On visual grounding at the age of deep learning

Monthly online ILFC Seminar

Denis Paperno, 18.11.2024

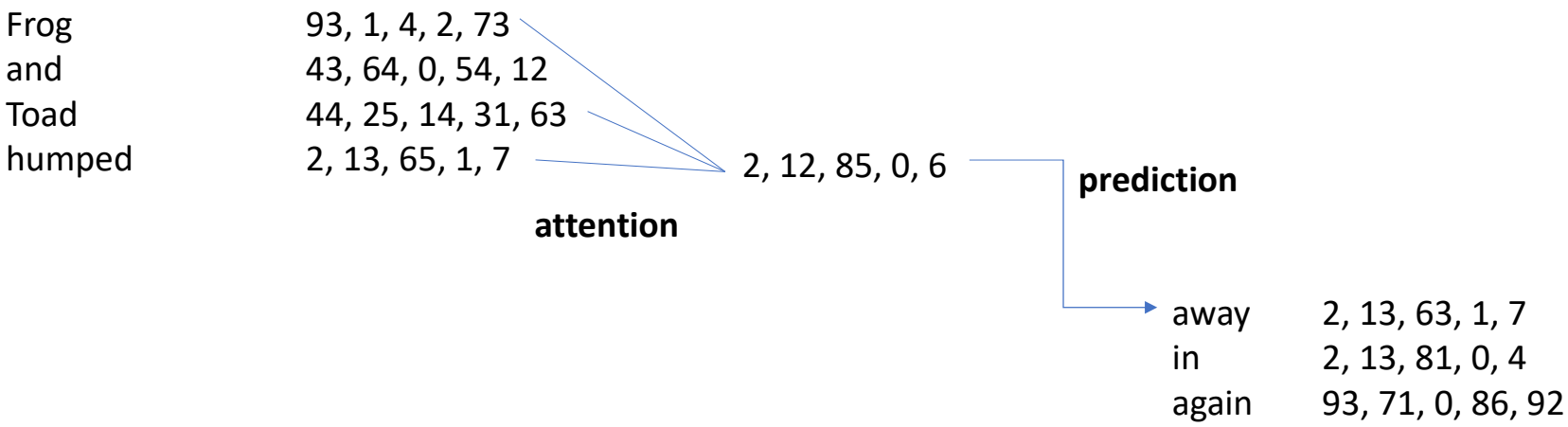
Goals of the talk

Widespread assumptions:

- Models for language based on text inherently limited
- Grounding them e.g. in visual modality is qualitatively different
- In particular, representation of meaning qualitatively different

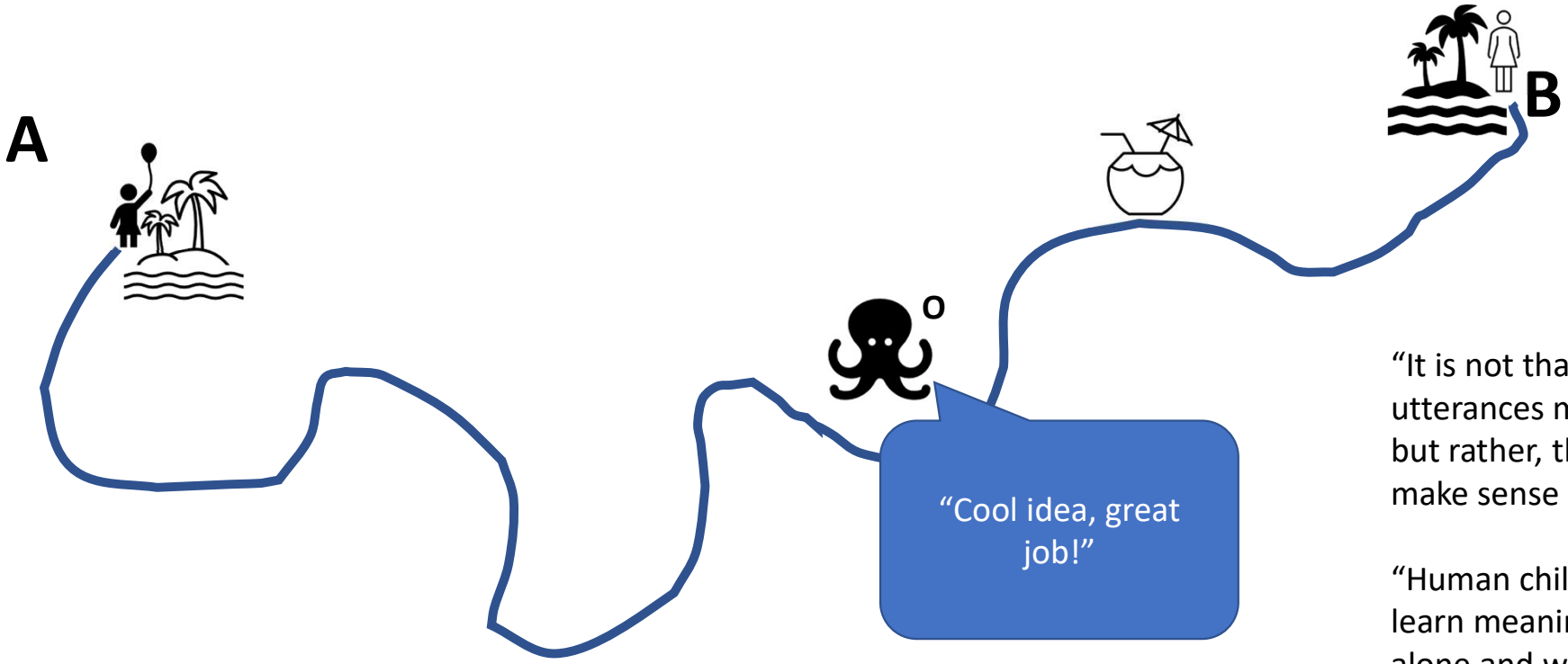
- How can we check this empirically?
- What changes when we switch to multimodality?
- How good/humanlike is multimodal models' grasp of meaning?

Transformer Language Models



The Octopus thought experiment

(Bender and Koller 2020)



"It is not that O's utterances make sense, but rather, that A can make sense of them."

"Human children do not learn meaning from form alone and we should not expect machines to do so either"

MULTIMODAL MODELS

GPT4, Google Gemini

12,87,66,93,98
18,20,67,29,98

I	93, 1, 4, 2, 73
invented	43, 64, 0, 54, 12
another	44, 25, 14, 31, 63
way	4, 15, 30, 51, 72
to	2, 13, 65, 1, 7

attention

2, 12, 85, 0, 6

prediction

open	2, 13, 63, 1, 7
start	2, 13, 81, 0, 4
be	93, 71, 0, 86, 92
…	

CLIP

12,87,66,93,98

match

coconut 18,20,67,29,98

Grounding and representations: controlled model learning

Timothee Mickus, Elaine Zosa, and Denis Paperno. Grounded and well-rounded: A Methodological Approach to the Study of Cross-modal and Cross-lingual Grounding. In Findings of EMNLP 2023.

Vatex dataset

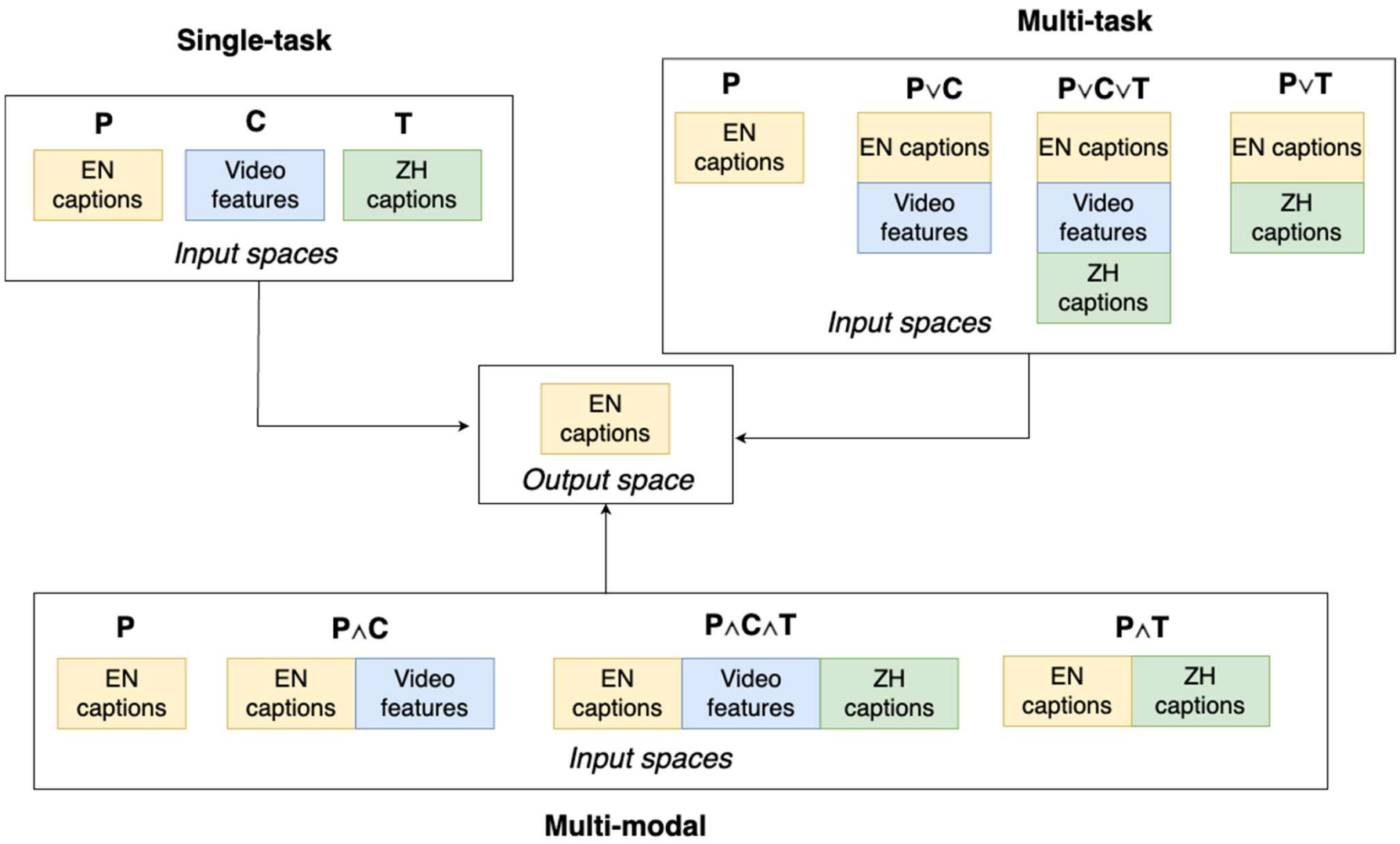
- Wang et al.2019

10 English Descriptions:

- A person wearing a bear costume is inside an inflatable play area as they lose their balance and fall over.
- A person in a bear costumer stands in a bounce house and falls down as people talk in the background.
- A person dressed in a cartoon bear costume attempts to walk in a bounce house.
- A person in a mascot uniform trying to maneuver a bouncy house.
- A person in a comic bear suit falls and rolls around in a moonbounce.
- A person dressed as a teddy bear stands in a bouncy house and then falls over.
- Someone dressed in a bear costume falling over in a bouncy castle.
- A person dressed up as a bear is standing in a bouncy castle and falls down.
- A man in a bear costume is balancing in a bouncy castle before they tumble to the floor.
- A man in costume was trying to stand straight on a bouncy castle but fell.

10 Chinese Descriptions:

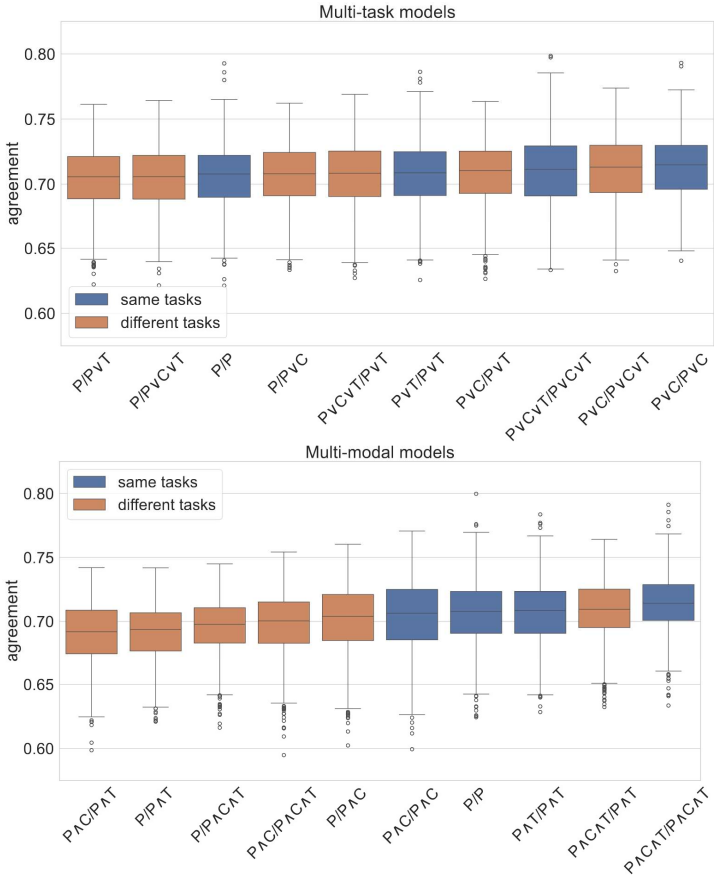
- 一个人穿着熊的布偶外套倒在了蹦床上。
- 一个人穿着一套小熊服装在充气蹦蹦床上摔倒了。
- 一个穿着熊外衣的人在充气垫子上摔倒了。
- 一个穿着深色衣服的人正在蹦蹦床上。
- 在一个充气大型玩具里,有一个人穿着熊的衣服站了一下之后就摔倒了。
- 一个打扮成泰迪熊的人站在充气房上，然后摔倒了。
- 有个穿着熊装的人在充气城堡摔倒了。
- 一个装扮成熊的人站在充气蹦床里，然后摔倒了。
- 一个穿着熊服装的人在一个有弹性的城堡里平衡，然后他们就倒在了地板上。
- 一个穿着布偶熊的人试图站在一个充气城堡上，但却摔倒了。

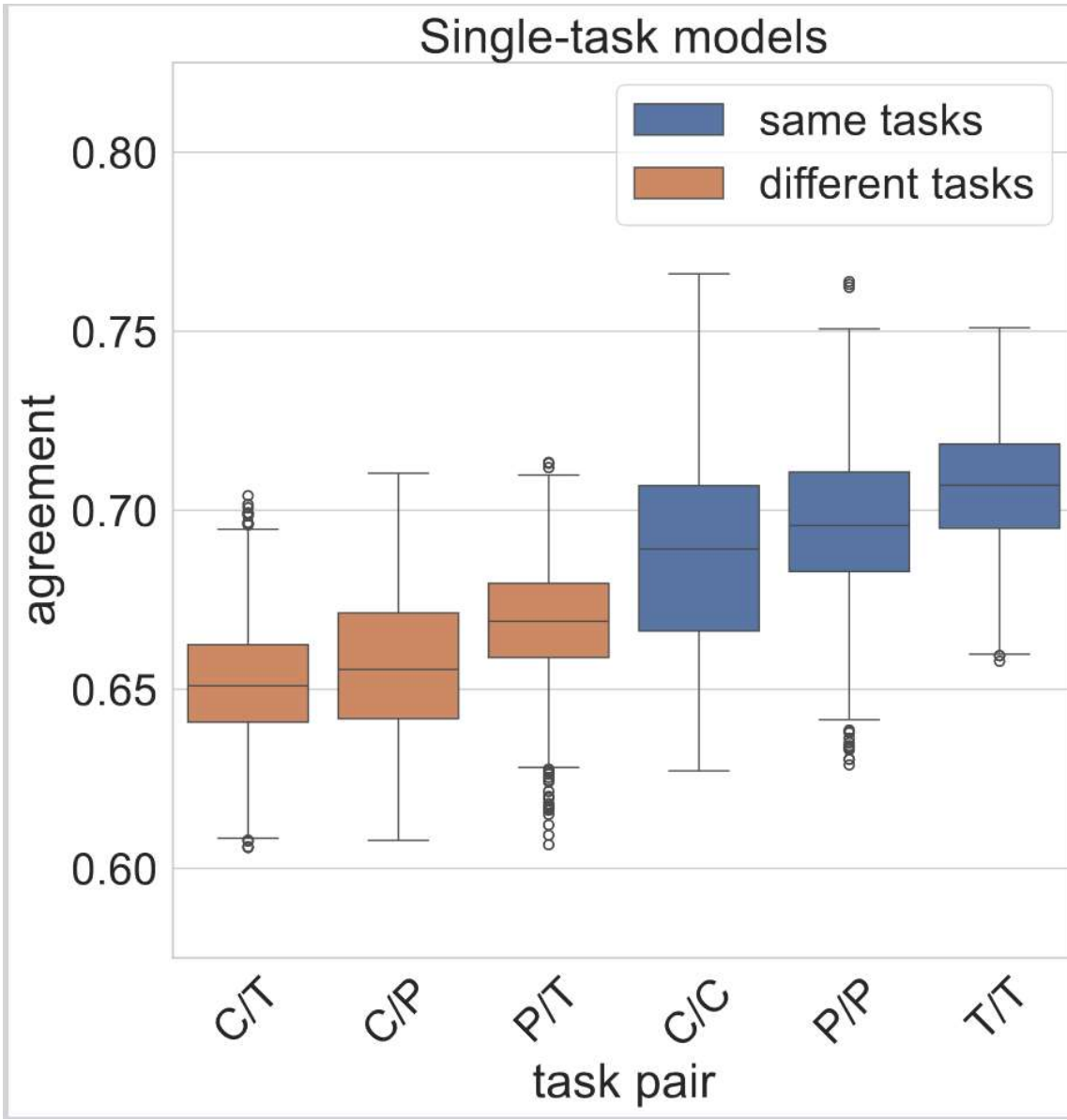

Single-task
P
C
T
EN captions
Video features
ZH captions
Input spaces
Multi-task
P
P∨C
P∨C∨T
P∨T
EN captions
EN captions
EN captions
EN captions
Video features
Video features
ZH captions
ZH captions
Input spaces
EN captions
Output space
P
P∧C
P∧C∧T
P∧T
EN captions
EN captions
Video features
EN captions
Video features
ZH captions
EN captions
ZH captions
Input spaces
Multi-modal

Experimental setup

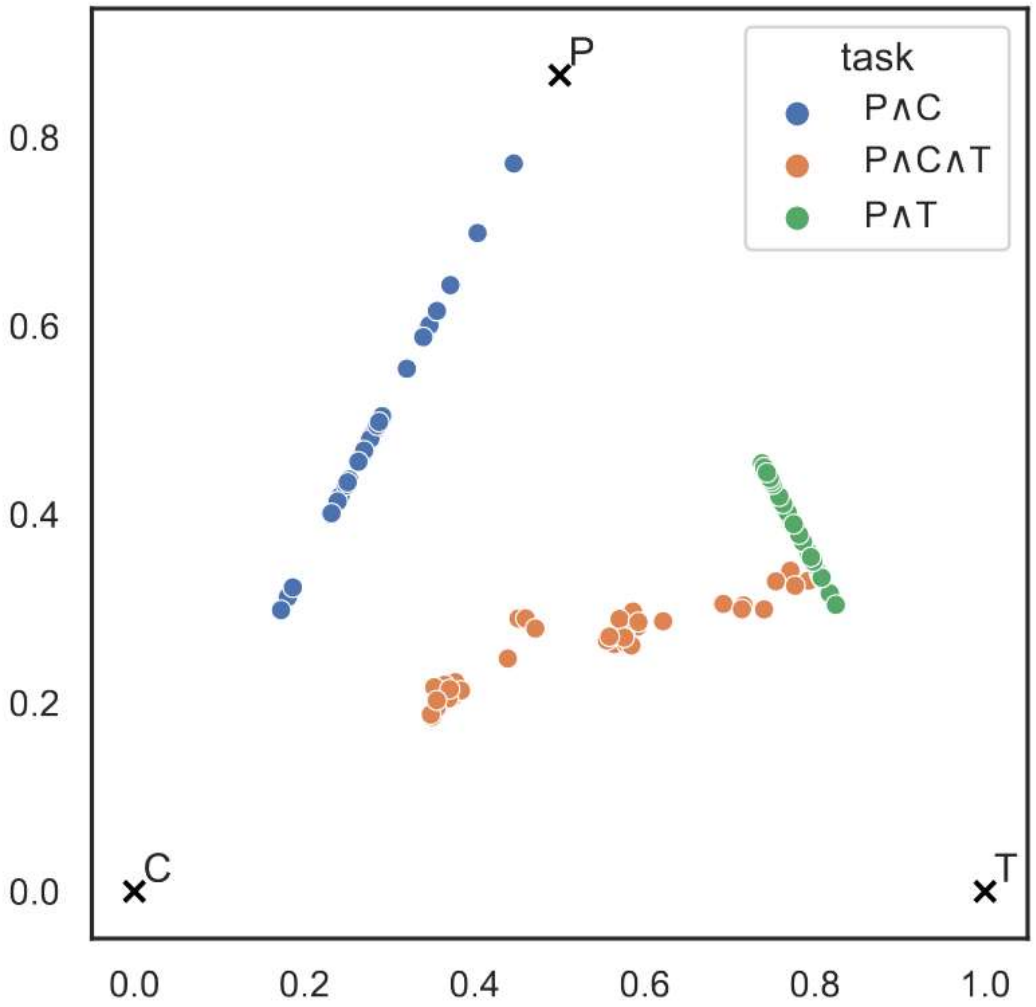
- For each type, train 25 model instances
- All models share the same architecture (64M parameters):
  - 6 layers
  - 8 heads per multihead sublayer
  - hidden dimensions of 512
  - latent feedforward dimensions of 2048
- Add varying degrees of noise to ensure comparable accuracy

Results: agreement between models





Results: attention patterns



Results: concreteness in word embeddings

Model	Silhouette		Purity		Inverse Purity	
	input	last hidden state	input	last hidden state	input	last hidden state
Single-task						
P	0.021(±0.004)	0.054(±0.007)	0.746(±0.021)	0.887(±0.005)	0.062(±0.015)	0.049(±0.005)
C	**0.026**(±0.002)	0.049(±0.006)	**0.760**(±0.013)	**0.890**(±0.005)	**0.080**(±0.025)	**0.068**(±0.014)
T	0.023(±0.004)	**0.056**(±0.006)	0.748(±0.016)	0.884(±0.005)	0.060(±0.014)	0.047(±0.006)
Multitask						
P	0.023(±0.004)	0.051(±0.005)	0.757(±0.017)	0.888(±0.005)	0.073(±0.021)	0.050(±0.008)
P∨C	0.024(±0.004)	0.056(±0.009)	0.752(±0.021)	0.891(±0.007)	0.070(±0.021)	**0.061**(±0.011)
P∨C∨T	**0.027**(±0.004)	0.055(±0.007)	**0.765**(±0.016)	0.889(±0.006)	0.073(±0.016)	0.052(±0.009)
P∨T	0.024(±0.004)	0.053(±0.008)	0.754(±0.019)	0.889(±0.006)	0.069(±0.016)	0.050(±0.007)
Multimodal						
P	0.025(±0.004)	0.051(±0.007)	0.759(±0.022)	0.890(±0.006)	0.074(±0.023)	**0.054**(±0.010)
P∧C	0.026(±0.003)	0.052(±0.006)	0.761(±0.018)	**0.891**(±0.007)	0.073(±0.019)	0.052(±0.006)
P∧C∧T	0.024(±0.004)	**0.056**(±0.007)	0.759(±0.017)	0.889(±0.0067)	0.072(±0.032)	0.048(±0.005)
P∧T	0.026(±0.004)	0.051(±0.008)	0.765(±0.017)	0.886(±0.007)	0.065(±0.014)	0.046(±0.005)

Interim conclusions

- Grounding leads to quantitatively and qualitatively different models
- Multimodal vs multilingual grounding: qualitatively different
- Multimodality: same input, qualitatively different representations

Grounding and representations: pretrained models

Aleksey Tikhonov, Lisa Bylinina and Denis Paperno. Leverage Points In Modality Shifts: Comparing Language-Only and Multimodal Word Representations. Proceedings of the 12th Joint Conference on Lexical and Computational Semantics: 11–17.

Goals

- There are pretrained models for both text and multimodal input.
- Models used in this study:
  - CLIP, OpenCLIP, Multilingual CLIP

  For each: 2 ways of extracting word embeddings **iso** (word in isolation), **avg** (over 10 sentences)
  - fastText, multilingual BERT, XLM-RoBERTa

  For BERT and RoBERTa: 3 ways of extracting word embeddings **iso**, **avg-last**, **avg-bottom**

- Unimodal vs. multimodal representation spaces: different structure?

- What semantic factors account for differences in structure?

Methodology

- Use a sample of word pairs
- For each model M, calculate cosine distance of pairs $dist^{M}(w_i, w_j)$
- Control for differences in embedding space, e.g. anisotropy:
  - Rank pairs by shift of distance between two models $\frac{dis^{M1}(w_i,w_j)}{dist^{M2}(w_i,w_j)}$
  - Regression analysis: predict rank of shifts from linguistic properties

- Hypothesis: some properties explain more of a difference when models belong to different classes (multimodal vs textual)

Factors considered

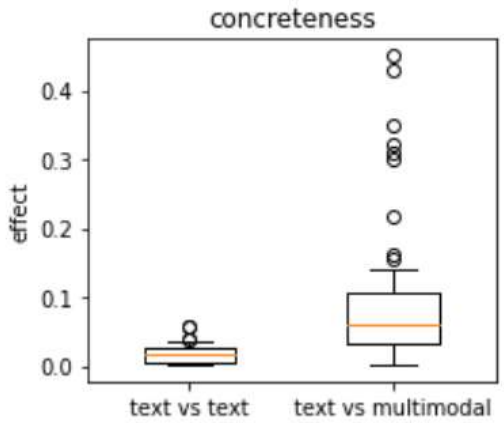
- Frequency
- Concreteness (Ghent norms)
- Ontological category (WordNet supersenses)
- Affective meaning (VAD: Valence, Arousal, Dominance)
- Relation within word pair
  - WordNet relations (e.g. hypernymy)
  - ConceptNet relations (e.g. *is used for*)

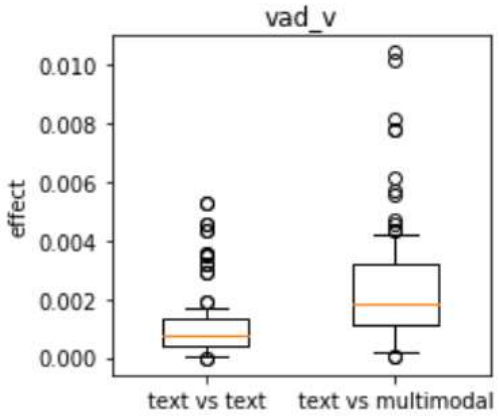
Illustration: method of model comparison

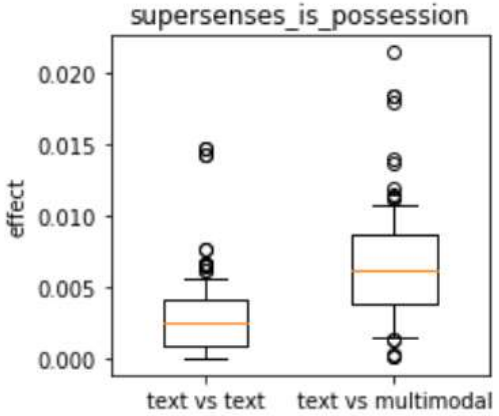
- Variance of distance ratio rank explained:

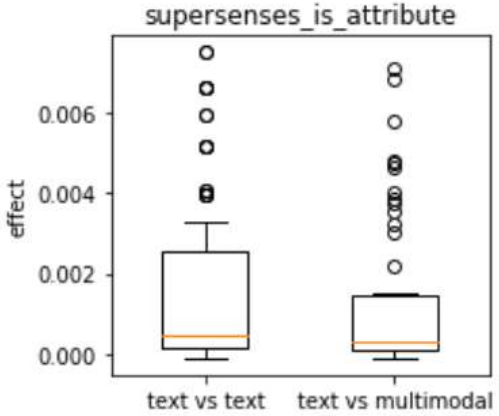
CLIP-iso vs.	**XLMR-iso**	**mBERT-iso**	**BERT-avg-last**	**fastText**
Baselines				
concreteness	9.5	11.68	2.27	8.71
frequency	5.43	7.81	1.91	0.45
concreteness+frequency	16.73	17.16	3.65	9.54
+taxonomic	21 (+4.27)	20.35 (+3.19)	5.43 (+1.78)	19.50 (+9.96)
+VAD	17.36 (+0.63)	17.49 (+0.33)	4.62 (+0.97)	10.78 (+1.24)
+WordNet relations	18.47 (+1.74)	17.36 (+0.2)	10.05 (+6.4)	10.34 (+0.8)
+ConceptNet relations	19.8 (+3.07)	17.47 (+0.31)	8.84 (+5.19)	10.26 (+0.72)

Illustration: specific features

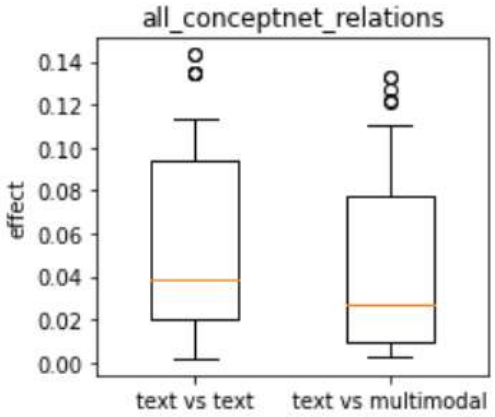


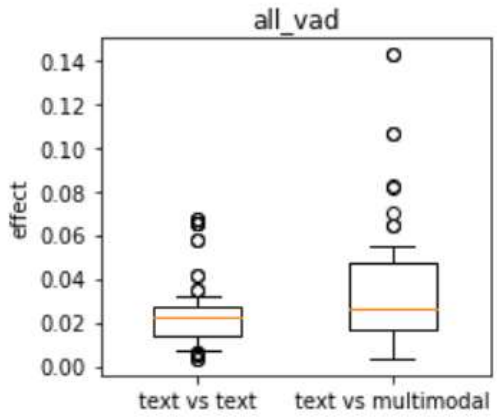


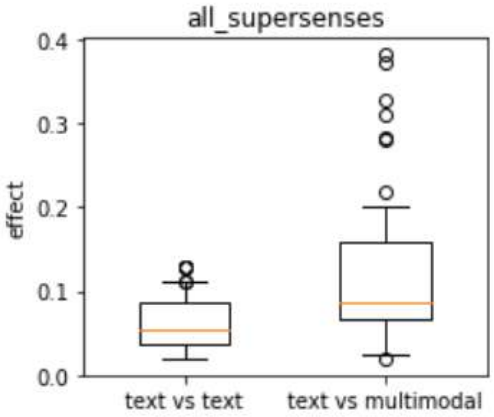


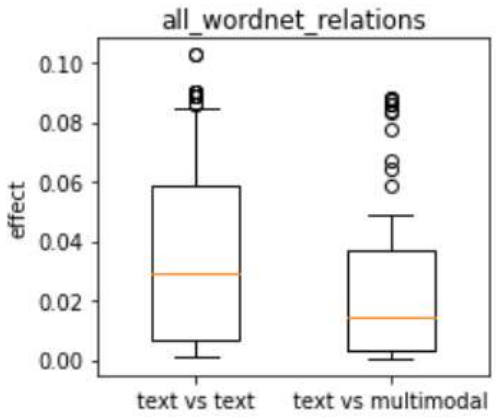


Results: global view









Results summary

- **Concreteness** plays a major role in explaining modality shifts, in line with results of previous studies.
- Combined **WordNet supersenses** have a significant effect for many, although different subsets of features prove significant in different pairs of models.
- WordNet and ConceptNet **relations** tend to be significant when aggregated, although no individual relation has a systematic effect across model pairs.
- **VAD** features produce varied effects, with **valence** showing the most consistent modality difference.

Case study: derivational contrast

Claudia Tagliaferri, Sofia Axioti, Albert Gatt and Denis Paperno. The Scenario Refiner: Grounding subjects in images at the morphological level. LIMO workshop (Linguistic Insights from and for Multimodal Language Processing @KONVENS 2023).

Methodology

- Collect human ratings on image-text match on a Likert scale

- -er nouns vs -ing forms of verbs from four domains:

professional	*sports*	*artistic*	*general*
baker vs bake	runner vs run	painter vs paint	supporter vs support
teacher vs teach	surfer vs surf	singer vs sing	reader vs read
cleaner vs clean	skier vs ski	dancer vs dance	lover vs love
	…		…

- Contrast human judgments to VLM predictions of match

The task:

the woman with pink gloves is a driver	**4.4**
the woman with pink gloves is driving	4.3

the man and the woman are supporters	6.2
the man and the woman are supporting	**6.4**

Human vs human (correlation, two groups)

Domain	Derived noun	Verb	Morphological contrast
Professional domain	0.76	0.84	0.75
Sport domain	0.69	0.70	0.60
Artistic domain	0.79	0.31	0.51
General	0.92	0.88	0.94
All domains	0.80	0.81	0.78

Statistics of preference

	Derived noun	Verb
Humans	8.3%	91.7%
CLIP ViT-L/14@336px	51.9%	48.1%
CLIP RN50x64	52.8%	47.2%
CLIP ViT-B/32	49.1%	50.9%
ViLT	47.2%	52.8%
LXMERT	51.9%	48.1%

Model vs human correlations overall

Model	Derived noun	Verb	Morphological contrast
CLIP ViT-L/14@336px	0.13	0.08	0.15
CLIP RN50x64	0.09	0.08	-0.01
CLIP ViT-B/32	0.09	0.18	0.08
ViLT	0.07	0.26	0.32
LXMERT	0.16	0.03	0.21

Model vs. human correlations per domain

	Sport domain			Professional domain		
	Deriv. noun	**Verb**	**Morph. contrast**	**Deriv. noun**	**Verb**	**Morph. contrast**
CLIP ViT-L/14@336p	0.02	-0.25	-0.06	-0.04	0.19	0.33
CLIP RN50x64	0.02	-0.31	-0.11	-0.22	0.33	0.23
CLIP ViT-B/32	-0.04	-0.26	-0.11	-0.16	0.23	0.40
ViLT	-0.01	-0.04	0.45	0.30	0.45	0.68
LXMERT	0.08	-0.32	0.17	-0.10	0.18	0.40
	Artistic domain			**General**		
	Derived noun	**Verb**	**Contrast**	**Derived noun**	**Verb**	**Contrast**
CLIP ViT-L/14@336p	-0.03	0.01	0.22	0.28	0.18	-0.06
CLIP RN50x64	0.06	0.21	0.27	0.08	0.23	0.10
CLIP ViT-B/32	-0.09	-0.04	-0.007	0.23	0.25	-0.06
ViLT	0.44	0.15	0.39	-0.06	0.05	0.25
LXMERT	0.29	0.26	0.42	-0.01	-0.25	0.12

Conclusions from the three studies

1. Methodology for testing qualitative differences between modalities

2. Identify which dimensions of meaning affected by modality

3. Meaning in multimodal models tested in a novel way, not humanlike

Thank you!

- Collaborators:

Albert Gatt

Aleksey Tikhonov

Claudia Tagliaferri

Elaine Zosa

Lisa Bylinina

Sofia Axioti

Timothee Mickus