding similarity of the input question and document chunks
- Enrich input context with the retrieved passages

```
<IMAGE>\nGiven the following context:\n<R1>\n<R2>\n<R3>\n <QUESTION>\n\nGive a short answer. ASSISTANT:
```

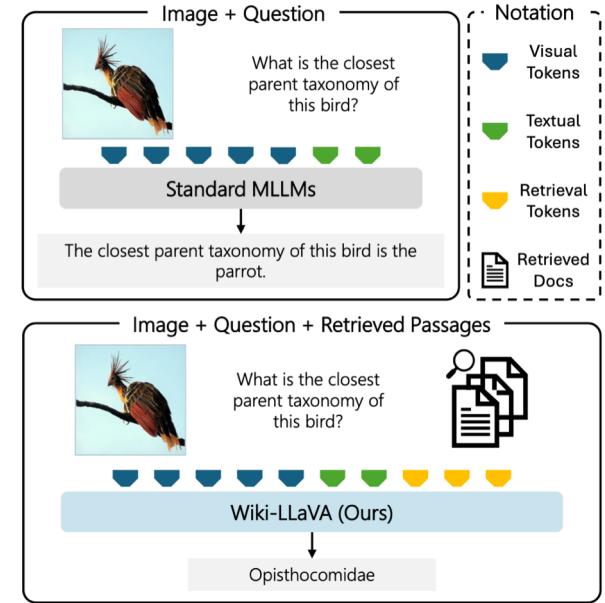


Figure 1. Comparison between a standard multimodal LLM and Wiki-LLaVa. Our model integrates knowledge retrieved from an external knowledge base of documents through a hierarchical retrieval pipeline. As a result, it provides more precise answers when tasked with questions that require external knowledge.

Wiki-LLaVA: Hierarchical RAG for Multimodal LLMs



In what state is this building located?

LLaVA-1.5:
California ✗
Wiki-LLaVA:
Washington ✓



When was this building constructed?

LLaVA-1.5:
1970 ✗
Wiki-LLaVA:
1927 ✓



Which geographic area is this fish found?

LLaVA-1.5:
Gulf of Mexico ✗
Wiki-LLaVA:
Brazil ✓



What is the oldest age of this animal?

LLaVA-1.5:
10 years ✗
Wiki-LLaVA:
24.9 ✓



Which culture is associated with this place?

Ancient Greek

LLaVA-1.5:
Roman ✗
Wiki-LLaVA:
Nuragic Civilization ✗



What is the name of the main club of this stadium?

FC Rotor

LLaVA-1.5:
Real Madrid ✗
Wiki-LLaVA:
FC Dynamo Kyiv ✗

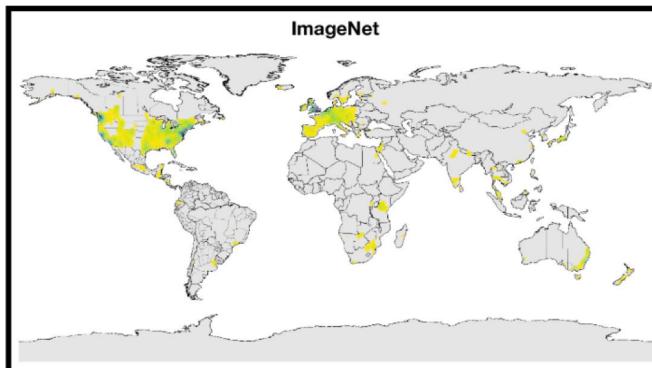
Ethical issues
surrounding
multimodal data

Ethical issues

- Data provenance
 - Very often, data is scraped from the internet with limited information on license or copyright
 - The data may be public, but using such data without legal authorization may infringe regulations
- Data diversity
 - Is the data representative of the population it intends to depict or serve?
- Data quality
 - Does the dataset construction process lead to perpetuating harmful biases?

Data diversity

- Datasets are mostly in English, or a few Indo-European languages
- Some datasets are translated from English
- The image sources mostly reflect North American and European cultures
- Some concepts are most immediately understood within a cultural background



Density map of geographical distribution of images in ImageNet (DeVries+, 2019)



ENG: An unusual looking vehicle ...

NLD: Een mobiel draaiorgel ...

Example from van Miltenburg+ 2017

MaRVL: Multicultural Reasoning over Vision and Language



Representative of annotators'
cultures



5 typologically diverse languages
Independent, culture-specific annotations



MaRVL-id Bola basket



MaRVL-sw Mpira wa kikapu



MaRVL-tr Basketbol

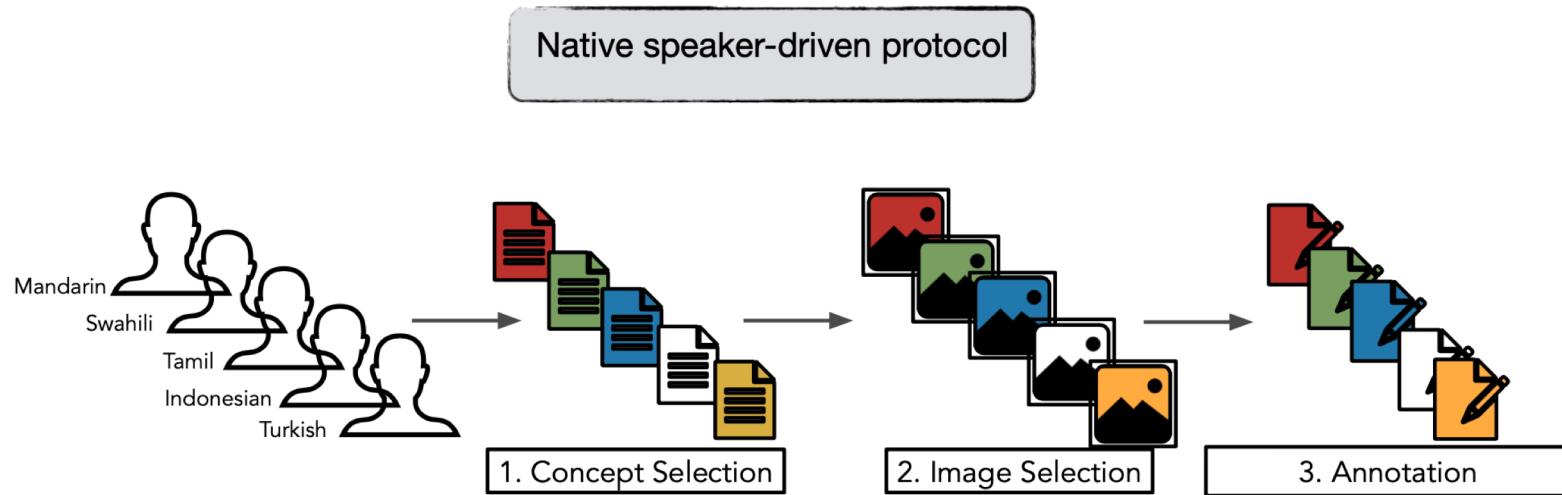


MaRVL-zh 篮球



MaRVL-ta കൂട്ടെപ്പബ്ന്താട്ടമ്

MaRV dataset construction



Visual reasoning task (Suhr et al. ACL 2019)

- two images paired with a sentence
- Predict whether the sentence is a true description of the pair

MaRVL examples

Visual reasoning task (Suhr et al. ACL 2019)

- two images paired with a sentence
- Predict whether the sentence is a true description of the pair

MaRVL-tr Kanun (çalgı)



Görsellerden birinde dizlerinde kanun bulunan birden çok insan var

(In one of the images, there are multiple people with qanuns on their knees)

Label: True

MaRVL-ta தை (Vada)



இரண்டு படங்களிலும் நிறைய மசால் வடைகள் உள்ளன

(Both images contain a lot of masala vadas)

Label: False

VizWiz-VQA: questions asked by people who are blind

- Blind people taking photos and recording a spoken question about them
- 10 crowdsourced answers per visual question
- Tasks:
 - Predict the answer to a question
 - Predict whether a question cannot be answered



Q: Does this foundation have any sunscreen?

A: yes



Q: What is this?

A: 10 euros



Q: What color is this?

A: green



Q: Please can you tell me what this item is?

A: butternut squash red pepper soup



Q: Is it sunny outside?

A: yes



Q: Is this air conditioner on fan, dehumidifier, or air conditioning?

A: air conditioning

Data quality

- Large models require huge datasets for training (e.g., LAION)
- Data scaling makes proper data curation extremely difficult
- As a result, models are often trained on data of very dubious quality with serious ethical implications.



(Eileen Collins, American astronaut)

$$\cos(\cdot, \cdot)$$

0.276

This is a portrait of an astronaut
with the American flag

0.308

This is a photograph of a smiling
housewife in an orange jumpsuit with
the American flag

Evaluation

Different types of evaluation

Extrinsic

- Evaluating the performance of models on multimodal downstream tasks

Intrinsic

- Evaluating the quality of the representations learned by the models (e.g., in terms of human likeness)
- Assessing the skills of pretrained models through challenging test sets

Extrinsic: Classic reference-based evaluation

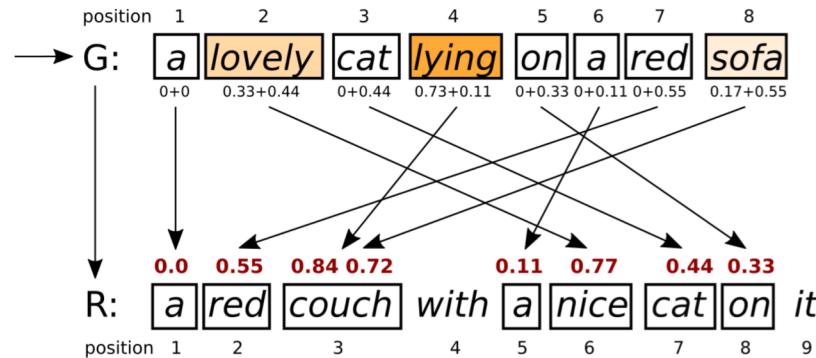
NLG metrics that rely on comparing generated text to a reference gold standard are often used for tasks such as image captioning or visual storytelling

- BLEU, ROUGE, CIDEr, METEOR

It is well known that these metrics are problematic:

Extrinsic: Classic reference-based evaluation

- The same message can be conveyed in very different ways!
 - **BERTScore** aims to address this (semantic similarity)
 - **SSD** (Semantic and Sequential Distance) extends this to take into account word order – important for models aiming to capture sequential cross-modal alignment (guided by gaze)



- These metrics only consider the language modality, ignoring the image
 - **CLIPScore** aims to address this

Visual Storytelling

Input: sequence of images



Task: to generate a textual story consistent with the input

Human-annotated story: *It's parade day, and the whole town turns out to watch. There are those who serve our country, and the crowds cheer. There are the bands, and the music is loud but thankfully well performed. The flags are always fun to watch. And of course you get the old cars and their owners traveling through.*

Challenge: plausibility of several creative stories for a single given image sequence, makes reference-based NLG metrics (e.g., METEOR) inappropriate for the task.

Visual Storytelling

Visual storytelling requires more evaluation dimensions (Wang et al. 2022; Surikuchi et al. 2023, 2024)

- **Coherence:** LM probability of the next sentence given the context
- **Degree of repetition:** Jaccard similarity between context and next sentence
- **Visual grounding:** CLIP-based cosine similarity between noun phrases and object bounding boxes, weighted by noun concreteness.

Distance between humans and models

$$\begin{aligned} d_{HM}^C &= |C_H - C_M|, & d_{HM} &= (d_{HM}^C + d_{HM}^G + d_{HM}^R)/3 \\ d_{HM}^G &= |G_H - G_M|, \\ d_{HM}^R &= |R_H - R_M| \end{aligned}$$

Visual Question Answering

VQA has traditionally been operationalized as a classification task, evaluated with accuracy.

Does the visual information matter to perform the task?

- VQA dataset: around 600k image-question pairs
 - Imbalances: e.g., 41% of questions starting with “What sport is...” have “tennis” as the correct answer
- VQA.v2 dataset: 1.1M image–question pairs with balanced distribution of answers

Who is wearing glasses?

man

woman



Is the umbrella upside down?

yes

no



Evaluation gone wrong: Ad understanding as an example

- Capabilities of models may be overestimated if evaluation setups are not properly designed.
- Consider the task of image-based advertisement understanding



Ads are interesting because they may contain atypical objects and figurative language, plus they are present in our daily lives.

Ad understanding evaluation

The Pitt Ads Retrieval Task (Hussain et al., CVPR, 2017)

- Positive explanations are crowdsourced
- Negative explanations are randomly sampled from different ads

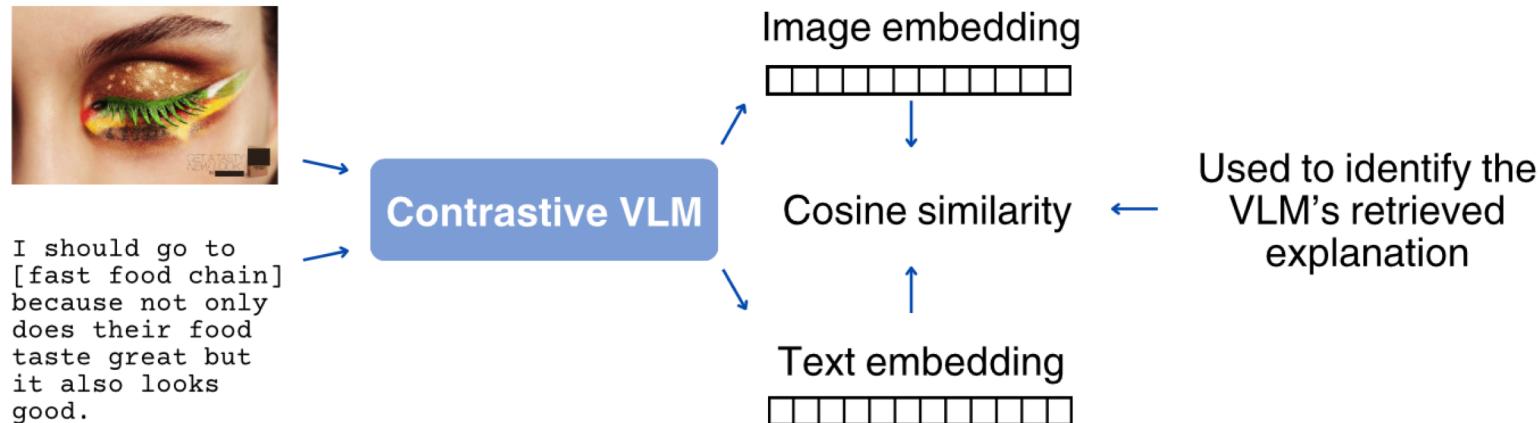


Positive explanation: *I should go to [fast food chain] because not only does their food taste great but it also looks good.*

Negative explanation: I should buy this skateboard because it is strong.

Ad understanding evaluation

Jia et al., *ACL*, 2023



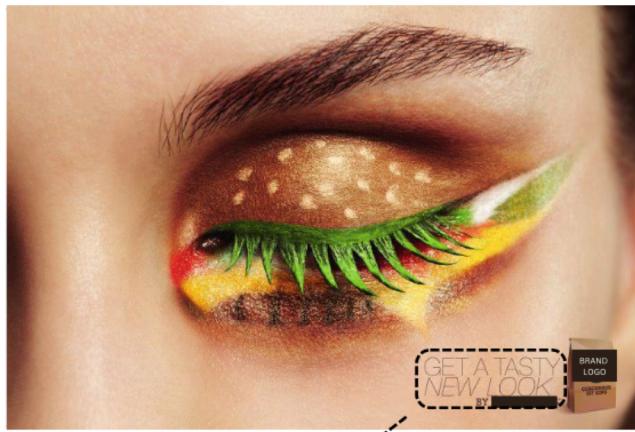
Highest zero-shot accuracy: **95.2%**

Ad understanding evaluation

TRADE: Truly adversarial AD-understanding Evaluation (Bavaresco et al., ACL 2024)

Features:

- 300 ads from the Pitt Ads Dataset
- Carefully curated negatives by in-house annotators



Positive:

I should go to [fast food chain] because not only does their food taste great but it also looks good.

Negatives:

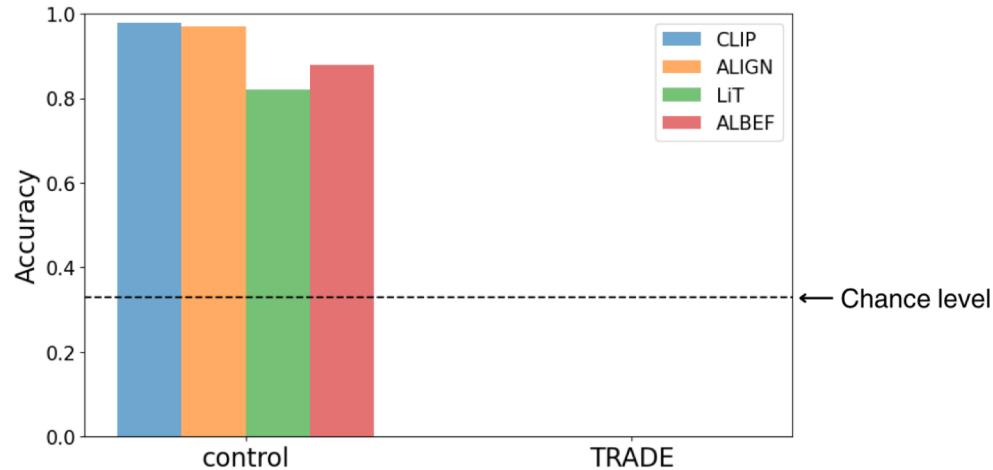
I should go to [fast food chain] because my eyelashes need a new look.

I should go to [fast food chain] because tasty burgers must look like these eyelashes.

Ad understanding evaluation

TRADE: Truly adversarial AD-understanding Evaluation (Bavaresco et al., ACL 2024)

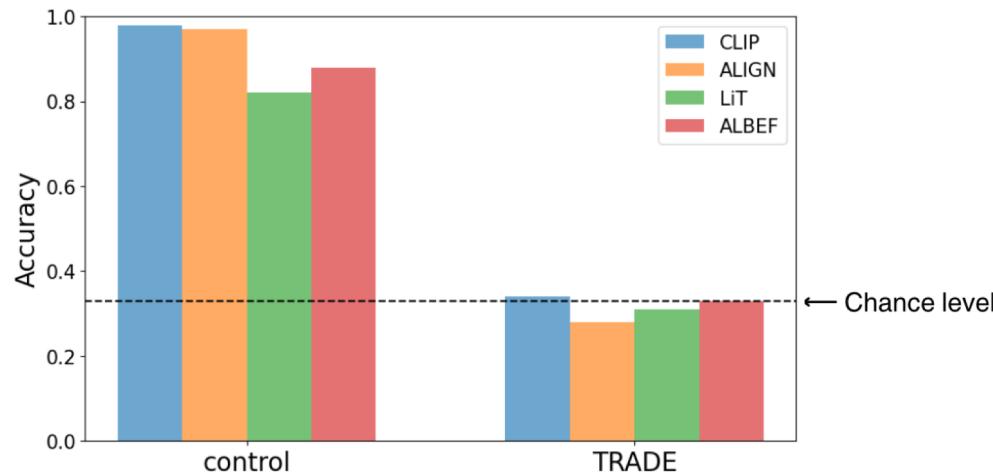
Contrastive VLMs performance on TRADE



Ad understanding evaluation

TRADE: Truly adversarial AD-understanding Evaluation (Bavaresco et al., ACL 2024)

Contrastive VLMs performance on TRADE



Ad understanding evaluation

TRADE: Truly adversarial AD-understanding Evaluation (Bavaresco et al., ACL 2024)

Examples of selected explanations

TRADE can't be solve by
exploiting grounding

Evaluating ad understanding
calls for operationalizations
beyond retrieval



🔍 GET A NEW TASTY
LOOK. By [fast food chain]

Positive

I should go to [fast food chain]
because not only does their food
taste great but it also looks
good.

→ Humans

Negatives

I should go to [fast food chain]
because my eyelashes need a
new look.

→ CLIP, ALIGN

I should go to [fast food chain]
because tasty burgers must look
like these eyelashes.

→ ALBEF, LiT

In sum, no automatic metric is perfect for
image-condition text generation tasks.

Human evaluation remains key.

Intrinsic evaluation

- By grounding language into vision, arguably pre-trained multimodal models have an advantage over text-only models.
 - They lead to better performance in downstream tasks.
 - However, task performance conflates different sources of error and does not allow for analysis of different skills.
-
- Do pre-trained general purpose multimodal models learn representations that better align with human multimodal knowledge?

Intrinsic evaluation

Representational quality: Similarity with human representations

- Explicit human similarity judgements
- Behavioural human cognitive signals such as eye-tracking
- Human brain signals

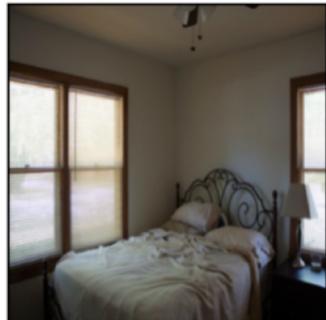
Model skills: What skills have general purpose multimodal models acquired and where do they fail?

- Developing challenge datasets that target specific skills is a vibrant area of research

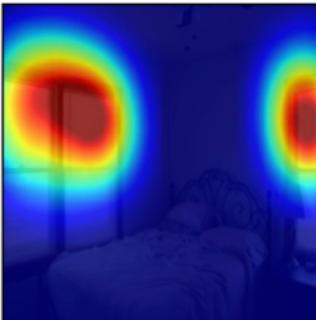
Intrinsic evaluation of task-based representations

Comparing attention patterns in VQA

Where humans look

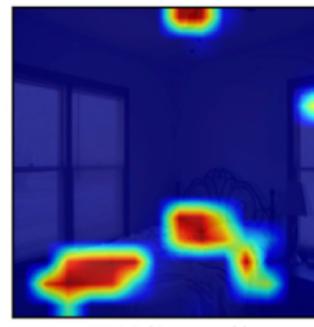


What is covering the windows? blinds

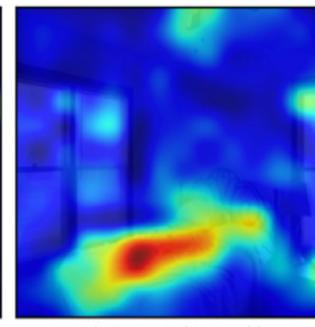


Human Attention

Where machines look



SAN-2 (Yang et al.)
Correlation: -0.495



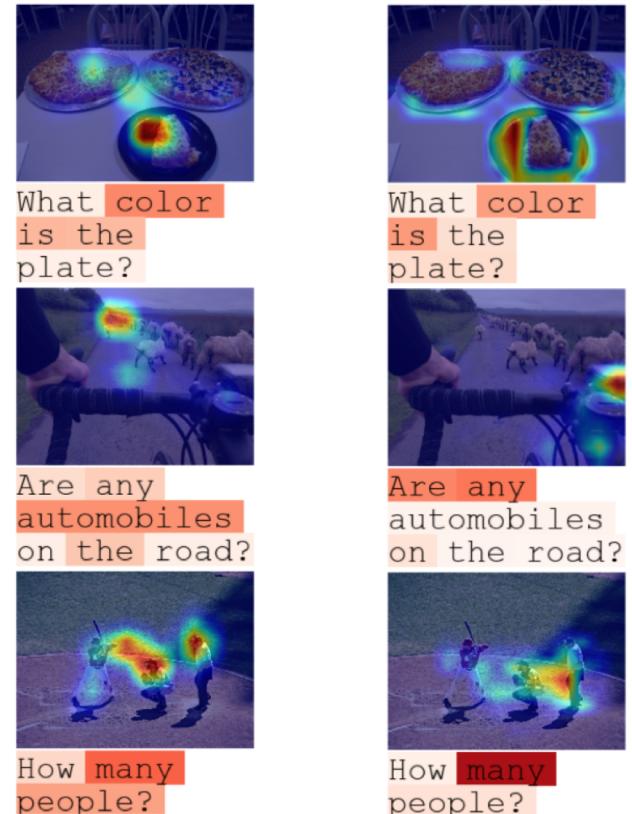
HieCoAtt-Q (Lu et al.)
Correlation: -0.440

- Low correlation between human and machine attention: 0.256
- Inter-human correlation: 0.623

Intrinsic evaluation of task-based representations

Comparing attention patterns in VQA

- Higher correlation with visual and text attention is a significant predictor of VQA performance



Sood et al. (2021).

Semantic similarity judgements

- Common evaluation of text-only embeddings (e.g., static vs. contextualized)

Semantic similarity:

man, person: **similar**

dog, airplane: **dissimilar**

- *donut, muffin* = 0.8
- *car, train* = 0.5
- *dog, airplane* = 0.1
- ...

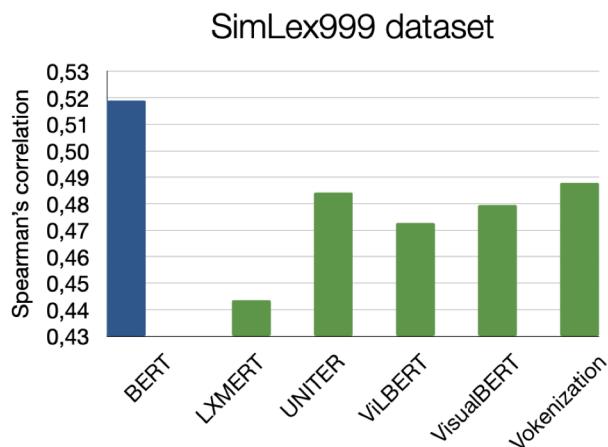
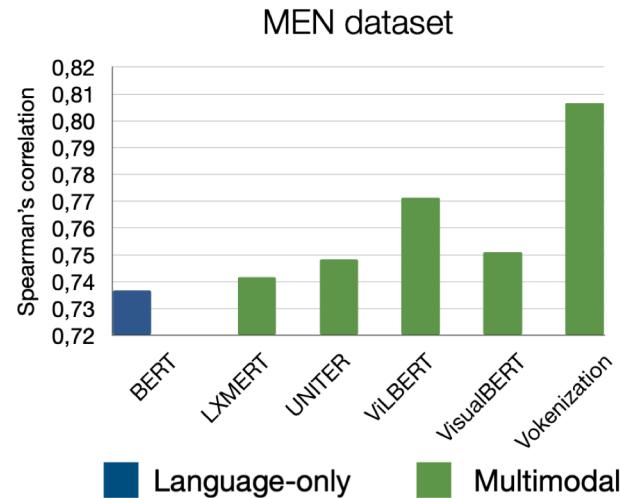
Spearman's correlation between:

- **human semantic similarity judgements** and
- **cosine similarity** between pairs of **model word representations**

Comparison of the
semantic spaces

Semantic similarity judgements

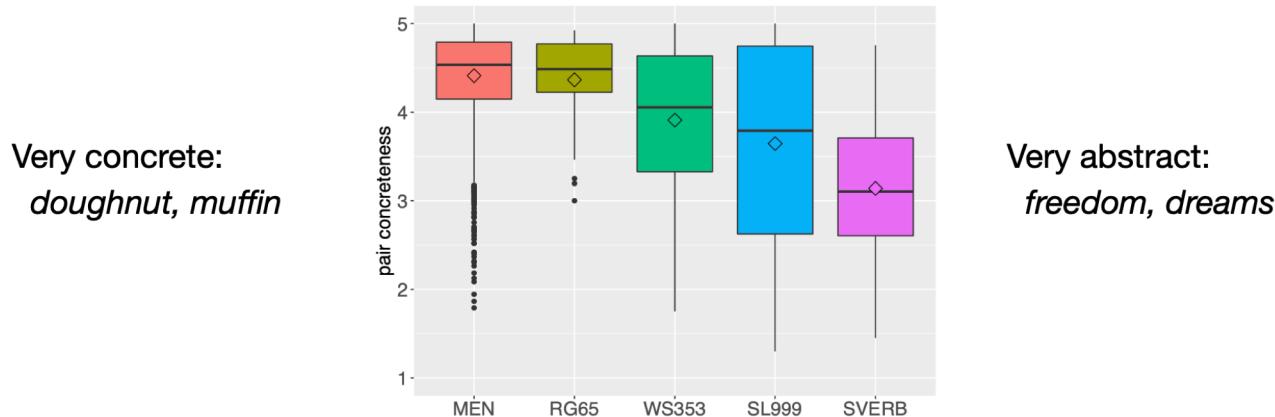
- Common evaluation of text-only embeddings (e.g., static vs. contextualized)



Semantic similarity judgements

- Common evaluation of text-only embeddings (e.g., static vs. contextualized)

The level of concreteness of the words being judged varies per dataset



Multimodal models are better than text-only ones at approximating similarity judgements of concrete words

Intrinsic evaluation

Representational quality: Similarity with human representations

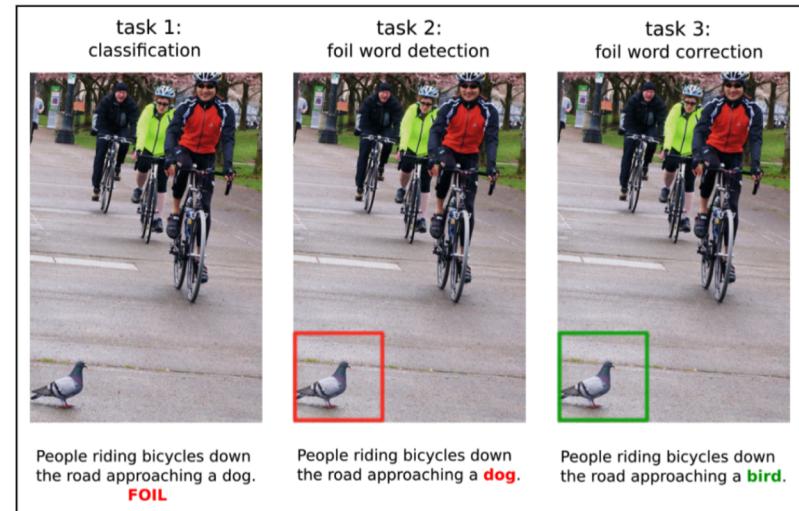
- Explicit human similarity judgements
- Behavioural human cognitive signals such as eye-tracking
- Human brain signals

Model skills: What skills have general purpose multimodal models acquired and where do they fail?

- Developing challenge datasets that target specific skills is a vibrant area of research

FOIL captions

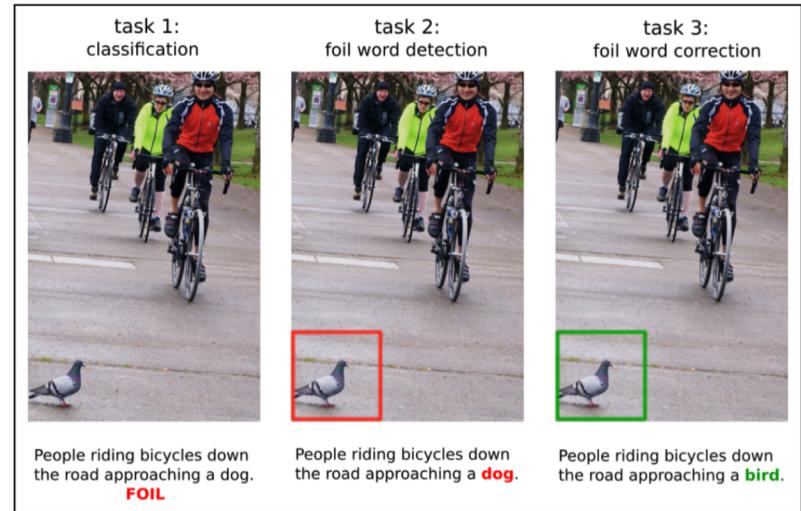
- Do V&L models really understand the relationship between words and images?
- Crowdsource datasets that contain contextually plausible but incorrect image–text pairs, focusing on nouns.



FOIL captions

T1: Classification task			
	Overall	Correct	Foil
Blind	55.62	86.20	25.04
CNN+LSTM	61.07	89.16	32.98
IC-Wang	42.21	38.98	45.44
LSTM + norm I	63.26	92.02	34.51
HieCoAtt	64.14	91.89	36.38
Human (<i>majority</i>)	92.89	91.24	94.52
Human (<i>unanimity</i>)	76.32	73.73	78.90

- Very challenging at the time, but has since been essentially solved.
- It's a good sanity check!



Subject-Verb-Object Probes

- SVO-Probes: subject-verb-object sentences, with focus on **verbs**
- Models largely fail to distinguish images with fine-grained **verb** differences
- Accuracy below chance on negative pairs
- Verb understanding is harder than subject or object understanding

A person *sings* at a concert.

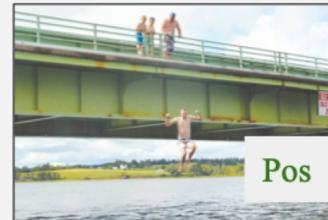


person, sing, concert



person, dance, concert

A man *jumping* into a river.



man, jump, river



man, kayak, river

Winoground

- 1,600 text-image pairs to evaluate compositional understanding
- Images sourced with permission from Getty
- Models struggle, often performing below chance



(a) some plants surrounding a lightbulb



(b) a lightbulb surrounding some plants

BLA: Basic Language Abilities

Focus on simple construction that preschool children can understand



Active-Passive voice

- T:** the **woman** feeds the **man**.
T: the **man** is fed by the **woman**.
F: the **man** feeds the **woman**.
F: the **woman** is fed by the **man**.



Coordination

- T:** the **man** wears a wetsuit and carries a surfboard.
T: the **woman** wears a red bikini and rides a red bike.
F: the **man** wears a wetsuit and rides a red bike.
F: the **woman** carries a surfboard and wears a red bikini.



Relative Clause

- T:** the **man** who wears a gray polo holds a stuffed bear.
T: the **man** who wears a striped shirt holds a cow.
F: the **man** who wears a gray polo holds a cow.
F: the **man** who wears a striped shirt holds a stuffed bear.

VALSE Benchmark

Evaluation of multiple linguistic phenomena

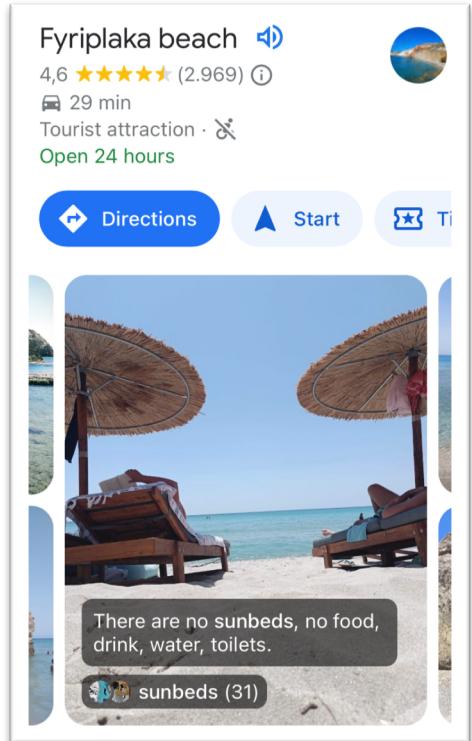
existence	plurality	counting	relations	actions	coreference
<i>There are no animals / animals shown.</i>	A small copper vase with some flowers / exactly one flower in it.	There are <i>four / six</i> zebras.	A cat plays with a pocket knife <i>on / underneath</i> a table.	A <i>man / woman</i> shouts at a <i>woman / man</i> .	Buffalos walk along grass. Are they in a zoo? <i>No / Yes</i> .



Metric	Model	Avg.
	Random	50.0
	GPT1*	60.7
	GPT2*	60.1
	CLIP	64.0
acc_r	LXMERT	59.6
	ViLBERT	63.7
	12-in-1	75.1
	VisualBERT	<u>46.4</u>

$$p(\text{caption}, \text{img}) > p(\text{foil}, \text{img})$$

- Lexical semantics is reasonably robust
- Multimodal compositionality is far more challenging



LLaVA: Large Vision and Language Assistant

- How are the newest instruction tuned multimodal models evaluated
- LLaVA evaluation: GPT4 as a judge, scoring answers for helpfulness, relevance and accuracy, together with an explanation. Also comparing to GPT4 answers.

Question:

Which material is this rocking chair made of?

Options: (A) wood (B) silk

LLaVA answer:

LECTURE: A material is a type of matter.

Wood, glass, metal, and plastic are common materials.

Some objects are made of just one material.

Most nails are made of metal.

Other objects are made of more than one material.

This hammer is made of metal and wood.

SOLUTION: Look at the picture of the rocking chair. The rocking chair is made of two different materials. The legs are made of wood, and the back and seat are made of silk.

The answer is B.



GPT-4 (text-only) answer:

The answer is A. BECAUSE: Rocking chairs are typically made of sturdy materials that can support the weight of a person and withstand the rocking motion. Wood is a common material used for making rocking chairs due to its strength, durability, and natural appearance.

GPT-4 (text-only) judge:

Since there is no context or image to refer to, I cannot independently analyze the rocking chair in question. However, I can try to analyze the reasoning given by the assistants.

Assistant 1, who claims to have seen the image, mentions that the legs of the rocking chair are made of wood, and the back and seat are made of silk, implying that it's a combination of two materials. However, this response seems less plausible, as silk is not a common material used for the construction of rocking chairs due to issues with stability and durability. Typically, one would expect the frame, legs, and seat or back to be made of wood, with cushions and padding made of fabrics, such as cotton or polyester.

Assistant 2 argues that rocking chairs are typically made of sturdy materials, and wood is a common material used for making rocking chairs due to its strength, durability, and natural appearance. This reasoning stands true and is based on common knowledge.

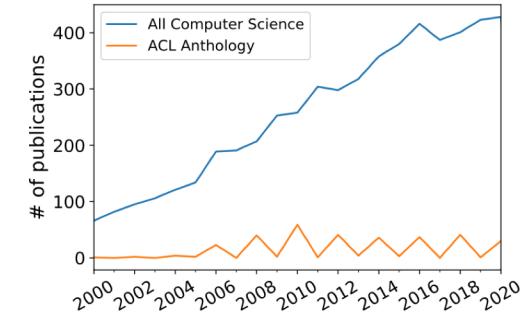
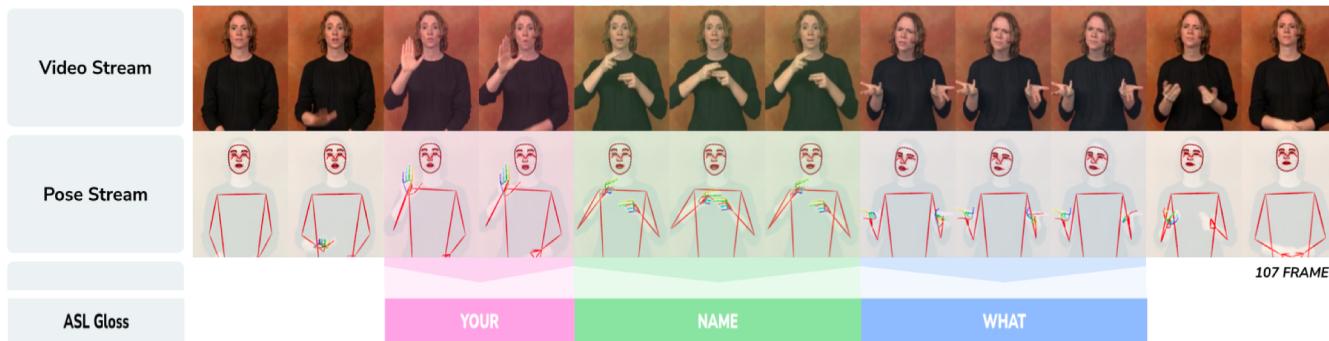
Given the information available, the most plausible answer is that the rocking chair is made of wood, as suggested by Assistant 1. The answer is A.

Table 10: One example on how the text-only GPT-4 acts as a judge to ensemble the predictions from LLaVA and a text-only GPT-4, and gives a correct final answer.

New directions

Sign language processing

Sign languages are the primary means of communication for many deaf and hard of hearing individuals.

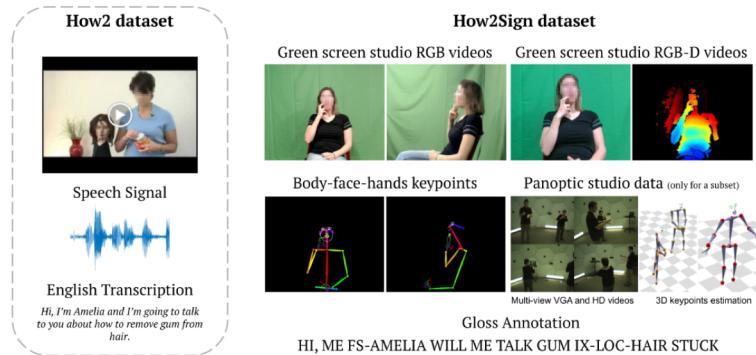


More work on SLP within computer vision, far less in NLP

Thanks to the rise of multimodal NLP, this is changing

More sign language datasets are being released

- BBS-Oxford British Sign Language dataset (Albanie et al. 2021): 1,400 hours of signed shows (factual, entertainment, drama, comedy, children's shows)
- How2Sign (<https://how2sign.github.io/>)



We introduce How2Sign, a multimodal and multiview continuous American Sign Language (ASL) dataset, consisting of a parallel corpus of more than 80 hours of sign language videos and a set of corresponding modalities including speech, English transcripts, and depth.

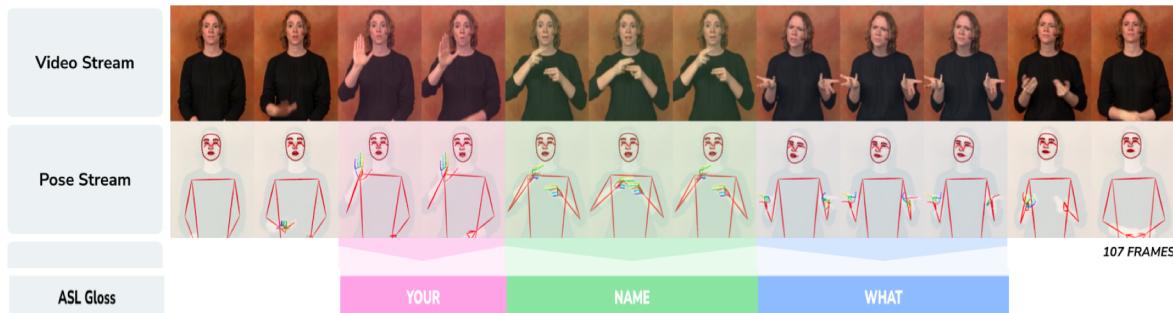
A three-hour subset was further recorded in the Panoptic studio enabling detailed 3D pose estimation.

This dataset is publicly available for research purposes only.

Sign language processing tasks

- Detection
 - Is sign language being used in a video?
- Identification
 - Which sign language is being used?
- Segmentation
 - Detecting boundaries of meaningful units

- Recognition
 - Recognizing which sign is being used
- Translation
 - From sign to spoken language (glosses)
- Production
 - From spoken to sign language (poses)



Any efforts must involve the Deaf community

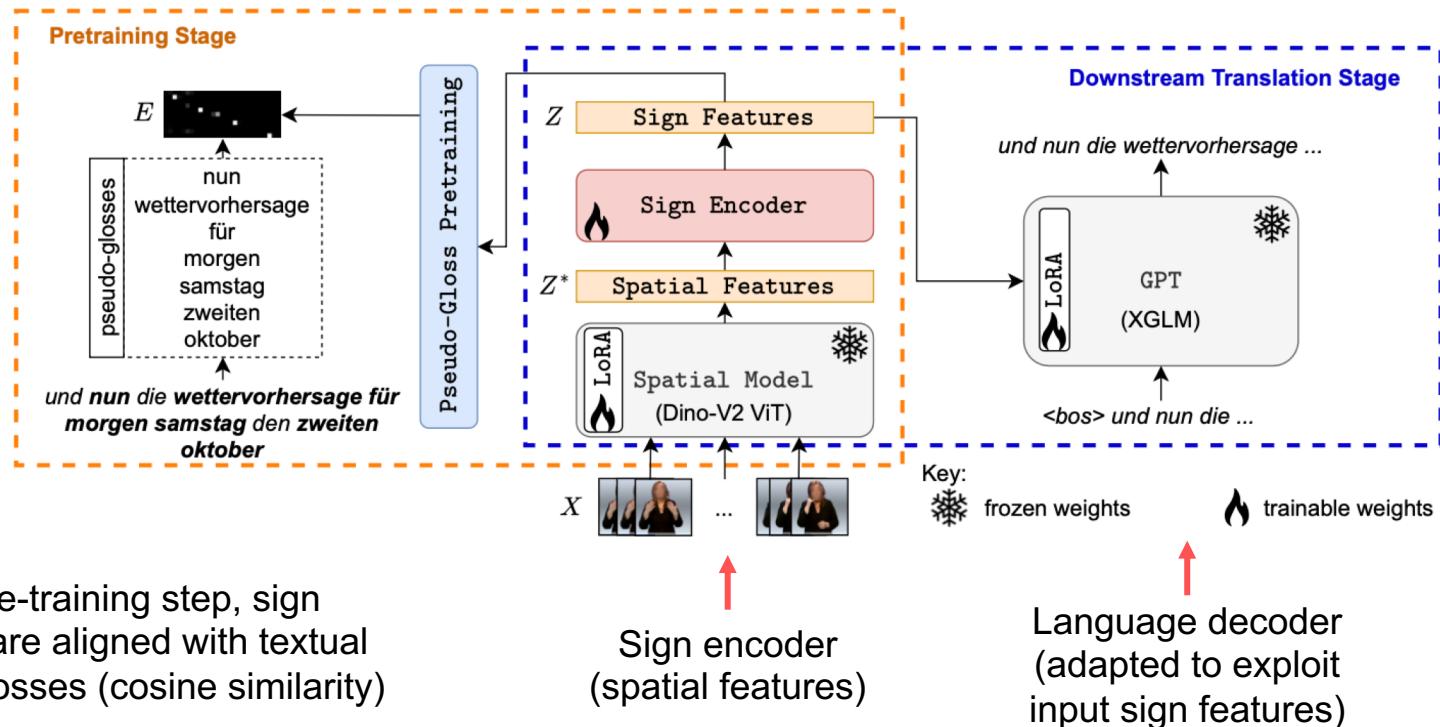
Generating Sign Language Instructions in Large Scale Dialogue Systems (Inan et al. NAACL-2024)



Figure 1: An overview of our multimodal dialogue system, capable of giving signed instructions to Deaf or Hard-of-Hearing users in ASL. We first translate task instructions to an intermediate textual representation called **glosses** using Large Language Models; then, we fetch token-level sign videos to display on the screens of Amazon Alexa Echo Show.

Prompting GPT-3 turbo + rule-based post-processing

Sign2GPT: Leveraging LLMs for Gloss-free Sign Language Translation (Wong et al. ICLR-2024)

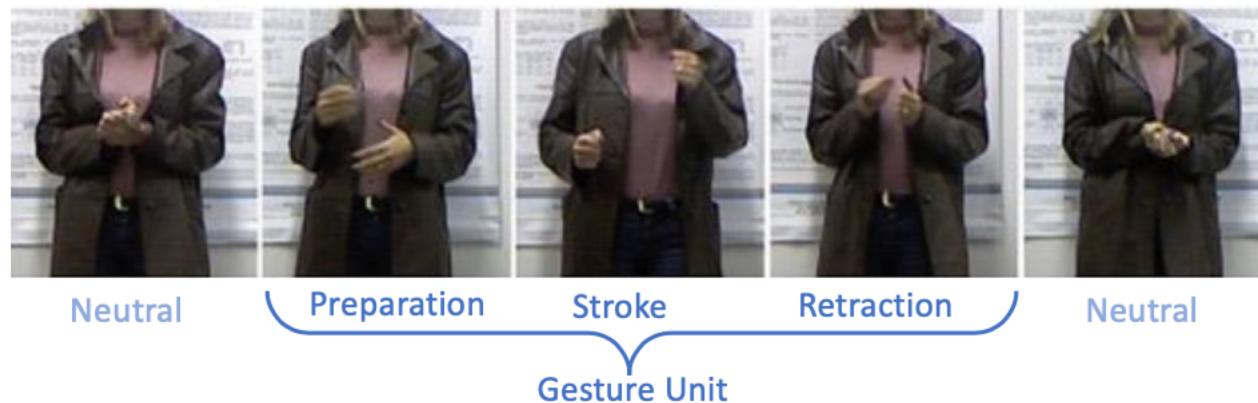


Face to face communication: co-speech gestures

Different kinds of gestures: emblems, beat gestures, representational/iconic

Most work in computer vision on isolated emblems

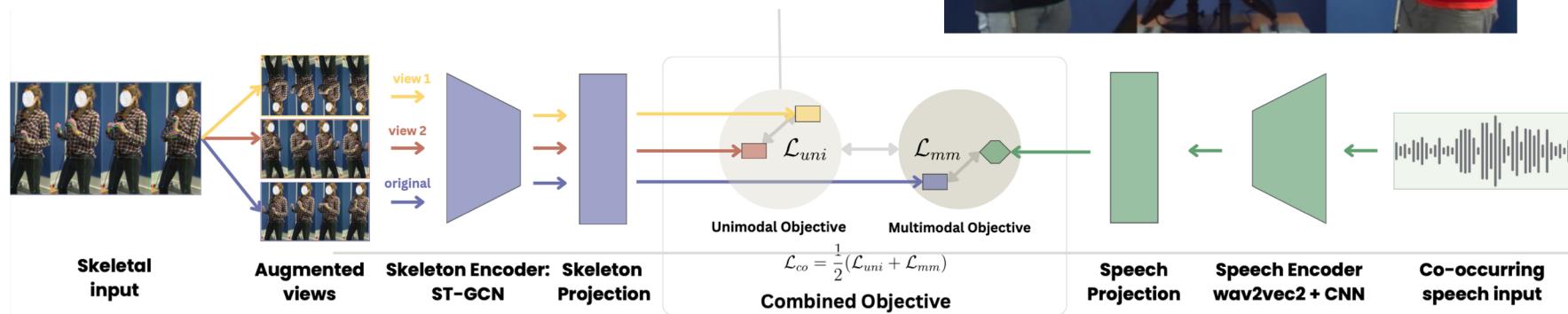
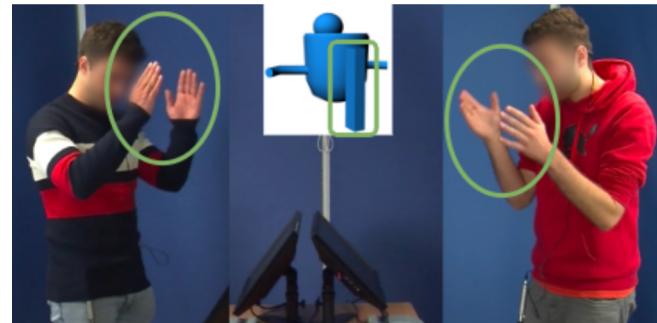
Beat & representational/iconic gestures more interesting from the NLP perspective



Kendon. Gesture units, gesture phases and speech. In *Gesture: Visible Action as Utterance*. 2004, CUP.

Self-supervised learning of iconic co-speech gestures

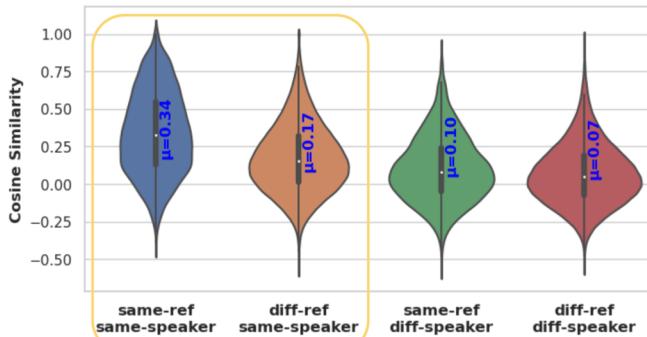
- How can we learn general (domain-specific) gesture representations?
- Our approach (Ghaleb et al., 2024) : self-supervised multimodal contrastive learning
 - Data augmentation with different gesture views
 - Gestures grounded in speech



Self-supervised learning of iconic co-speech gestures

- How can we learn general (domain-specific) gesture representations?
- **How can we evaluate such representations?**
 - **Intrinsically:** do they comply with similarity-based well motivated hypotheses?
 - Extrinsically: are they useful for downstream tasks? (work in progress)

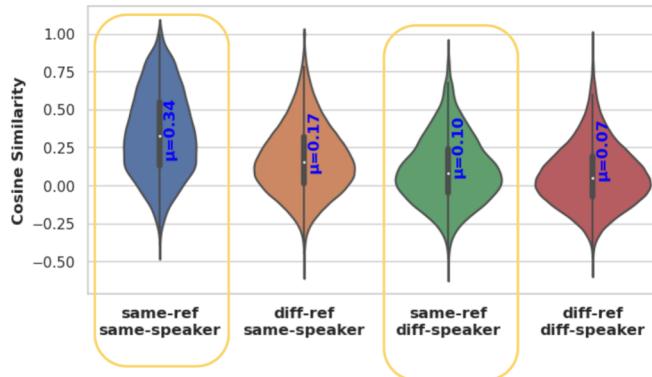
Hypothesis 1:
representations of
gestures by the same
speaker will be more
similar if they refer to
the same object



Self-supervised learning of iconic co-speech gestures

- How can we learn general (domain-specific) gesture representations?
- **How can we evaluate such representations?**
 - **Intrinsically:** do they comply with similarity-based well motivated hypotheses?
 - Extrinsically: are they useful for downstream tasks? (work in progress)

Hypothesis 1:
representations of
gestures by the same
speaker will be more
similar if they refer to
the same object

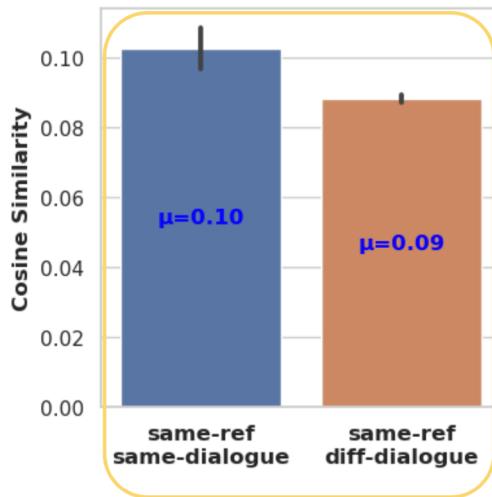


Hypothesis 2:
representation of
gestures with the same
referent will be more
similar if the gestures
are produced by the
same speaker

Self-supervised learning of iconic co-speech gestures

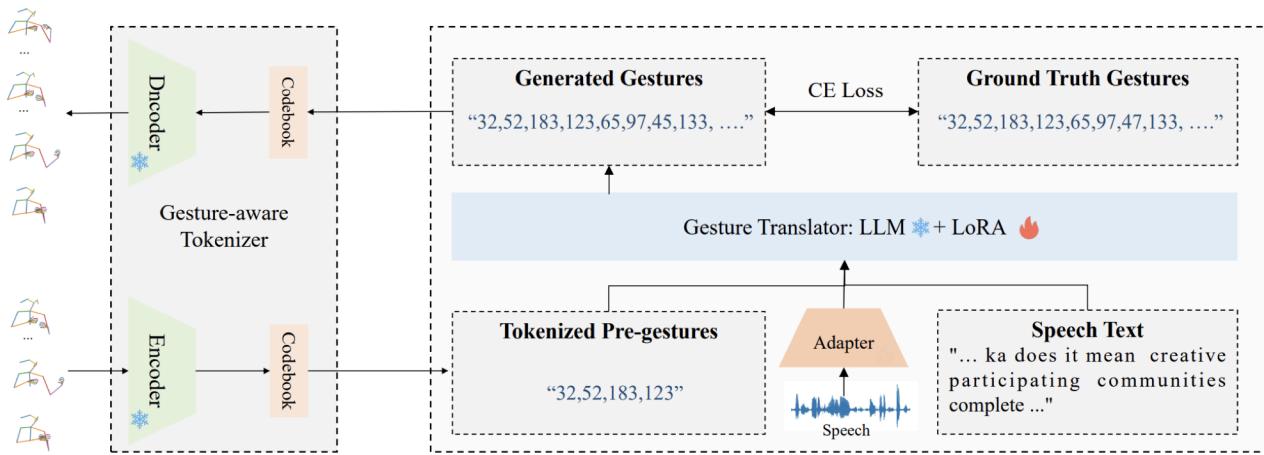
- How can we learn general (domain-specific) gesture representations?
- **How can we evaluate such representations?**
 - **Intrinsically:** do they comply with similarity-based well motivated hypotheses?
 - Extrinsically: are they useful for downstream tasks? (work in progress)

Hypothesis 3:
representations of gestures by
different speakers will be more
similar when the two speakers
are dialogue partners



GenTran: co-speech gesture generation

- Compile a “gesture lexicon”
- Adapt a generative LLM to use both speech and gesture tokens



Shu et al. LLM knows body language too: Translating speech voices into human gestures. ACL, 2024.

The future?

- Huge foundation models that can simultaneously model (read and generate) all sorts of modalities: images, text, speech, code, actions
- More attention to NLP for more *natural* language
 - speech, and face-to-face language use including speech & gesture
- Embodied agents that act in the world, virtually (avatars) or physically (robots).

The future?

- Huge foundation models that can simultaneously model (read and generate) all sorts of modalities: images, text, speech, code, actions
- More attention to NLP for more *natural*
 - speech, and face-to-face language use inclu
- Embodied agents that act in the world,
- Learning like a child?

> [Science](#). 2024 Feb 2;383(6682):504-511. doi: 10.1126/science.adl1374. Epub 2024 Feb 1.

Grounded language acquisition through the eyes and ears of a single child

[Wai Keen Vong](#) ¹, [Wentao Wang](#) ¹, [A Emin Orhan](#) ¹, [Brenden M Lake](#) ^{1, 2}

Abstract

Starting around 6 to 9 months of age, children begin acquiring their first words, linking spoken words to their visual counterparts. How much of this knowledge is learnable from sensory input with relatively generic learning mechanisms, and how much requires stronger inductive biases? Using longitudinal head-mounted camera recordings from one child aged 6 to 25 months, we trained a relatively generic neural network on 61 hours of correlated visual-linguistic data streams, learning feature-based representations and cross-modal associations. Our model acquires many

Acknowledgements

- Some of the present slides are inspired or directly taken (with permission) from the excellent LXMLS tutorial on Vision & Language by Desmond Elliott
<https://elliottd.github.io/vlprimer/>
- Thanks to the members of my research group for feedback and their awesome work!

