Assessing Language and Vision-Language Models on Event Plausibility

Maria Cassese[1], Alessandro Bondielli[1,2] and Alessandro Lenci[1]

[1]CoLing Lab, Department of Philology, Literature, and Linguistics, University of Pisa, 36 S. Maria St, Pisa, I-56126, Italy
[2]Department of Computer Science, University of Pisa, 3 Largo Bruno Pontecorvo, Pisa, 56127, Italy

Abstract

Transformer-based *Language Models (LMs)* excel in many tasks, but they appear to lack robustness in capturing crucial aspects of event knowledge due to their reliance on surface-level linguistic features and the mismatch between language descriptions and real-world occurrences. In this paper, we investigate the potential of Transformer-based *Vision-Language Models (VLMs)* in comprehending *Generalized Event Knowledge (GEK)*, aiming to determine whether the inclusion of a visual component affects the mastery of GEK. To do so, we compare multimodal Transformer models with unimodal ones on a task evaluating the plausibility of curated minimal sentence pairs. We show that current VLMs generally perform worse than their unimodal counterparts, suggesting that VL pre-training strategies are not yet as effective to model semantic understanding and resulting models are more akin to bag-of-words in this context.

Keywords

multimodal semantics, vision language models, language models, generalized event knowledge,

1. Introduction

Humans have rich knowledge about events and their typical participants. This is known as **Generalized Event Knowledge** (GEK) [1]. GEK is a fundamental part of commonsense knowledge, and plays a key role in language processing as well as in reasoning. For instance, GEK supports our intuitions about likely events (e.g., *A cop arrested a thief*), possible but implausible events (e.g., *A thief arrested a cop*), and impossible events (e.g., *A stone arrested a thief*). Event knowledge is intuitive for humans because we perceive the world by simultaneously processing information from different modalities such as textual, visual, and auditory [2]. In fact, GEK is acquired through linguistic (e.g., reading and talking about events) and sensorimotor experiences based on observing and participating in real-world events.

Several works have investigated to what extent Language models (LMs) possess GEK [3, 4]. These analyses reveal that LMs have remarkable aspects of GEK, though with important differences with respect to humans. This prompts the question whether such differences might stem from the way LMs acquire their knowledge. In fact, even the most recent Transformer-based ones [5], do not possess the same level of multimodal integration of human learners, since they are trained solely on textual data, lacking key visual information like an object's shape and color. In this context, it is natural to ask whether the recently introduced Vision-Language Models (VLMs) possess capabilities that surpass those of text-only LMs in modelling GEK due to their multimodal knowledge of the world. Recent literature has shown that language interpretation appear to not be improved using multimodal architectures [6], and that in some cases VLMs behave as bag-of-words models when it comes to interpreting texts [7, 8].

We contribute to this line of research by carrying out a comparative study of the performance of LMs and VLMs in recognizing event plausibility. The dataset is formed by sentences that differ for the degree of plausibility of the event they express and the argument animacy. Furthermore, we explore the effect of event concreteness on the performances of the models. Finally, we evaluate the impact of actually including images describing test events sentences as inputs for multimodal models. Our analyses reveal that VLMs do not exhibit better performances than LMs on semantic plausibility recognition, with or without images as inputs. Further, we show how more challenging sentences impact the performances of VLMs, suggesting that they are less capable than LMs in recognizing semantic differences that are affected by word orders (e.g., with subject-patient inversion).

This paper is organized as follows. In Section 2 we describe related work. Then, Section 3 details the datasets (Sec. 3.1), the tested models (Sec. 3.2), and the evaluation procedure (Sec. 3.3). We show and discuss the obtained results in Section 4. Finally, Section 5 draws some conclusions and highlights possible future works.

CLiC-it 2023: 9th Italian Conference on Computational Linguistics, Nov 30 — Dec 02, 2023, Venice, Italy
m.cassese4@studenti.unipi.it (M. Cassese);
alessandro.bondielli@unipi.it (A. Bondielli);
alessandro.lenci@unipi.it (A. Lenci)
0009-0007-4765-1221 (M. Cassese); 0000-0003-3426-6643 (A. Bondielli); 0000-0001-5790-4308 (A. Lenci)

2. Related Work

The introduction of multimodal models in NLP stems from the intrinsic limitations of computational models that are trained exclusively on distributional statistics extracted from textual data [9, 3, 10]. In fact, they lack referential competence [11], which prevents them from grounding linguistic structures onto real-world experiences [12, 13].

The earliest multimodal distributional models already showed the ability to improve the semantic representation of concrete concepts and properties [14], as well as abstract verbs that lack direct perceptual information, but benefit from integrating linguistic inputs and perceptual information [15]. However, they proved to be less effective in representing verbs, adjectives, and abstract concepts [16].

The introduction of Transformer-based VLMs such as Visual-BERT [17] and FLAVA [18], and effective techniques for Vision-Language Pre-training [19] paved the way for new research in a multimodal setting. While numerous studies has shown VLMs success on different multimodal tasks, less effort has been put in analyzing their differences with unimodal counterparts on natural language understanding (NLU). [20] show that both dual-stream and single-stream VLMs are equally capable of preserving NLU capabilities. The analysis conducted by [6] shows that multimodal models do not significantly outperform the text-only variants in a language-only setting. This was attributed to the use of narrow domain data and direct extensions of NLP architectures. Our work support these findings by focusing on understanding the plausibility of events.

3. Experiments

Our goal is to evaluate the ability of LMs and VLMs to predict the semantic plausibility of sentences with respect to human judgements. In the following, we describe the data used in the experiments (Sec. 3.1), the models we considered in the evaluation (Sec. 3.2), and detail the evaluation procedure itself (Sec. 3.3).

3.1. Data

We sourced our data from a number of existing datasets containing pairs of sentences describing transitive event distinguished by patient plausibility within the context of the sentence. Plausibility is rated by humans and expressed on a 1-7 Likert scale. Formally, each data point consists of a plausible sentence S_p and its corresponding implausible one S_i obtained through a modification of S_p.

We considered the following datasets:

DTFit [21]. It includes past tense sentences distinguished by patient prototipicality. For each plausible sentence, the implausible (i.e., atypical) one is obtained by replacing the patient with an atypical filler for that role (e.g., *The actor won the award* vs *The actor won the battle)*.

EventsAdapt [22]. It includes pairs of plausible-implausible sentences where the implausible one is obtained by reversing the noun phrases (e.g., *The cop arrested the criminal* vs. *The criminal arrested the cop*). The dataset is divided into two sub-datasets. In the former, henceforth referred to as $EventsAdapt_{AN-AN}$, both the agent and the patient are animate. In the latter, denoted as $EventsAdapt_{AN-IN}$, the agent of the original sentence is animate, while the patient is not. Thus the implausible sentence is also semantically impossible.

EventsRev [23]. It includes concrete sentences describing events in the present progressive tense. Like EventsAdapt, implausible sentences are obtained by reversing the noun phrases, which in this case always depict animate entities (e.g., *The cat is chasing the mouse* vs. *The mouse is chasing the cat*). Each sentence, both plausible and implausible, is accompanied by an image depicting the interaction between the two animated participants described in the sentence. The images are simple black and white drawings.

As we are interested in considering also the effect of concreteness in VLMs' ability to recognize plausibility, we further grouped sentences of EventsAdapt (and its subgroups) and DTFit into concrete and abstract ones. We categorized the sentences based on the level of concreteness of the verb, subject, and object in each sentence. We chose to consider sentences that refer to abstract concepts with high imageability as concrete (e.g., *The priest celebrated the marriage*).

To the best of our knowledge, none of the data used in this study was included in the training set of the evaluated models.

3.2. Models

We test various popular multimodal VLMs and compare them with baseline unimodal LMs: BERT [24] and RoBERTa [25]. As for the VLMs, our analysis includes:

VisualBERT [17]. A single-stream early fusion encoder model initialized from pre-trained BERT-base weights and further trained on multimodal datasets. Visual features are extracted from a pre-trained Faster R-CNN network [26] and fed into the transformer model alongside the text.

LXMERT [27]. A dual-stream early fusion encoder model including some modality-specific layers and allowing cross-attention in specific co-attention layers. Visual features are extracted with a Faster R-CNN network.

ViLT [28]. A single-stream model employing a BERT model for textual feature extraction and a ViT model for visual feature extraction, respectively. Resulting representations are then concatenated and fed into the final model.

FLAVA [18]. A foundation VLM including an image encoder, a textual encoder, and a multimodal encoder. It is jointly pre-trained on both unimodal and multimodal data, thus learning high-quality visual and textual representations. It is capable of achieving both crossmodal alignment and multimodal fusion objectives.

To adapt multimodal models to the text-only task, we simply modified the inputs, e.g. by feeding them empty image tensors. FLAVA does not require to be adapted to text-only inputs, as it can directly be evaluated by using only the textual encoder.

All models and their pre-trained weights are available on Huggingface Transformers [29].[1]

3.3. Evaluation procedure

To evaluate the ability of a LM (or VLM) to distinguish between plausible and implausible sentences, we first have to compute a plausibility score for each sentence. Since we are dealing with bi-directional masked language models, we can approximate this plausibility score via pseudo-log-likelihood (PLL), defined as the sum of logarithmic probabilities of each token based on the remaining tokens in the sentence [30]. To avoid bias favoring multi-token words, we apply an additional mask that covers tokens to the right of the target, as proposed in [4]. To compare PLL scores with human judgements expressed on a Likert scale, we normalized both using a min-max scaler function.

First, we evaluated the models using an accuracy metric. Specifically, considering all (S_p, S_i) sentence pairs for a dataset, we computed accuracy as the percentage of cases for which $PLL(S_p) > PLL(S_i)$.

To provide a more detailed analysis of the performances, we further evaluate the models via distribution analyses. We used the Pearson correlation coefficient between each model's score for the plausible and implausible sentences. More in detail, for each pair of (S_p, S_i), we plot the correlation between normalized $PLL(S_p)$ and $PLL(S_i)$. High correlation implies similar scores for plausible and implausible sentences, indicating that the model is less able to distinguish between them. Thus, negative correlation values indicate good performances. We also analyzed the density of the distributions for PLLs. This is essential to comprehend how humans and models differentiate between plausible and implausible classes, aiding in evaluating sentence complexity and comparing model behavior to humans'.

4. Results and Discussion

We first verify the performances of VLMs in plausibility recognition via accuracy. Results of all models on the datasets are reported in Table 1.

Both LMs and VLMs show significantly higher performances on $EventsAdapt_{AN-IN}$, where implausible sentences describe impossible events, than on $EventsAdapt_{AN-AN}$ where implausible sentences depict unlikely but not impossible events. On AN-AN sentences, BERT, RoBERTa, VisualBERT, and FLAVA performed above chance levels, while ViLT and LXMERT performed at chance. This indicates that extracting information about AN-AN sentence plausibility is generally challenging, and more so for VLMs. Among VLMs, FLAVA performs best, with results generally close to RoBERTa.

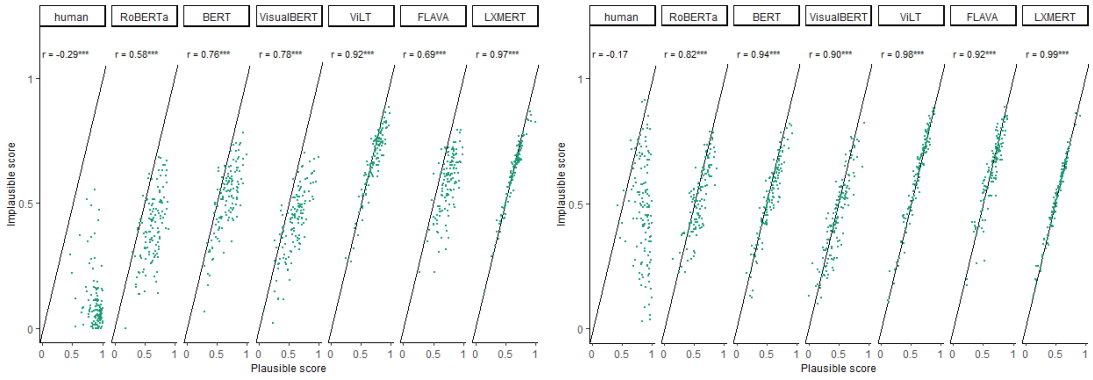
Going further, we consider EventsAdapt and we provide the density plot of PLLs divided by plausibility for each model and human raters in Figure 2, and plot the correlation between PLLs of plausible and implausible sentences in Figure 1.

Both LMs and VLMs do not clearly distinguish between the two classes and exhibit very similar distributions for plausible and implausible sentences. The complexity of the task affect the results as well: for tasks where humans have no difficulty in distinguishing between the two classes, as the implausible sentence violates the verb selection preferences (AN-IN), the models can better identify patterns that differentiate the two sentences (Fig 1a); for tasks where even humans are more uncertain (AN-AN), the models tend to assign very similar scores to the two sentences (Fig. 1b). This is also clearly shown by the density distribution plot in Figure 2. for the AN-IN case, the density distribution for humans show a clear separation, while models show more modest but still evident signs of separation. The density distribution for AN-AN sentences shows a less separated distribution for human scores and almost entirely overlapped distributions for models. One possible reason for this is that the grammaticality of a sentence depends on syntactic rules that can be more easily detected through statistical inference. In contrast, linguistic acceptability may depend on extralinguistic information requiring multiple inference levels.

[1] https://huggingface.co/

Table 1
Model accuracy on the different datasets

Dataset	Size	Human	BERT	RoBERTa	VisualBERT	LXMERT	ViLT	FLAVA
$DTFit$	395	0.99	0.85	0.89	**0.90**	0.70	0.80	0.86
$EventsAdapt_{AN-IN}$	128	1.00	0.93	**0.95**	0.93	0.72	0.84	**0.95**
$EventsAdapt_{AN-AN}$	129	0.95	**0.78**	**0.78**	0.64	0.53	0.50	0.66
$EventsRev$	38	1.00	0.76	**0.79**	0.76	0.66	0.76	**0.79**



(a) AN-IN (128 tuples)

(b) AN-AN (129 tuples)

Figure 1: Correlation plot on the EventsAdapt dataset for plausible and implausible sentences tuples. Significant differences are marked with asterisks ($*p < 0.05, **p < 0.01, ***p < 0.001$).

The role of concreteness. The experiments show that VLMs do not show improved abilities to deal with GEK and event plausibility with respect to textual LMs. However, we could expect that this might also depend on the event concreteness, as concrete concepts are more directly grounded on visual information than abstract ones. The concreteness of an event depends on the the predicate itself, as well on its arguments. For instance, the verb *to fight* has a concrete use in the sentence *The wrestler fought the opponent* and an abstract use in *The patient fought the cancer*. Cognitive research has shown that abstract concepts require more linguistic experiences to be understood [31]. Thus they are generally more difficult to acquire and process. This is influenced by two main factors, namely imageability and familiarity [32]. For instance, the abstract verb *to celebrate* becomes more concrete in the context of *The priest celebrated the wedding* because it is easier to form mental images of the event and it is very frequent in language use.

To evaluate how concreteness affects the models' abilities, we first compute the accuracy on concrete and abstract subsets of the dataset. Results are reported in Table 2. Multimodal models seem to perform worse on abstract sentences with a higher degree of complexity: on the $EventsAdapt_{AN-AN}$ dataset, the average performance gap between abstract and concrete sentences is higher for VLMs than for LMs (0.06 for LMs, 0.09 for VLMs); when considering the simpler sentences of $EventsAdapt_{AN-IN}$, the differences are less marked. On the other hand, multimodal models demonstrate excellent recognition of abstract events in the DTFit dataset. Note however that abstract sentences are an order of magnitude less than concrete ones in the dataset.

We also show a comparison of Pearson correlation scores of results between $EventsAdapt^{concr}_{AN-AN}$ and $DTFit^{concr}$, shown respectively in Figures 3a and 3b. While VLMs exhibit high correlation values, i.e. less prowess on the task, values for DTFit are generally lower, suggesting a better ability to assess plausibility. VLMs' performance difference in the two datasets may be due to how implausible sentences are generated. EventsAdapt uses noun phrase order reversal, while DTFit only replaces the typical patient with an incompatible one. If VLMs behave more like bag-of-words models, they may struggle to recognize semantic differences between sentences with the same words but different order. This would explain their worse performances on $EventsAdapt_{AN-AN}$.

The impact of images Finally, we analyze whether including images of the (im)plausible test events in the input is beneficial for VLMs. We provide accuracy scores

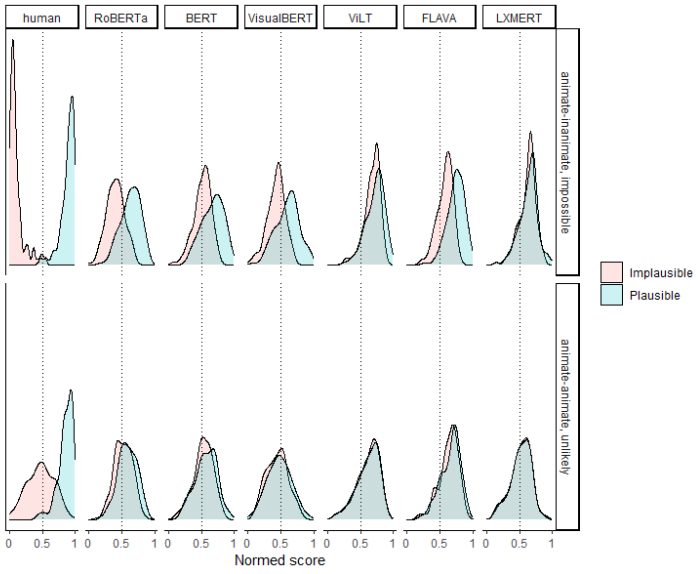


Figure 2: Density plots (EventsAdapt AN-IN (128 pairs) and AN-AN (129 pairs))

Table 2
Accuracy on DTFit and EventsAdapt sentences distinguished by concreteness.

Dataset	Size	Human	BERT	RoBERTa	VisualBERT	LXMERT	ViLT	FLAVA
$DTFit^{abstr}$	45	0.99	0.89	0.86	**0.93**	0.55	0.80	**0.93**
$DTFit^{concr}$	350	0.99	0.85	**0.90**	0.89	0.72	0.80	0.85
$EventsAdapt^{abstr}_{AN-IN}$	31	1.00	0.87	0.94	0.90	0.71	0.71	**0.97**
$EventsAdapt^{concr}_{AN-IN}$	97	1.00	**0.95**	**0.95**	0.94	0.72	0.88	**0.95**
$EventsAdapt^{abstr}_{AN-AN}$	64	0.96	0.75	**0.80**	0.56	0.47	0.47	0.62
$EventsAdapt^{concr}_{AN-AN}$	65	0.96	**0.82**	0.75	0.70	0.57	0.55	0.67

for VLMs on the EventsRev dataset in Table 3. Including event images does not lead to any improvement: performances either remain the same or slightly degrade.

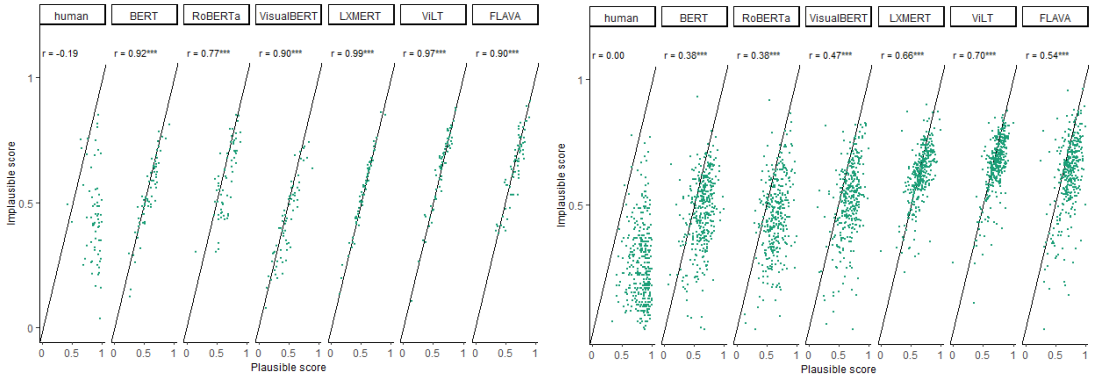
Dataset	VisualBERT	LXMERT	ViLT	FLAVA
$EventsRev_t$	0.76	0.66	0.76	**0.79**
$EventsRev_{t+i}$	0.61	0.66	0.71	**0.79**

Table 3
Accuracy of VLMs on EventsRev with $(t + i)$ and without (t) images in the input.

4.1. Discussion

Several interesting findings have emerged from our analysis. First, VLMs do not achieve significantly higher accuracy values than unimodal ones in a semantic plausibility recognition task. Second, we saw that performances of VLMs is worse when dealing with more challenging sentences represented by $EventsAdapt_{AN-AN}$, exhibiting lower accuracy and a high correlation between plausible and implausible sentences. Third, we saw that including images of events in the input does not lead to improved model performances.

We discuss a possible interpretation of these findings in the following. First, the generally high correlation between PLL scores for pairs of (S_p, S_i) for VLMs suggest that these models struggle to recognize semantic differences, especially between sentences with different word orders (e.g., with subject-patient inversion), and relationships between sentence components, like semantic roles. This may be further indication that VLMs model language in a bag-of-words fashion [7, 8]. The pre-training method used in masked language modelling for VLMs, adding visual features to language models already specialized on linguistic tasks, may also compromise learning as suggested by [33]. The high-dimensional space learnt by these models could make it difficult to identify se-



(a) $EventsAdapt^{concr}_{AN-AN}$.

(b) $DTFit^{concr}$.

Figure 3: Correlation plots for sentences in $EventsAdapt^{concr}_{AN-AN}$ and $DTFit^{concr}$.

mantic errors. Moreover, models using pre-trained LM weights for text processing may face limitations in the type of visual information they can capture during training. Some models rely on object categories trained on bounding boxes. This is computationally expensive, and the learned representations may not adequately capture shapes and relationships. Other models, such as ViLT, that leverage ViT representations and use a linear function to extract embeddings for image patches, are less costly but may result in lower-quality representations. These results are in line with [33].

A possible explanation of why VLMs do not benefit from including test images is that in this specific case (minimal sentence pairs with subject-object inversion) the images for both sentences are very similar, and differ only for the relationship between the entities. The visual encoders of the models might be too weak to differentiate substantially similar images, leading the models to rely on their LM priors and make random choices. Finally, we saw that even the foundation Large VLM we considered – FLAVA – does not show significantly improved accuracy compared to other VLMs.

5. Conclusions and Future Works

In this paper, we presented a set of experiments aimed at evaluating the ability of VLMs to model event plausibility in both language-only and vision-language tasks against LMs. We find that VL pre-training does not lead to a significant improvement compared to unimodal LMs in this task aiming at testing their GEK. Specifically, we observed that VLMs tend to perform worse when the implausible sentence has a higher semantic complexity, because it contains two animate nouns. Our analysis also brings further support to argument that VLMs models still behave similarly to Bag-of-Words models, regardless of the degree of concreteness of the events.

In the future, we plan to focus on the analysis of models with visual grounding as their training objective, such as PaLM-E [34], a large embodied multimodal language model that directly incorporates real-world continuous sensor modalities into language processing. This may shed more light into the abilities of large multimodal models to achieve more human-level grounded language understanding.

Acknowledgments

Research partially supported by the Italian Ministry of University and Research (MUR) in the framework of the PON 2014-2021 “Research and Innovation" resources – Innovation Action - DM MUR 1062/2021 - Title of the Research: “Modelli semantici multimodali per l’industria 4.0 e le digital humanities.”, and by PNRR - M4C2 - Investimento 1.3, Partenariato Esteso PE00000013 - “FAIR - Future Artificial Intelligence Research" - Spoke 1 “Human-centered AI", funded by the European Commission under the NextGeneration EU programme.

References

[1] K. McRae, K. Matsuki, People use their knowledge of common events to understand language, and do so as quickly as possible., Lang Linguist Compass. 3(6) (2009) 1417–1429. doi:10.1111/j.1749-818X.2009.00174.x.

[2] T. Baltrušaitis, C. Ahuja, L.-P. Morency, Multimodal machine learning: A survey and taxonomy, IEEE transactions on pattern analysis and machine intelligence 41 (2018) 423–443.

[3] P. Pedinotti, G. Rambelli, E. Chersoni, E. Santus, A. Lenci, P. Blache, Did the cat drink the coffee? challenging transformers with generalized event knowledge, arXiv preprint arXiv:2107.10922 (2021).

[4] C. Kauf, A. A. Ivanova, G. Rambelli, E. Chersoni, J. S. She, Z. Chowdhury, E. Fedorenko, A. Lenci, Event knowledge in large language models: the gap between the impossible and the unlikely, 2022. doi:10.48550/ARXIV.2212.01488.

[5] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin, Attention is all you need, in: Proceedings of the 31st International Conference on Neural Information Processing Systems, NIPS'17, Curran Associates Inc., Red Hook, NY, USA, 2017, p. 6000–6010.

[6] T. Yun, C. Sun, E. Pavlick, Does vision-and-language pretraining improve lexical grounding?, in: Findings of the Association for Computational Linguistics: EMNLP 2021, Association for Computational Linguistics, Punta Cana, Dominican Republic, 2021, pp. 4357–4366. doi:10.18653/v1/2021.findings-emnlp.370.

[7] S. Castro, O. Ignat, R. Mihalcea, Scalable performance analysis for vision-language models, in: Proceedings of the The 12th Joint Conference on Lexical and Computational Semantics (*SEM 2023), Association for Computational Linguistics, Toronto, Canada, 2023, pp. 284–294.

[8] T. Thrush, R. Jiang, M. Bartolo, A. Singh, A. Williams, D. Kiela, C. Ross, Winoground: Probing vision and language models for visio-linguistic compositionality, 2022. arXiv:2204.03162.

[9] J. Gordon, B. Van Durme, Reporting bias and knowledge acquisition, in: Proceedings of the 2013 Workshop on Automated Knowledge Base Construction, AKBC '13, Association for Computing Machinery, New York, NY, USA, 2013, p. 25–30. doi:10.1145/2509558.2509563.

[10] A. Lenci, M. Sahlgren, Distributional Semantics, Cambridge University Press, Cambridge, 2023.

[11] D. Marconi, Lexical Competence, The MIT Press, Cambridge, MA, 1997.

[12] S. Harnad, The symbol grounding problem, Physica D: Nonlinear Phenomena 42 (1990) 335–346. doi:https://doi.org/10.1016/0167-2789(90)90087-6.

[13] E. M. Bender, A. Koller, Climbing towards nlu: On meaning, form, and understanding in the age of data, in: Proc. ACL, Seattle, WA, 2020, pp. 5185–5198.

[14] E. Bruni, G. Boleda, M. Baroni, N.-K. Tran, Distributional semantics in technicolor, Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (2012) 136–145.

[15] F. Hill, A. Korhonen, Learning abstract concept embeddings from multi-modal data: Since you probably can't see what I mean, in: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, Doha, Qatar, 2014, pp. 255–265. doi:10.3115/v1/D14-1032.

[16] R. Shekhar, S. Pezzelle, Y. Klimovich, A. Herbelot, M. Nabi, E. Sangineto, R. Bernardi, FOIL it! find one mismatch between image and language caption, in: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Vancouver, Canada, 2017, pp. 255–265. doi:10.18653/v1/P17-1024.

[17] L. H. Li, M. Yatskar, D. Yin, C.-J. Hsieh, K.-W. Chang, Visualbert: A simple and performant baseline for vision and language, arXiv preprint arXiv:1908.03557 (2019).

[18] A. Singh, R. Hu, V. Goswami, G. Couairon, W. Galuba, M. Rohrbach, D. Kiela, Flava: A foundational language and vision alignment model, 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2021) 15617–15629.

[19] Z. Gan, L. Li, C. Li, L. Wang, Z. Liu, J. Gao, et al., Vision-language pre-training: Basics, recent advances, and future trends, Foundations and Trends® in Computer Graphics and Vision 14 (2022) 163–352.

[20] T. Iki, A. Aizawa, Effect of visual extensions on natural language understanding in vision-and-language models, in: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 2021, pp. 2189–2196. doi:10.18653/v1/2021.emnlp-main.167.

[21] P. Vassallo, E. Chersoni, E. Santus, A. Lenci, P. Blache, Event Knowledge in Sentence Processing: A New Dataset for the Evaluation of Argument Typicality, in: LREC 2018 Workshop on Linguistic and Neurocognitive Resources (LiNCR), Miyazaki, Japan, 2018.

[22] E. Fedorenko, I. A. Blank, M. Siegelman, Z. Mineroff, Lack of selectivity for syntax relative to word meanings throughout the language network, Cognition 203 (2020) 104348. doi:https://doi.org/10.1016/j.cognition.2020.104348.

[23] A. A. Ivanova, Z. Mineroff, V. Zimmerer, N. Kanwisher, R. Varley, E. Fedorenko, The Language Network Is Recruited but Not Required for Nonverbal Event Semantics, Neurobiology of Language 2 (2021) 176–201. doi:10.1162/nol_a_00030. arXiv:https://direct.mit.edu/nol/article-pdf/2/2/176/189

[24] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for

language understanding, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 4171–4186. doi:10.18653/v1/N19-1423.

[25] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, Roberta: A robustly optimized bert pretraining approach, 2019. Cite arxiv:1907.11692.

[26] S. Ren, K. He, R. Girshick, J. Sun, Faster r-cnn: Towards real-time object detection with region proposal networks, Advances in neural information processing systems 28 (2015).

[27] H. Tan, M. Bansal, Lxmert: Learning cross-modality encoder representations from transformers., in: K. Inui, J. Jiang, V. Ng, X. Wan (Eds.), EMNLP/IJCNLP (1), Association for Computational Linguistics, 2019, pp. 5099–5110.

[28] W. Kim, B. Son, I. Kim, Vilt: Vision-and-language transformer without convolution or region supervision, in: International Conference on Machine Learning, 2021.

[29] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. Le Scao, S. Gugger, M. Drame, Q. Lhoest, A. Rush, Transformers: State-of-the-art natural language processing, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, Association for Computational Linguistics, Online, 2020, pp. 38–45. doi:10.18653/v1/2020.emnlp-demos.6.

[30] J. Salazar, D. Liang, T. Q. Nguyen, K. Kirchhoff, Masked language model scoring, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Online, 2020, pp. 2699–2712. URL: https://aclanthology.org/2020.acl-main.240. doi:10.18653/v1/2020.acl-main.240.

[31] G. Vigliocco, L. Meteyard, M. Andrews, S. T. Kousta, Toward a theory of semantic representation, Language and Cognition 1 (2009) 219–247.

[32] G. Löhr, Does the mind care about whether a word is abstract or concrete? why concreteness is probably not a natural kind, Mind & Language n/a (2023). doi:https://doi.org/10.1111/mila.12473.

[33] C. Fields, C. Kennington, Vision language transformers: A survey, 2023. arXiv:2307.03254.

[34] D. Driess, F. Xia, M. S. M. Sajjadi, C. Lynch, A. Chowdhery, B. Ichter, A. Wahid, J. Tompson, Q. Vuong, T. Yu, W. Huang, Y. Chebotar, P. Sermanet, D. Duckworth, S. Levine, V. Vanhoucke, K. Hausman, M. Toussaint, K. Greff, A. Zeng, I. Mordatch, P. Florence, Palm-e: An embodied multimodal language model, 2023. arXiv:2303.03378.