Are language-and-vision transformers sensitive to discourse?
A case study of ViLBERT

Ekaterina Voloshina [†] and **Nikolai Ilinykh** [*] [†] and **Simon Dobnik** [*] [†]
[*]Centre for Linguistic Theory and Studies in Probability (CLASP),
Department of Philosophy, Linguistics and Theory of Science (FLoV),
[†]University of Gothenburg, Sweden
gusvolek@student.gu.se,
nikolai.ilinykh@gu.se, simon.dobnik@gu.se

Abstract

Language-and-vision models have shown good performance on tasks such as image-caption matching and caption generation. However, it is challenging for such models to generate *pragmatically* correct captions, which adequately reflect what is happening in one or several images. Here we explore to what extent contextual language-and-vision models are sensitive to different discourses, both textual and visual. In particular, we employ one of the multimodal transformers (ViLBERT) and test if it can match descriptions and images, differentiating them from distractors of different degree of similarity that are sampled from different visual and textual contexts. We place our evaluation in the multi-sentence and multi-image setup, where images and sentences are expected to form a single narrative structure, e.g. discourse. We show that the model can distinguish different situations but it is not sensitive to differences within one narrative structure. We also demonstrate that the model's performance depends on the task itself, for example, the effect of what modality remains unchanged in non-matching pairs or how similar non-matching pairs are to the original pairs.

1 Introduction

Large language models are considered "black boxes" as it is often hard to explain their predictions. Therefore, it is essential to evaluate such models on tasks different from downstream applications to see if they have acquired necessary knowledge, including linguistic knowledge, during the pre-training process (Liu et al., 2019; Belinkov, 2022; Elazar et al., 2021). Moreover, model performance on linguistic tasks has been shown to correlate with downstream applications (Saphra, 2021). Although the models have been tested on morphology, semantics, and syntax (Conneau et al., 2018; Warstadt et al., 2020; Taktasheva et al., 2021; Stańczak et al., 2022; Maudslay and Cotterell, 2021; Lasri et al., 2022), little has been done on the evaluation of discourse (Hong et al., 2020; Liang et al., 2022). Models such as GPT-3 are good at generating long, coherent sequences but they are often not sensitive to intents or communicative purposes that the narratives are written with (Ruis et al., 2022).

In this work, we focus on the evaluation of the sensitivity of multi-modal models to *discourse*. We understand discourse as contextual information defined by the situation of communication connected to social and cultural background. Discourse can be **local** defined by individual items such as an utterance or an image or **narrative** defined by a sequence of several items. Discourse meaning includes but is not limited to:

- **Textual discourse**: linguistic relations and dependencies that exist between linguistic units across words in a single utterance;
- **Visual discourse**: images represent parts of reality by being focused on specific situations involving entities and events among all that are visually available;
- **Situation-level discourse**: operates at a level of larger structures such as narratives and requires world knowledge and awareness of the social context.

As we are interested in how a language-and-vision model in general captures all three components of discourse outlined above and to investigate the effects and the role of discourse on the generation of text across different discourses, for experiments we evaluate a single multi-modal transformer, ViLBERT (Lu et al., 2019), that was trained on a (simple) discourse task of image-text matching. This is also one of the first transformer-based models with a cross-modal alignment pre-training objective. For the data we use the Visual Storytelling dataset (VIST) (Huang et al., 2016) which

Proceedings of the Workshop on Multimodal, Multilingual Natural Language Generation and Multilingual WebNLG Challenge (MM-NLG 2023), pages 28–38
September 12, 2023 ©2023 Association for Computational Linguistics

consists of descriptions of images at two different levels of annotation and therefore discourses: *descriptions-in-isolation* and *stories-in-sequence*. While *descriptions-in-isolation* were collected following standard instructions for image captioning in MS-COCO (Lin et al., 2014) (for example, annotators have to name all important parts, i.e. *local discourse*), for *stories-in-sequence* crowd workers were asked to write a story about a sequence of images (*narrative discourse*, see Figure 1). *Stories-in-sequence* do not tend to name objects in images but rather connect images to a coherent situation, and therefore we expect them to be more challenging for a language-and-vision model.

We assume that during pre-training a model acquires the knowledge of situation described by a specific image-text pair. Since we understand discourse as a situation itself rather than a sequence of images, we adapt a standard image-caption matching task to our experiments by creating different perturbations, i. e. non-matching pairs that we call **distractors**, by permuting data in several ways, across stories and within one story. We investigate the model's abilities of grounding discourse at different levels:

1. Can a language-and-vision model perform *visual grounding of individual descriptive captions* when non-matching images or captions are taken from different situations?

2. Can a language-and-vision model perform *story visual grounding*, i. e. ground captions that are parts of a narrative but there is a mismatch between images or captions and the situation?

3. Can a language-and-vision model understand the narrative structure, i. e. the model can detect if two parts of the same story form a coherent whole?

In the first two cases we are testing the model's ability to identify distractors across different stories and situations, hence testing the model's ability to identify *local discourse*. In the third case, we are testing the model's ability to identify distractors within the same narrative and hence we are testing the model's ability to identify *narrative discourse*.

We introduce a new method of constructing evaluation tasks for language-and-vision models. We then analyse the role of different discourse structures and different modalities on the performance on our tasks. Our results reveal details about the behaviour of large multi-modal transformer for a setup that is beyond simple image-text matching.

2 Related Work

Little has been done on evaluating models' knowledge of discourse. For language models Ettinger (2020) shows that BERT produces pragmatically incorrect outputs as it does not take into account broader context. Most of the previous work on discourse evaluation of BERT and BERT-like models was focused on local discourse structures. For example, Nie et al. (2019) evaluate models on explicit discourse relations expressed with conjunctions. Chen et al. (2019) propose a benchmark for model evaluation on different discourse tasks such as prediction of implicit discourse relations based on the Penn Discourse Treebank annotation (Prasad et al., 2008), discourse coherence, and others. Araujo et al. (2021) attempt to improve the results on discourse tasks from DiscoEval by changing the pretraining objective of models. Recently, Hong et al. (2023a) introduced a dataset for image-based story generation and proposed a character-based story generation model that is evaluated based on the coherence of generated texts.

Multimodal models were shown to ground semantic and syntactic knowledge. Ilinykh and Dobnik (2022a) provide results that the models ground both semantic and syntactic relations. Parcalabescu et al. (2021) argue that models can ground objects but struggle with interdependence relations. Most of the work on discourse evaluation of language-and-vision models focuses on downstream tasks such as Visual Question Answering (VQA) and Visual Coreference. Bernardi and Pezzelle (2021) provide an overview of VQA systems and their challenges related to reasoning, language ambiguity etc. Several studies address the problem of discourse-coherent image generation. Takmaz et al. (2020) introduce a generation mechanism that produces captions grounded not only in the visual context but also in the established common ground. Alikhani et al. (2020) improve the quality of generated captions by feeding models with additional information on the types of connections between two clauses. Ilinykh and Dobnik (2022b) show how different decoding strategies for image captioning reflect discourse structure in comparison to reference captions.

As for existing image-caption datasets, Alikhani and Stone (2019) argue that existing datasets do not reflect all possible coherence types and are limited in the word usage. However, even existing datasets are challenging for models, as (Alikhani et al., 2023) show models such as CLIP or ViLBERT do not capture differences in coherence relations. However, recently discourse has started to receive increased attention in the research community, including development of new approaches to building discourse datasets (Hong et al., 2023b).

3 Models

We examine ViLBERT (Vision-and-Language BERT) (Lu et al., 2019), a dual-stream multi-modal BERT-based model. Unlike single-stream models that encode both modalities at the same time, dual-stream models initially represent each modality separately. These models then learn cross-modal grounding which should include some knowledge of discourse structure, depending on the context in which data was collected. ViLBERT performs well on the image-text matching task and it is frequently used in the studies of multi-modality. One important feature of ViLBERT is that its text module's parameters are initialised with BERT. In comparison, LXMERT (Tan and Bansal, 2019), a very similar model to ViLBERT, learns its text module from scratch. Although there are more recent transformers trained for different discourse contexts we focus on a transformer that was trained on a simple referring context (the properties of which are well-known to us) in order to investigate how well the knowledge of that discourse transfers to other (more complex) discourses. We expect that the performance of a model will vary depending on how well its pre-training objective(s) and datasets that it was trained on match the visual storytelling task.

ViLBERT was trained on three objectives: masked language modelling, image region masking, and an image-text matching task. In the latter task, the model has to predict if a given sequence of tokens consisting of a visual input (an image) and a language input (a caption) match or do not match. This objective helps the model to ground descriptions in images and differentiate them from non-matching counter-parts. This task therefore also involves learning about contextual sensitivity of descriptions to discourse and therefore, provided that non-matching items are controlled, we see it as a suitable task for our investigation.

We test the ViLBERT model from VOLTA (Bugliarello et al., 2021), a framework that provides the code base for several transformer-based language-and-vision encoders and allows working with custom datasets. The model takes a set of pre-processed features: masked sentences, token ids, visual features, image location, masked images with regions of interest and their object labels. Textual features are generated by the BERT tokeniser[*] that returns a sequence of token ids. To extract image features, we use the Caffe VG Faster R-CNN implementation (Anderson et al., 2018)[†]. The model extracts 36 proposal boxes with features of dimension 2048 and object labels. We do not mask any tokens or regions as we only focus on one output head of ViLBERT that predicts if an image and a text match.

4 Data

For experiments we use images and descriptions from the Visual Storytelling Dataset (VIST) (Huang et al., 2016). The dataset includes stories from 10,117 Flickr albums containing 210,819 images split into train (80%), dev (10%), and test (10%) samples. Stories reflect narrative structure and sentences are linked with discourse relations. An example of a story taken from the dataset is illustrated in Figure 1.

Although there are other datasets that include narrative captions, such as RecipeQA (Yagcioglu et al., 2018), we choose VIST as it includes several types of captions, which allows us to look at different discourse structures. The dataset includes three levels of description annotation, *descriptions-in-isolation, descriptions-in-sequence*[‡], and *stories-in-sequence*, by crowd-workers hired through Amazon's Mechanical Turk. For *stories-in-sequence* a worker selected at least 5 images and wrote a story about them. Then, another worker received the same images and wrote their story. For *descriptions-in-isolation* workers followed the instructions of image captioning tasks from MS COCO (Lin et al., 2014). For each album, 5 stories were collected, and for each image 3 workers wrote *descriptions-in-isolation*. Not to repeat images, we limit our data to only one annotation.

We expect that the first type of captions,

[*] https://huggingface.co/bert-base-cased

[†] https://github.com/airsplay/py-bottom-up-attention

[‡] According to the authors of the dataset, this layer of annotation has been lost.

Desc-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.
Desc-in-Sequence	A roof top with a black frisbee laying on the top of the edge of it.	A man is standing in the grass in front of the house kicking a soccer ball.	A man is in the front of the house throwing a soccer ball up.	A blue and white soccer ball and black Frisbee are on the edge of the roof top.	Two soccer balls and a Frisbee are sitting on top of the roof top.
Story-in-Sequence	**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.**

Figure 1: A story from the VIST dataset with three different layers of annotations: *descriptions-in-isolation*, *descriptions-in-sequence*, and *stories-in-sequence*.

descriptions-in-isolation, is more descriptive, in other words, annotators tend to mention objects on a picture rather than describe events. *Stories-in-sequence*, on the other hand, might not be directly related to pictures, and annotators can omit some information that could be extracted from visual context. Hence, two types of annotation reflect different types of image-text coherence.

5 Experimental setup

We run three experiments based on type of descriptions in VIST and whether descriptions were considered as a part of local or narrative discourse as shown in Table 1.

Captions	Local discourse	Narrative discourse
Descriptions-in-isolation	Experiment I	-
Stories-in-sequence	Experiment II	Experiment III

Table 1: Summary of experiments

In Experiment I and II we test the model on the image-caption matching task where items are selected in one of five different conditions according to which distractors are selected (Figure 2). In addition to a random assignment (condition 5) we also use similarity scores to control for different degree of distraction (conditions 1–4). To every item we match either a caption or an image with one of highly similar or dissimilar captions or images from the entire dataset. For each condition we create a separate dataset that consists of both matching and non-matching pairs. In Experiment III, we create distractors by selecting randomly an image or a caption from the same story. In every experiment, we test the model on each item (image or caption) *twice*: with its original caption or image and with its non-matching version and the task of the model is to identify a match or mismatch. In that way, we make our evaluation datasets balanced and this setup allows us to compare scores for match or mismatch across different setups.

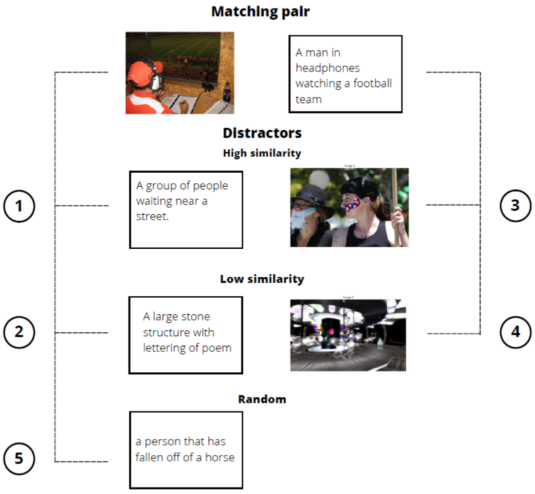
Matching pair
A man in headphones watching a football team
Distractors
High similarity
A group of people waiting near a street.
Low similarity
A large stone structure with lettering of poem
Random
a person that has fallen off of a horse
1
2
3
4
5

Figure 2: The dataset construction based on different similarities scores and modalities. The 5th condition repeats the procedure of assigning random captions in the pre-training objective of ViLBERT.

For constructing non-matching pairs based on similarity, we extract textual and visual representations of original descriptions or images and calculate pairwise scores between each dataset item of the same type with cosine similarity. We then take the upper quartile of the distribution of similarity scores as a threshold for high similarity items and the lower quartile as a threshold for low similar-

Distractors	High Similarity	Low similarity
Textual	0.92	0.95
Visual	0.94	0.95

Table 2: The overall performance on the image-caption matching task with *descriptions-in-isolation* based on accuracy on a balanced dataset. The accuracy of the baseline on random distractor descriptions is **0.94**.

ity items. For details see Appendix A. For each item, we construct lists of items with high and low similarity distractors from which we select items randomly. This way, we make our data more diverse as the same distractor is only used in one comparison task.

To extract representations for texts we use the Hugging Face implementation of BERT (Devlin et al., 2018) [§]. For image representations we use ResNet-101 (He et al., 2016), a deep residual network based on a CNN architecture. We choose these models since these are used by most language-and-vision transformers.

6 Results

6.1 Experiment I: Local discourse and descriptions in isolation

Here we test the model on images and their descriptions by distracting it with out-of-story captions permuted according to the five conditions outlined in Figure 2. Distractors are selected according to the modality (text or vision), the degree of similarity (high or low) or they are picked randomly.

The model shows high performance on *descriptions-in-isolation*: in all five conditions the accuracy is greater than 0.9 (Table 2).[¶] As for differences in similarity, the model performs better when distinguishing distractors of low similarity, possibly because such items are easier to differentiate. In terms of the modality, the difference in performance is observed only with high similarity distractors: the performance is slightly lower with textual distractors compared to visual distractors. This indicates the sensitivity of the model to changes in text where individual words carry high information content and, possibly, this result might also be indicative that the model is relying on text much more than on vision. This is a well-known challenge for multi-modal architectures – models are known to ground text in images better than images in text (Agrawal et al., 2018; Thomason et al., 2019; Ilinykh et al., 2022). In general, we cannot see a large difference in the results of the experiment under all five conditions.

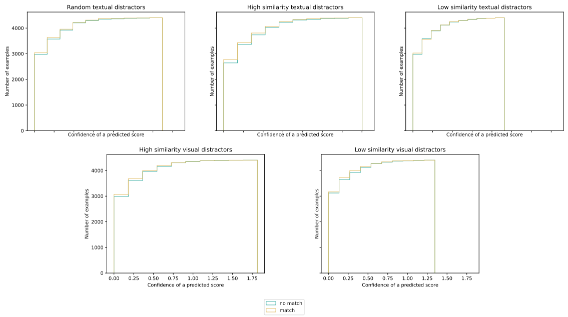

Figure 3: The confidence of the model in predicting a correct answer (match / no match) in an image-caption matching task for *descriptions-in-isolation*.

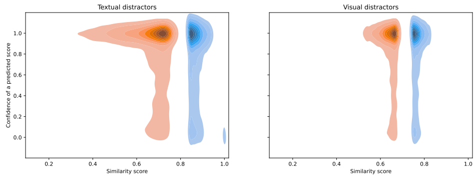

Figure 4: The relation between similarity scores and the model's confidence on correct answers for matching distractors for *descriptions-in-isolation*.

To understand the uncertainty of the model when predicting a match or non-match we look at its predictions on individual classes. If the model is confident in its predictions, predictions will be evenly distributed as the classes are balanced. Differences in performance on individual classes might reveal any differences in model uncertainty in cases involving high and low similarity distractors or between textual or visual distractors.

Figure 3 demonstrates that a predominant majority of the examples can be classified with high confidence. There is almost no difference between the different conditions except for random distractors where the model is less confident on non-matching labels than in the other four conditions. Figure 4 illustrates the relationship between similarity scores of distractors and confidence of the model. For a large number of examples the model is confident in identifying distractors of both high and low similarity. However, we can see that in the range of similarity scores where two classes meet the confidence drops on both classes with some examples. This is also the range where most of the examples

[§] https://huggingface.co/bert-base-uncased

[¶] Note that since both classes are balanced accuracy is a suitable measure here.

lie. When examples are more dissimilar the model retains high confidence.

Distractors	Condition	T-test
Similarity	captions: similar vs dis-similar	<0.001
	images: similar vs dis-similar	<0.001
Modality	similar: captions vs images	<0.001
	dis-similar: captions vs images	0.8

Table 3: The p-values of a *Student t-test* on model performance scores on *descriptions-in-isolation* using different distractors.

Finally, to demonstrate any differences between distractors in terms of degree of similarity or modality we run a *Student t-test* on prediction scores for non-matching examples. The results of the tests are summarised in Table 3 and indicate a significant difference in model's performance on different distractor types in three out of four cases. The t-test found no significant difference between dissimilar captions vs images. In the remaining cases, distractors of different types have a different effect on the model's performance.

6.2 Experiment II: Local discourse and stories in sequence

Distractors	High Similarity	Low similarity
Textual	0.78	0.85
Visual	0.81	0.85

Table 4: The overall performance on the image-caption matching task with *stories-in-sequence* based on accuracy on accuracy on a balanced dataset. The results on random distractor descriptions are **0.82**.

We implement the same experimental setup for *stories-in-sequence* to test the model's sensitivity to local discourse structure on captions that were originally part of a narrative. As seen in Table 4, the results are worse than on *descriptions-in-isolation*. The model performs slightly better on dissimilar distractors than on similar distractors. Modality matters for similar distractors, images are more easily identifiable than captions.

In Figure 5 we take a closer look at the model's confidence on predicting matches or mismatches. Overall, the model is quite certain in its predictions. The majority of cases are identified with high confidence scores. The model is more uncertain in matching distractors than true examples. Note however, that the confidence on true examples is less distributed with random distractors compared to distractors created with our similarity scores which indicates that the task we created is more difficult. Figure 6 shows the relation between similarity scores and the model confidence for matching distractors. The trends are similar to the ones reported for *descriptions-in-isolation*.

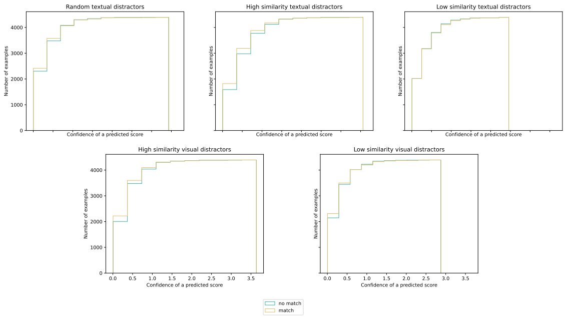

Figure 5: The confidence of the model in predicting a correct answer (match /no match) in an image-caption matching task for *stories-in-sequence*.

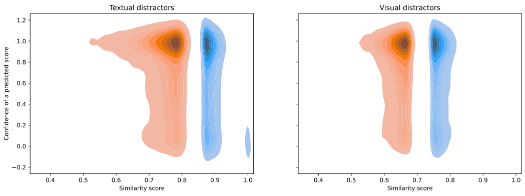

Figure 6: The relation between similarity scores and the model's confidence on correct answers for matching distractors for *stories-in-sequence*.

Table 5 shows a Student t-test of the model's performance identifying distractors of different kinds. The results are identical to the ones obtained for *descriptions-in-isolation* condition.

Distractors	Condition	T-test
Similarity	captions: similar vs dis-similar	<0.001
	images: similar vs dis-similar	<0.001
Modality	similar: captions vs images	<0.001
	dis-similar: captions vs images	0.08

Table 5: The p-values of a *Student t-test* on model performance scores on *stories-in-sequence* using different distractors.

6.3 Experiment III: narrative discourse and stories-in-sequence

The last task includes image-caption matching of sentences from the *stories-in-sequence* annotation layer where distractors are randomly sampled from the same story. The results can be seen in Table 6. Unlike in the previous experiments the model's performance is higher on textual distractors than on visual distractors. A possible reason might be that images within one story are more similar than cap-

Distractors	Random
Textual	0.63
Visual	0.6

Table 6: The overall performance on the image-caption matching task with *stories-in-sequence*. Distractors are randomly sampled from the same story.

tions but also our earlier observation that images are less informative for the model than texts. Another reason might be the way *stories-in-sequence* are collected. Some workers were not the ones who combined pictures in stories and were only asked to write a story about a pre-chosen sequence of images. Different crowd-workers would have different perspectives and understanding of sequences of images, resulting that in some cases more descriptions are focused on creating coherent storyline rather than referring to images.

In comparison to Experiment II where *stories-in-sequence* were used but were taken from different stories, the performance here drops by 0.2 which indicates that the model struggles identifying distractors from the same story. As the same entities appear across the sequence of a story the model struggles to capture causal relations that could identify items from different parts of a story. Moreover, as seen in Figure 7, the model finds it harder to identify distractors than true labels: more than 1,500 distractors out of 2,500 are identified as true labels. Furthermore, the model does not only predict wrong answers but it gives high confidence scores to false positives (low confidence in the distractor label means high confidence in the true label). Overall, we can conclude that the model is struggling in identifying distractors from the narrative discourse.

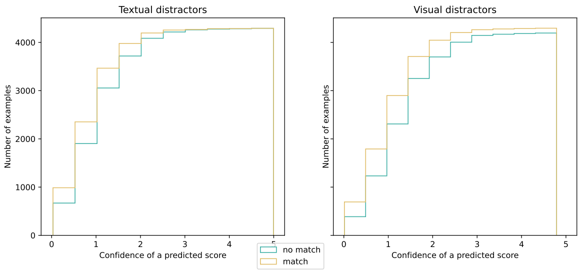

Figure 7: The confidence of the model in predicting a correct answer (match / no match) in an image-caption matching task for *stories-in-sequence*.

7 Discussion

7.1 The impact of non-matching discourse

Overall, for *local discourse* the evaluation results show that the model was not particularly challenged by distractors with high similarity, especially for *descriptions-in-isolation*. In other words, the model distinguishes different situations well regardless whether descriptions or images of the situations are close in similarity or not. However, for *stories-in-sequence*, the performance of the model on distractors of high similarity was slightly lower than those with distractors of low similarity. Table 5 confirms that this drop in performance is statistically significant. Here, the model can distinguish between two completely different situations but is less sensitive in identifying similar entities (whether textual or visual) in different situations.

Descriptions-in-isolation contain captions that are more grounded in images as they were not created as a part of narrative. For example, they refer to entities and their attributes and how they are related in images. On the other hand, *stories-in-sequence* are more abstract in the sense that they are more grounded in the narrative of the story. In such captions, annotators rely on visual attention of an interpreter to match the story with the images but such information which is part of human cognitive processing ability is not explicitly expressed in the data. Crowd-workers had different communication intents while writing these annotations. The model has thus less information to ground descriptions in images.

The differences between *descriptions-in-isolation* and *stories-in-sequence* on the local discourse task could be also due to the fact that the model was pre-trained on captions that were also produced in isolation although they were from a different dataset (Conceptual Captions rather than VIST).

7.2 The role of modalities

It has been shown that models rely more on the textual modality in its predictions than on the visual modality (Frank et al., 2021). Our results from Experiments I and II reveal the same tendency. The results on experiments where image distractors were tested are better on *descriptions-in-isolation* and *stories-in-sequence*. It is more difficult for the model to identify a mismatch between an image and a distractor caption than a mismatch between a caption and a distractor image. The distribution of similarity scores indicates that images are less diverse than captions. A correct caption will be more informative for the model to find sufficient information in the image to make a decision.

In Experiment III, however, the model performs better on textual distractors meaning that having a correct image is more informative to identify a mismatch of textual distractors. This means that here captions are more similar to each other which could be explained by the fact that they create a story narrative whereas images are different snapshots of the situations involved.

7.3 Are transformers sensitive to discourse?

The model shows better performance on captions taken from different stories than on ones taken from the same story which means that the model is less sensitive to distinctions made within a narrative. The model is good at capturing even fine-grained differences but the performance degrades when it is presented with more abstract captions that are grounded more in the narrative than in an image. It could be related to different types of image-text coherence (Alikhani et al., 2023) since in *stories-in-sequence* information that overlaps with visual context could be omitted. It makes the image-caption matching task more complicated for the model and reflects that the model struggles with complex coherence relations. However, taking into account the biases of the dataset, such as some sequences could have stronger visual discourse and otherwise, it is a complicated issue whether the model is not sensitive to visual, textual, or situation-level discourse.

However, the performance of the model on distractors sampled from the same story is still better than a random baseline which for a balanced binary classification task is 0.5. Looking at the example in Figure 1 we see that the story mentions the same objects, but the sentences in the story are not interchangeable so there is some discourse information there that the model can utilise. In other words, the model can differentiate different situations but it struggles with identifying fine-grained pragmatic distinctions through form alone when all items come from the same story.

8 Conclusion

In this work we examined whether a language-and-vision model is sensitive to discourse structure at different levels of granularity of distractors, random, similar and dissimilar to control for the difficulty of the task. We focused on local and narrative discourse in the task of image-caption matching which is one of the pre-training objectives of multimodal models. We ran three experiments on ViLBERT under different conditions using data from the Visual Storytelling dataset where images are collected into stories. The dataset has several layers of annotation including captions of images in isolation and captions of images in stories in sequence.

In Experiment I we test ViLBERT's image-caption matching performance on the captions of images in isolation under five different conditions with different distractors, random, similar and dissimilar. We observe that the model performs better than 90% in all cases.

In Experiment II we take the *stories-in-sequence* annotation level and test them for local discourse matching against distractors from different stories. The results drop by 0.1 in comparison with the previous experiment which shows that captions that are part of a narrative are harder to distinguish from captions from different narratives. However, the model still achieves high performance and has therefore learnt to distinguish the differences in local discourse structure by identifying from other local discourses.

In Experiment III we focus on *stories-in-sequence* discourses where distracting items are sampled from the same story. The model's performance drops to 0.6. The model assigns false positives to the matching class with high confidence and therefore finds it challenging to distinguish items from the same story discourse that refer to the same entities but over a sequence of relations. The model is more successful in distinguishing between different entities and situations as shown by Experiment I and II.

The effect of dis/similarity of distractors becomes more pronounced when the nature of discourse becomes more challenging. Overall, such behaviour of the model is expected since it has never been trained on captions that come from stories in sequence. While this might be achieved through fine-tuning which we will attempt in our future work it also raises a more profound question about the nature of semantic knowledge captured in large models. On the one hand, the models are exposed to contextual discourse knowledge while they are trained since *all* descriptions are made within some context. On the other hand, their performance quickly degrades when the nature of the context changes. Understanding the discourse context is therefore important both at the level of data creation and at the level of utilisation of pre-trained models. It is important to note that our results should be considered relative to the dataset and the model we investigated and variations are expected based on the contexts in which the model was trained and the contexts it is applied to. Assessing the similarities between discourses (even by expert linguists) may not be straightforward as parameters may not be directly observable or known.

In our future work, we are planning to expand our methods to other language-and-vision models, such as single-stream models, and other narrative discourse datasets, such as RecipeQA, which is an example of naturally created narrative in comparison with crowd-sourced VIST dataset.

Acknowledgments

Ilinykh and Dobnik were supported by a grant from the Swedish Research Council (VR project 2014-39) for the establishment of the Centre for Linguistic Theory and Studies in Probability (CLASP) at the University of Gothenburg. All authors would like to thank reviewers for their comments.

References

Aishwarya Agrawal, Dhruv Batra, Devi Parikh, and Aniruddha Kembhavi. 2018. Don't just assume; look and answer: Overcoming priors for visual question answering. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 4971–4980. Computer Vision Foundation / IEEE Computer Society.

Malihe Alikhani, Baber Khalid, and Matthew Stone. 2023. Image–text coherence and its implications for multimodal ai. *Frontiers in Artificial Intelligence*, 6.

Malihe Alikhani, Piyush Sharma, Shengjie Li, Radu Soricut, and Matthew Stone. 2020. Clue: Cross-modal coherence modeling for caption generation. *arXiv preprint arXiv:2005.00908*.

Malihe Alikhani and Matthew Stone. 2019. "caption" as a coherence relation: Evidence and implications. In *Proceedings of the Second Workshop on Shortcomings in Vision and Language*, pages 58–67, Minneapolis, Minnesota. Association for Computational Linguistics.

Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-up and top-down attention for image captioning and visual question answering. In *CVPR*.

Vladimir Araujo, Andrés Villa, Marcelo Mendoza, Marie-Francine Moens, and Alvaro Soto. 2021. Augmenting BERT-style models with predictive coding to improve discourse-level representations. *arXiv preprint arXiv:2109.04602*.

Yonatan Belinkov. 2022. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219.

Raffaella Bernardi and Sandro Pezzelle. 2021. Linguistic issues behind visual question answering. *Language and Linguistics Compass*, 15(6):elnc3–12417.

Emanuele Bugliarello, Ryan Cotterell, Naoaki Okazaki, and Desmond Elliott. 2021. Multimodal pretraining unmasked: A meta-analysis and a unified framework of vision-and-language BERTs. *Transactions of the Association for Computational Linguistics*, 9:978–994.

Mingda Chen, Zewei Chu, and Kevin Gimpel. 2019. Evaluation benchmarks and learning criteria for discourse-aware sentence representations. *arXiv preprint arXiv:1909.00142*.

Alexis Conneau, German Kruszewski, Guillaume Lample, Loïc Barrault, and Marco Baroni. 2018. What you can cram into a single vector: Probing sentence embeddings for linguistic properties. *arXiv preprint arXiv:1805.01070*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Yanai Elazar, Shauli Ravfogel, Alon Jacovi, and Yoav Goldberg. 2021. Amnesic probing: Behavioral explanation with amnesic counterfactuals. *Transactions of the Association for Computational Linguistics*, 9:160–175.

Allyson Ettinger. 2020. What BERT is not: Lessons from a new suite of psycholinguistic diagnostics for language models. *Transactions of the Association for Computational Linguistics*, 8:34–48.

Stella Frank, Emanuele Bugliarello, and Desmond Elliott. 2021. Vision-and-language or vision-for-language? on cross-modal influence in multimodal transformers. *arXiv preprint arXiv:2109.04448*.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Xudong Hong, Asad Sayeed, Khushboo Mehra, Vera Demberg, and Bernt Schiele. 2023a. Visual Writing Prompts: Character-Grounded Story Generation with Curated Image Sequences. *Transactions of the Association for Computational Linguistics*, 11:565–581.

Xudong Hong, Asad Sayeed, Khushboo Mehra, Vera Demberg, and Bernt Schiele. 2023b. Visual writing prompts: Character-grounded story generation with curated image sequences. *Transactions of the Association for Computational Linguistics*, 11:565–581.

Xudong Hong, Rakshith Shetty, Asad Sayeed, Khushboo Mehra, Vera Demberg, and Bernt Schiele. 2020. Diverse and relevant visual storytelling with scene graph embeddings. In *Proceedings of the 24th Conference on Computational Natural Language Learning (CoNLL 2020)*, pages 420–430, Online. ACL.

Ting-Hao Huang, Francis Ferraro, Nasrin Mostafazadeh, Ishan Misra, Aishwarya Agrawal, Jacob Devlin, Ross Girshick, Xiaodong He, Pushmeet Kohli, Dhruv Batra, et al. 2016. Visual storytelling. In *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: Human language technologies*, pages 1233–1239.

Nikolai Ilinykh and Simon Dobnik. 2022a. Attention as grounding: Exploring textual and cross-modal attention on entities and relations in language-and-vision transformer. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 4062–4073.

Nikolai Ilinykh and Simon Dobnik. 2022b. Do Decoding Algorithms Capture Discourse Structure in Multi-Modal Tasks? A Case Study of Image Paragraph Generation. In *Proceedings of the 2nd Workshop on Natural Language Generation, Evaluation, and Metrics (GEM)*, pages 480–493, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Nikolai Ilinykh, Yasmeen Emampoor, and Simon Dobnik. 2022. Look and answer the question: On the role of vision in embodied question answering. In *Proceedings of the 15th International Conference on Natural Language Generation*, pages 236–245.

Karim Lasri, Tiago Pimentel, Alessandro Lenci, Thierry Poibeau, and Ryan Cotterell. 2022. Probing for the usage of grammatical number. *arXiv preprint arXiv:2204.08831*.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, et al. 2022. Holistic evaluation of language models. *arXiv preprint arXiv:2211.09110*.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft COCO: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer.

Nelson F Liu, Matt Gardner, Yonatan Belinkov, Matthew E Peters, and Noah A Smith. 2019. Linguistic knowledge and transferability of contextual representations. *arXiv preprint arXiv:1903.08855*.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. ViLBERT: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, 32.

Rowan Hall Maudslay and Ryan Cotterell. 2021. Do syntactic probes probe syntax? experiments with jabberwocky probing. *arXiv preprint arXiv:2106.02559*.

Allen Nie, Erin Bennett, and Noah Goodman. 2019. DisSent: Learning sentence representations from explicit discourse relations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4497–4510.

Letitia Parcalabescu, Michele Cafagna, Lilitta Muradjan, Anette Frank, Iacer Calixto, and Albert Gatt. 2021. Valse: A task-independent benchmark for vision and language models centered on linguistic phenomena. *arXiv preprint arXiv:2112.07566*.

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind K Joshi, and Bonnie L Webber. 2008. The Penn Discourse TreeBank 2.0. In *LREC*.

Laura Ruis, Akbir Khan, Stella Biderman, Sara Hooker, Tim Rocktäschel, and Edward Grefenstette. 2022. Large language models are not zero-shot communicators. *arXiv*, arXiv:2210.14986 [cs.CL]:1–45.

Naomi Saphra. 2021. *Training dynamics of neural language models*. Doctoral thesis, School of Informatics, University of Edinburgh, Edinburg, UK.

Karolina Stańczak, Edoardo Ponti, Lucas Torroba Hennigen, Ryan Cotterell, and Isabelle Augenstein. 2022. Same neurons, different languages: Probing morphosyntax in multilingual pre-trained models. *arXiv preprint arXiv:2205.02023*.

Ece Takmaz, Mario Giulianelli, Sandro Pezzelle, Arabella Sinclair, and Raquel Fernández. 2020. Refer, reuse, reduce: Generating subsequent references in visual and conversational contexts. *arXiv preprint arXiv:2011.04554*.

Ekaterina Taktasheva, Vladislav Mikhailov, and Ekaterina Artemova. 2021. Shaking syntactic trees on the sesame street: Multilingual probing with controllable perturbations. *arXiv preprint arXiv:2109.14017*.

Hao Tan and Mohit Bansal. 2019. LXMERT: Learning cross-modality encoder representations from transformers. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5100–5111, Hong Kong, China. Association for Computational Linguistics.

Jesse Thomason, Daniel Gordon, and Yonatan Bisk. 2019. Shifting the baseline: Single modality performance on visual navigation & QA. 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)*, pages 1977–1983, Minneapolis, Minnesota. Association for Computational Linguistics.

Alex Warstadt, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R Bowman. 2020. BLiMP: The benchmark of linguistic minimal pairs for english. *Transactions of the Association for Computational Linguistics*, 8:377–392.

Semih Yagcioglu, Aykut Erdem, Erkut Erdem, and Nazli Ikizler-Cinbis. 2018. Recipeqa: A challenge dataset for multimodal comprehension of cooking recipes. *arXiv preprint arXiv:1809.00812*.

A Study of similarity scores in Conceptual Captions and Visual Storytelling Datasets

We use similarity scores for constructing our tasks for the model. However, the dataset can be biased and this will affect the similarity scores needed for construction of tasks for the first two experiments. As seen in Figure 8, the similarity scores between images or texts vary between 0.6 and 0.9. Since these scores are not interpretable, i.e. we cannot compare their relative differences across modalities, we compare their distribution with the distribution in the Conceptual Captions dataset that language-and-vision models are trained on. Figure 8 shows that the distribution of similarities of captions is similar in the two datasets, although image similarities have a slightly narrower distribution in the VIST dataset. In other words, they are less diverse. This may lead to the model struggling to distinguish similar images, as they are more similar than in the pre-training data.

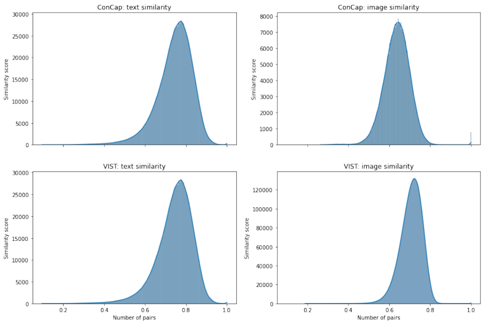

Figure 8: The distribution of similarity scores for captions and images for the Conceptual Captions (ConCap) and the Visual Storytelling dataset (VIST).