Contextual Understanding of Language in the World

Nikolai Ilinykh - Slutseminar
27.3.2024

Language in context

Language is used in the world, about the world

The world as *perceived* through senses, e.g. vision, hearing

The world as *learned* through accumulated experiences

Language is used by people:

- in shared contexts
- with shared common ground
- with shared history
- with shared goals
- but also with idiosyncratic and interesting variation

How to talk about an image

Possible task contexts/goals:

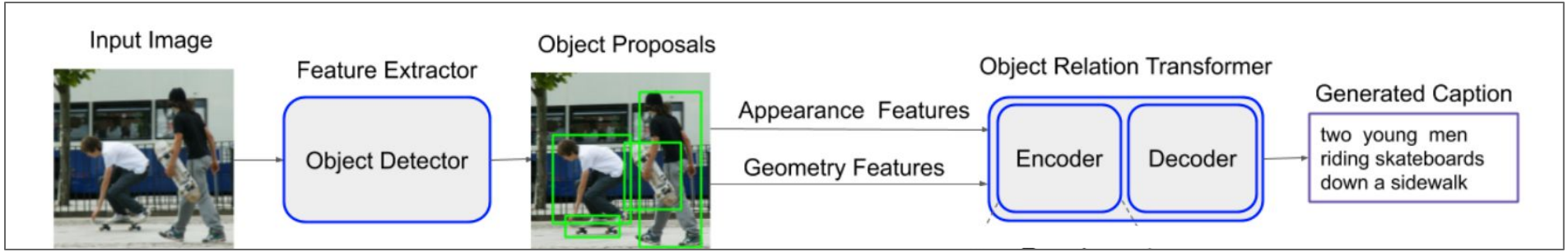
- identify this image among others
- describe the image in a sentence (captioning)
- ‘tell me more’: give further details in a paragraph
- identify a dog species
- suggest interior design styles
- guess what’s for dinner

Can computational models of language and vision incorporate contextual information from different modalities to produce a contextual description of the image?

Primary research question

A Language & Vision Transformer Model

Object Relation Transformer: Herdade et al, 2019



Object detector (Faster R-CNN) gives explicit labelled regions, also features

Encoder and Decoder are 6-layer Transformers with 8 heads

Self-attention in Encoder; Self + Cross-attention in Decoder

I: Learning Knowledge & Structures beyond Text

II: Language-and-Vision Representation Learning

III: Generating Language with Respect to the Task

I. Learning knowledge & structures beyond text

What can the decoder in language-and-vision models learn that language-only models can't?

What can we see in the learned *attention* patterns in the model components?

How Vision Affects Language: Comparing Masked Self-Attention in Uni-Modal and Multi-Modal Transformer. Nikolai Ilinykh and Simon Dobnik. 2021. In Proceedings of the 1st Workshop on Multimodal Semantic Representations (MMSR)

What Does a Language-And-Vision Transformer See: The Impact of Semantic Information on Visual Representations. Nikolai Ilinykh and Simon Dobnik. 2021. Frontiers in Artificial Intelligence

Attention as Grounding: Exploring Textual and Cross-Modal Attention on Entities and Relations in Language-and-Vision Transformer. Nikolai Ilinykh and Simon Dobnik. 2022. In Findings of the Association for Computational Linguistics: ACL 2022

Attention(s)!

Self-attention in image encoder and text decoder

Cross-attention:

- Keys and Values from image encoder
- Queries & residual from text decoder

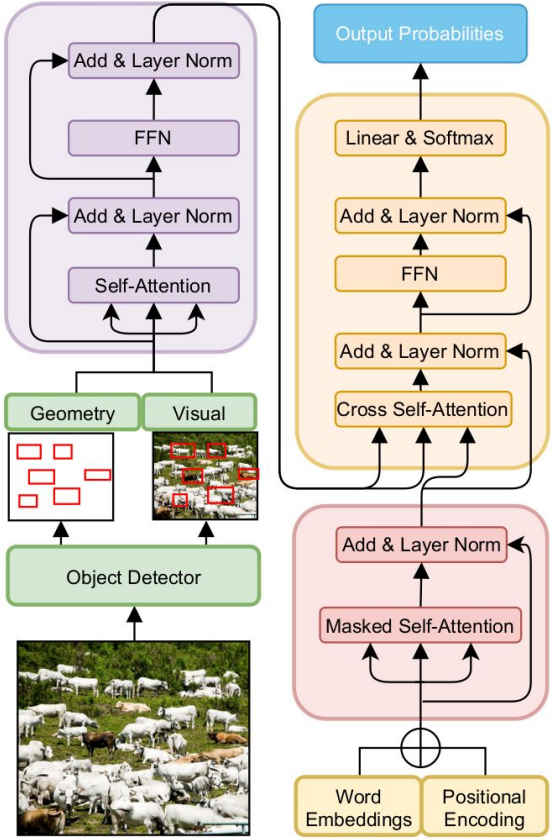
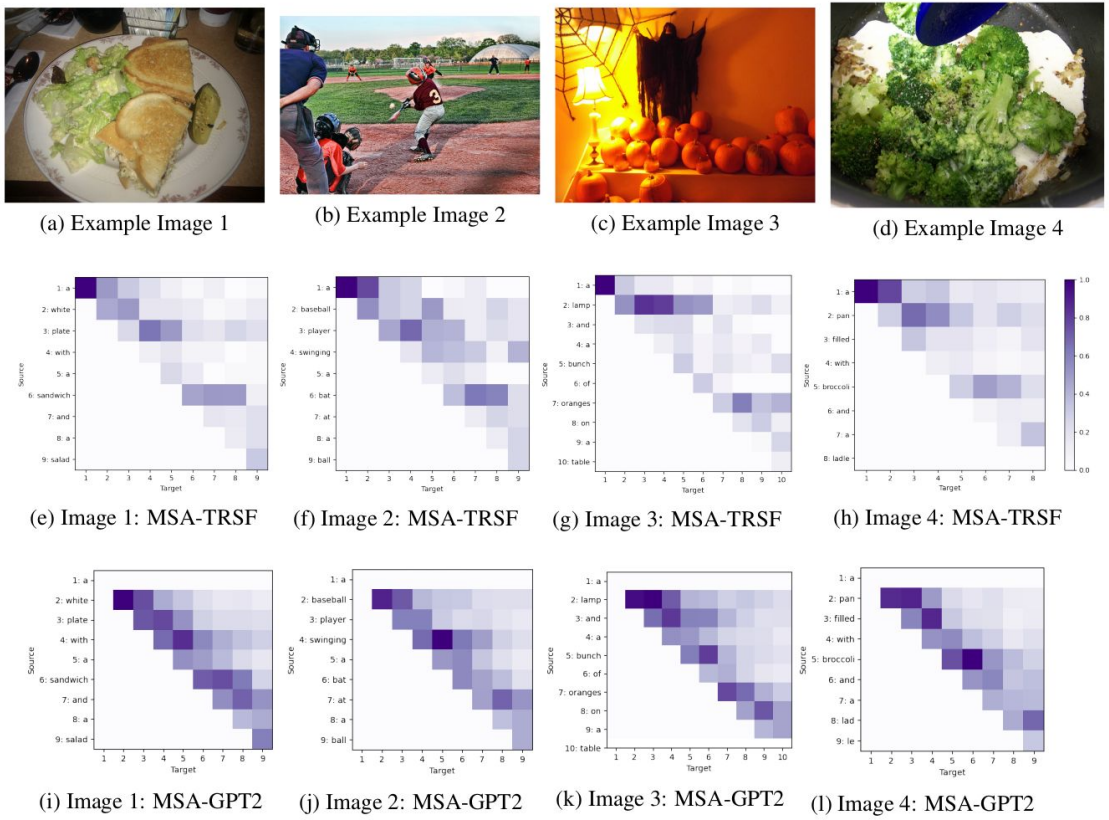


Fig 1 from Ilinykh & Dobnik, 2021

How Vision Affects Language (2021)

Masked self-attention patterns in text decoder looks quite different between multimodal model (top) and text-only GPT2 (bottom)

Multimodal model attention “jumps” from object noun to noun.
Text-only attention is more broadly dispersed



What does a L&V Transformer See? (2021)

Self-attention in image encoder: which regions/objects attend to each other?

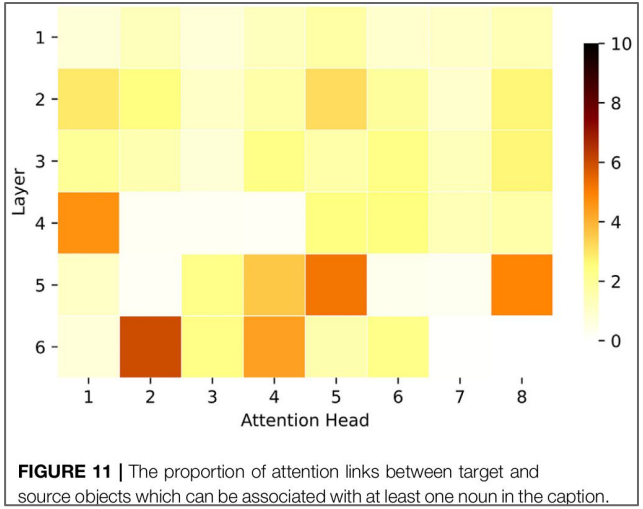


FIGURE 11 | The proportion of attention links between target and source objects which can be associated with at least one noun in the caption.

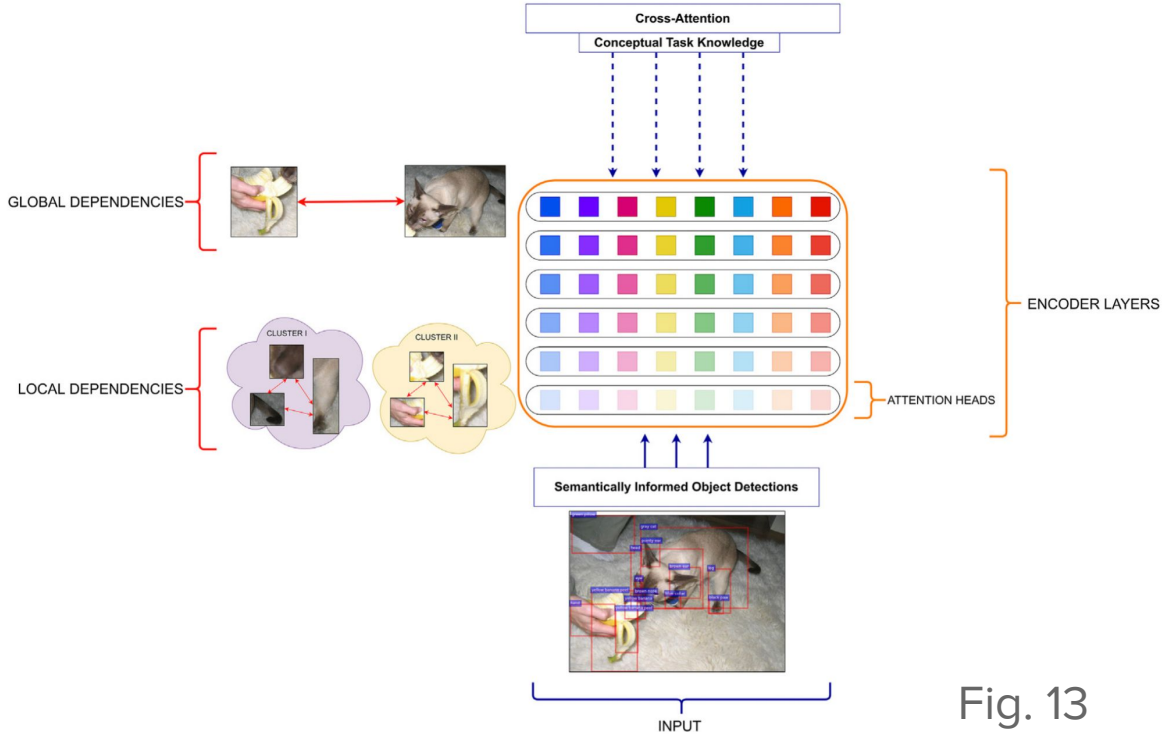


Fig. 13

Attention as Grounding (2022)

Cross-modal attention during generation/decoding

Q1: Does cross-attention look at the right object when generating a noun?
A: Yes, eventually!

Q2: Does attention during spatial relation generation look at the target? landmark? word or object?

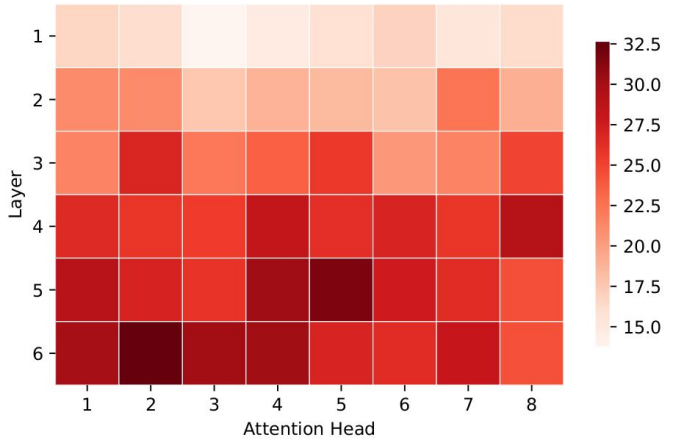
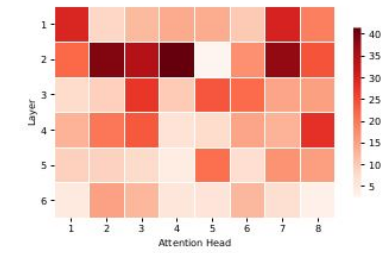
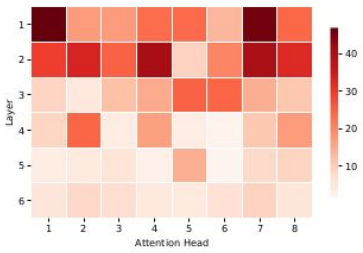


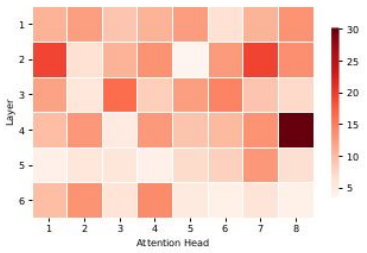
Figure 4: Attention proportions $\boldsymbol{P}$ on correct noun-object pairs (as determined by linking) for each attention head in the cross-modal attention. The darker the colour, the **bigger** the proportion. The proportions are averaged over the noun phrases in descriptions.



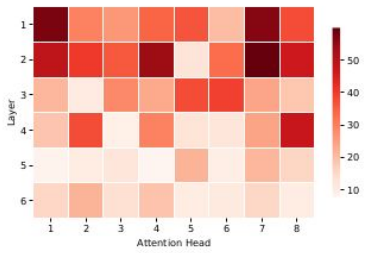
(a) $rel \rightarrow target$



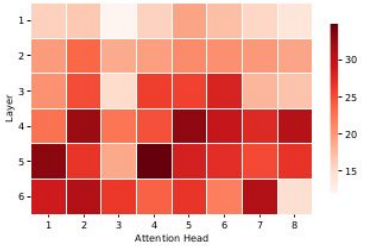
(b) $land \rightarrow rel$



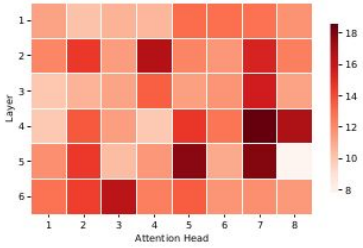
(c) $land \rightarrow target$



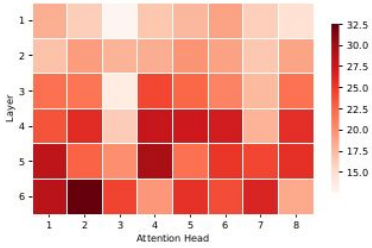
(d) $land \rightarrow rel + target$



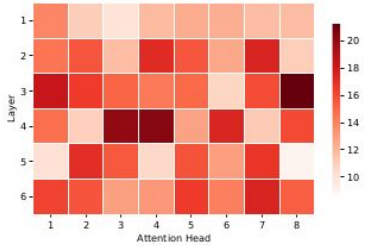
(e) $rel \rightarrow target$



(f) $rel \rightarrow land$



(g) $land \rightarrow target$



(h) $target \rightarrow land$

Figure 6: Heat-map visualisations of $\boldsymbol{P}$ for masked self-attention (**the top row**) and cross-attention (**the bottom row**) for different possible configurations of attention between words constituting spatial relations. All attention proportions are normalised by the number of spatial relations in the test set.

II. L&V representation learning

What kinds of features to use from multiple modalities?

How to effectively combine features from multiple modalities?

When an Image Tells a Story: The Role of Visual and Semantic Information for Generating Paragraph Descriptions. Nikolai Ilinykh and Simon Dobnik. 2020. In Proceedings of the 13th International Conference on Natural Language Generation.

Look and Answer the Question: On the Role of Vision in Embodied Question Answering. Nikolai Ilinykh, Yasmeen Emampoor, and Simon Dobnik. 2022. In Proceedings of the 15th International Conference on Natural Language Generation.

Context matters: evaluation of target and context features on variation of object naming. Nikolai Ilinykh and Simon Dobnik. 2023. In Proceedings of the 1st Workshop on Linguistic Insights from and for Multimodal Language Processing.

When an Image Tells a Story (2020)

Task: generate a multi-sentence paragraph description of an image.

Input image is represented as either visual features or textual region descriptions (or both): conjunction mostly performs best.

Visual input is given to paragraph-level sentence planner module, which either uses *attention* or *max-pooling*: attention has better diversity, worse metrics.

Evaluating a paragraph goes beyond BLEU: avoid repetition, want diversity

- Human evaluation for word choice, object salience, sentence structure, paragraph structure
- Human and automatic metrics disagree about best model:
  Humans prefer model with textual region features + attention

Look and Answer the Question (2022)

Task: Embodied Question Answering - navigate to a room to answer a question.

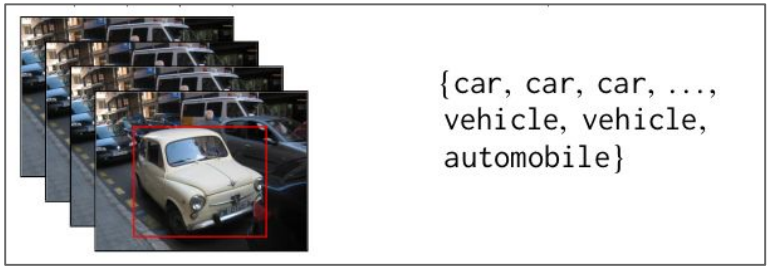
Test the effect of perturbing visual input.

Result: shuffled and blind models match original - indicates dataset is flawed.

Figure 1: Example of successive removal of context, content and structure. For each removal type, we show the first frame from the set of frames that the model takes to answer the question "What color is the stove in the kitchen?". From left to right: **original** (nothing is removed), **shuffled** (structure and content are present, but context is incorrect), **blind** (no content and context, but structure), **random** (most disturbed representation).

Context Matters (2023)

ManyNames Dataset



Naming Variation: same object can be identified as different nouns.

Q: What combination of context feature types (target, context, scene; text, visual) result in a model that matches human naming variation?

A: *multimodal target* features are most important for predicting most frequent name; adding *scene* features is better for capturing variation.

Match is calculated as rank correlation between dataset name distribution and model's predicted probability distribution over those names.

III. Generating Task-Oriented Language

Do text-generating multimodal models produce good language?

- Natural & correct *discourse structure*
- Applicability to *task context*
- *Variability*: human language is diverse, not monotone or repetitive
- *Complexity*: often need more than simplest possible utterance

Do Decoding Algorithms Capture Discourse Structure in Multi-Modal Tasks? A Case Study of Image Paragraph Generation. Nikolai Ilinykh and Simon Dobnik. 2022. In Proceedings of the 2nd Workshop on Natural Language Generation, Evaluation, and Metrics (GEM).

Describe Me an Auklet: Generating Grounded Perceptual Category Descriptions. Bill Noble* and Nikolai Ilinykh*. 2023. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing.

Approximation sacrifices variability: estimating the closeness between human and machine language in image descriptions. Nikolai Ilinykh and Simon Dobnik. 2024. To be published on arXiv.

Do Decoding Algorithms capture Discourse? (2022)

Different decoding methods lead to large variation in metric results.

However, human ratings of these sentences don't correlate (much).

Generated texts describe different sets of objects than reference texts.

		g			b2			s50			st50			n50			db2		
		P	*S*	*K*	*P*	*S*	*K*	*P*	*S*	*K*	*P*	*S*	*K*	*P*	*S*	*K*	*P*	*S*	*K*
R	BLEU_1	0.23	0.18	0.13	0.3	0.28	0.22	-0.01	-0.06	-0.03	-0.06	-0.03	-0.02	0.25	0.21	0.15	0.27	0.19	0.15
	BLEU_2	0.21	0.17	0.12	0.34	0.28	0.2	-0.04	-0.16	-0.1	-0.13	-0.15	-0.11	0.14	0.1	0.06	0.3	0.19	0.14
	BLEU_3	0.14	0.16	0.1	0.29	0.22	0.17	-0.05	-0.12	-0.07	-0.15	-0.2	-0.14	0.11	0.1	0.07	0.27	0.21	0.16
	BLEU_4	0.01	0.1	0.07	0.26	0.24	0.18	0.04	-0.1	-0.04	-0.12	-0.16	-0.12	0.19	0.11	0.08	0.2	0.22	0.16
	METEOR	-0.21	-0.18	-0.13	0.14	0.12	0.09	-0.21	-0.22	-0.16	-0.22	-0.32	-0.22	-0.05	-0.09	-0.06	-0.26	-0.24	-0.19
	ROUGE_L	0.18	0.15	0.11	0.22	0.23	0.16	0.06	0.02	0.02	-0.19	-0.22	-0.15	0.19	0.16	0.11	0.28	0.21	0.15
	CIDEr	0.02	0.15	0.1	0.33	0.17	0.12	-0.06	-0.17	-0.1	-0.15	-0.18	-0.11	0.23	0.23	0.17	0.16	0.19	0.14
	WMD	-0.0	0.0	-0.0	0.2	0.16	0.1	-0.14	-0.09	-0.06	-0.12	-0.14	-0.1	-0.09	-0.09	-0.06	-0.14	-0.12	-0.09
C	BLEU_1	0.14	0.13	0.09	0.11	0.12	0.09	0.02	0.01	0.0	0.06	0.11	0.07	0.22	0.19	0.13	0.19	0.18	0.12
	BLEU_2	0.12	0.08	0.06	0.15	0.15	0.12	-0.05	-0.12	-0.09	0.09	0.13	0.09	0.13	0.1	0.06	0.21	0.18	0.12
	BLEU_3	0.02	0.05	0.03	0.09	0.11	0.08	-0.09	-0.12	-0.09	0.11	0.13	0.08	0.12	0.1	0.06	0.18	0.18	0.13
	BLEU_4	-0.12	-0.02	-0.03	0.02	0.09	0.06	-0.0	-0.13	-0.1	0.1	0.14	0.09	0.22	0.15	0.11	0.17	0.22	0.16
	METEOR	-0.15	-0.13	-0.08	0.09	0.08	0.06	-0.19	-0.17	-0.13	-0.09	-0.1	-0.07	-0.16	-0.24	-0.16	-0.27	-0.27	-0.2
	ROUGE_L	0.13	0.19	0.14	0.06	0.08	0.05	-0.07	-0.09	-0.07	-0.02	0.02	0.02	0.22	0.19	0.14	0.16	0.17	0.11
	CIDEr	0.03	0.12	0.09	0.14	0.1	0.05	-0.0	-0.01	-0.01	-0.07	0.09	0.08	0.22	0.26	0.17	0.12	0.21	0.16
	WMD	-0.02	-0.03	-0.02	0.16	0.13	0.1	-0.22	-0.17	-0.12	-0.09	-0.07	-0.05	-0.22	-0.28	-0.21	-0.1	-0.09	-0.08
F	BLEU_1	0.41	0.37	0.27	0.42	0.4	0.31	-0.22	-0.24	-0.19	0.01	0.0	0.01	0.13	0.08	0.06	0.32	0.32	0.24
	BLEU_2	0.39	0.36	0.28	0.38	0.29	0.23	-0.18	-0.27	-0.21	-0.01	-0.04	-0.03	0.07	0.05	0.03	0.32	0.31	0.22
	BLEU_3	0.29	0.32	0.23	0.35	0.25	0.19	-0.22	-0.25	-0.18	0.01	-0.0	0.0	0.12	0.07	0.05	0.3	0.3	0.22
	BLEU_4	0.15	0.24	0.18	0.23	0.2	0.14	-0.01	-0.17	-0.12	0.03	0.05	0.03	0.19	0.06	0.04	0.22	0.24	0.17
	METEOR	-0.07	-0.07	-0.08	0.12	0.09	0.06	-0.11	-0.14	-0.1	-0.0	-0.06	-0.03	-0.12	-0.2	-0.16	-0.01	-0.01	-0.01
	ROUGE_L	0.31	0.29	0.22	0.24	0.24	0.18	-0.06	-0.1	-0.08	-0.08	-0.07	-0.05	0.16	0.12	0.09	0.28	0.29	0.19
	CIDEr	0.16	0.27	0.2	0.36	0.27	0.21	-0.35	-0.37	-0.28	0.01	0.02	0.02	0.02	0.1	0.07	0.13	0.29	0.23
	WMD	0.04	0.03	0.02	0.19	0.22	0.14	-0.14	-0.16	-0.12	-0.03	-0.02	-0.03	-0.02	-0.03	-0.03	0.1	0.05	0.03

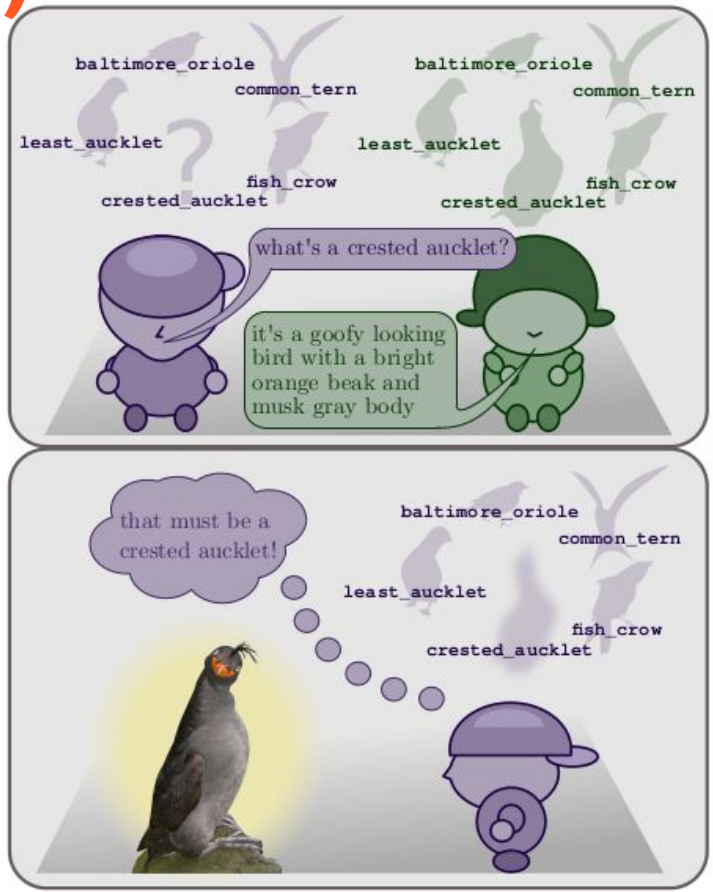
Table 3: Correlation scores between automatic metrics and human judgements across three criteria. **R**, **C** and **F** on the left side stand for relevance, correctness and composition (flow), corresponding to the type of questions that the crowdworkers were provided with. *P*, *S* and *K* stand for Pearson's, Spearman's and Kendall's correlations. We report correlation scores per search and per correlation metric. The scores coloured in red have $p < 0.05$.

Describe me an Auklet (2023)

Perceptual category recognition: Can you (textually) describe a new category to me so that I can (visually) recognise it?

Setup: Generator and Interpreter models, independently trained to either
GEN: image-classify & describe
IPT: listen and image-classify

Results are promising but mixed: decoding strategies matter, diversity is lacking.



Approximation sacrifices variability (2024)

Compare human and machine generated descriptions using a bigram model:

Train on (human), test on (human, model) and vice versa

Result: Human language is far more perplex/complex than model output

Metrics don't capture, or correlate with, complexity

		Human	BLIP				BLIP-2			
			M_{gr}	M_{be}	M_{co}	M_{nu}	M_{gr}	M_{be}	M_{co}	M_{nu}
	$\|\mathcal{V}\|$	6919	1222	1291	1300	2100	1476	1616	1711	2383
P	Human	***538.4***	**206.3**	**240.6**	**1248.5**	**791.5**	**188.2**	**183.5**	**419.9**	**614.0**
	M_{gr}	*357.0*	90.3	119.9	593.6	439.5	106.1	121.3	204.5	423.4
	M_{be}	*342.4*	90.8	93.6	563.8	401.6	109.3	110.2	203.5	397.3
	M_{co}	*448.6*	194.5	224.8	269.0	497.5	149.2	155.0	183.7	513.2
	M_{nu}	*438.0*	177.9	187.3	602.1	422.3	178.1	179.3	289.0	395.3
BERTSCORE ↑			0.843	0.847	0.787	0.797	0.864	0.868	0.844	0.818
MOVERSCORE ↑			0.609	0.615	0.567	0.583	0.623	0.629	0.613	0.597

Table 1: Performance of bi-gram language models trained on different speakers (Human, BLIP, BLIP-2) in terms of averaged perplexity **P**, BERTSCORE and MOVERSCORE. The columns and rows denote what the model has been trained and tested on, respectively. Models trained on machine-generated language were tested with different decoding algorithms: *gr* (greedy), *be* (beam), *co* (contrastive) or *nu* (nucleus)). $|\mathcal{V}|$ is the vocabulary size of the corresponding model. Numbers in **bold** are comparisons across models tested on human data, numbers in *italics* are perplexities of the model trained on human data and red cells denote perplexities of a model trained and tested on the same data. Numbers in blue are best-performing models.

What have I learned?

Need for task-specific models: because task context is such a strong constraint.

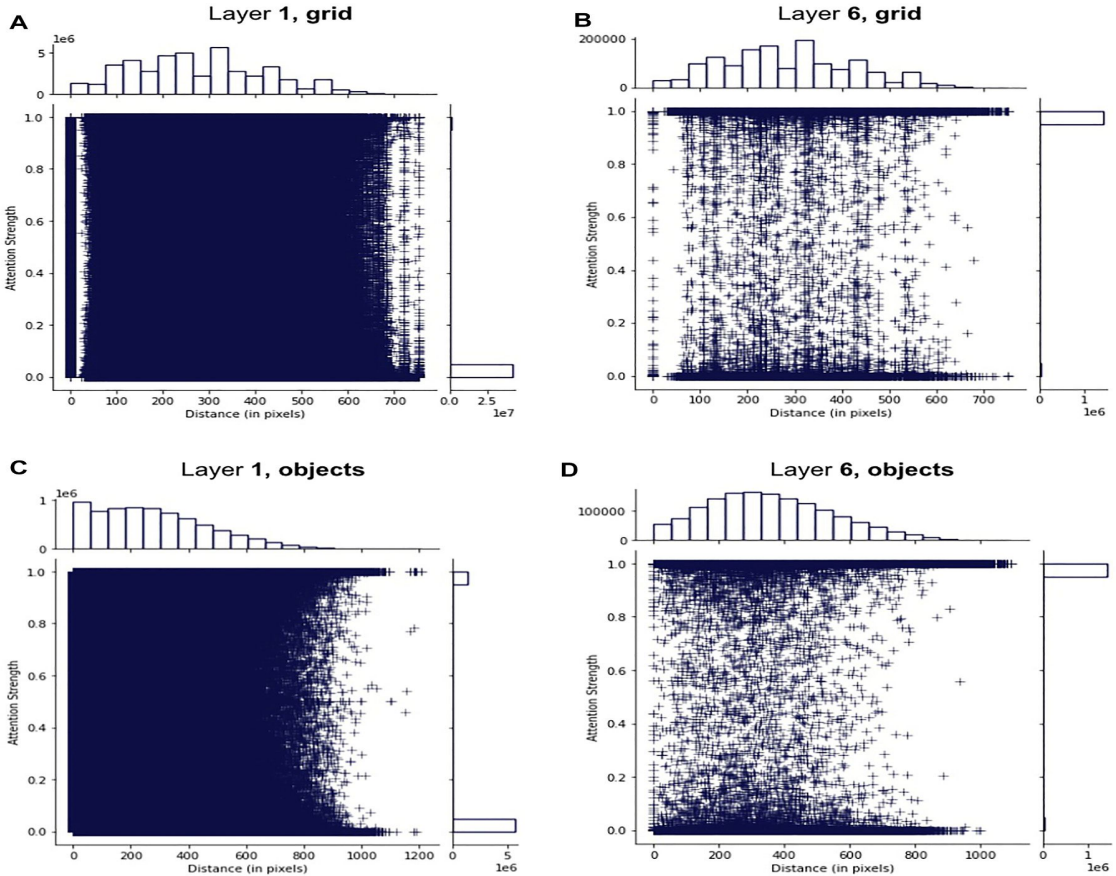
Abstracted visual features can be as effective as pixels.

Decoding algorithm matters hugely: even big models can produce small (boring, degenerate, overly-general) outputs.

Automatic evaluation measures correlate poorly with human desiderata

- different objects are named
- relevance/correctness are not captured (but maybe flow?)
- specificity, i.e. ability to discriminate, is essential, but only indirectly measured

What's next?



What Does a Language-And-Vision Transformer See: The Impact of Semantic Information on Visual Representations. Nikolai Ilinykh and Simon Dobnik. 2021. Frontiers in Artificial Intelligence

FIGURE 12 | The distribution of the attention links (depicted as +marks) in terms of their strength (the vertical axis) and distance between centres of two connected objects (the horizontal axis). **(A)** and **(B)** correspond to the patterns from the first and the last layer of the model for grid-based features (e.g., image patches), while **(C)** and **(D)** represent links when we use object representations as input features. We disregard objects' thematic clusters in these visualisations for a fair comparison with the grid-based approach. Information about other parts of the figures (e.g. histograms) is identical to the description in **Figure 6**.