Encoded Object Affordances in Text-Only and Multi-Modal Language Models

Klara Båstedt
Department of Philosophy, Linguistics and Theory of Science, University of Gothenburg
gusbaskl@student.gu.se

Abstract

This paper examines what knowledge of objects and their affordances is encoded in a large language model trained on text only and compare it with a multi-modal language model that benefit from vision. The question is whether a multi-modal model has more suitable word representations for assigning affordances to unseen objects since it has been trained on images and text jointly. To examine this, a simple probe is trained on the models' representations of objects and affordances and tested on unseen objects. The experiment shows that the probe performs similarly on the two models. However, the performance of the probe is too poor to draw conclusions about a multimodal model's potential advantage over a text-only model when it comes to hypothesizing about object affordances.

1 Introduction

Humans can use visual information about the traits of an object, such as material, shape, and size, to hypothesize about its affordances. In Zhu et al. (2014), this ability is modelled with a rule-based knowledge base. Their model can successfully assign affordance labels to unseen objects in images, based partly on their observable features and the knowledge base.

Since visual features are of importance for predicting the affordances of objects, it is plausible to assume that models that are trained on images and text jointly have this knowledge encoded to a larger extent than text-only models. Along these lines, Ilharco et al. (2020) use a probe to examine whether multimodal models have an advantage over unimodal models when it comes to mapping descriptions to corresponding images. Their investigation shows that while multimodal models benefit from having seen images during training, the representations from the text-only model are also suitable for identifying matching images. Even though the visually grounded models outperform contextual language models, they are still far away from human performance.

With this background, this experiment aims to compare the effectiveness of text-only and multi-modal word representations for assigning affordances to objects. To achieve this, a simple probe is trained on word embeddings from unimodal BERT (Devlin et al., 2019) and multimodal VisualBERT (Li et al., 2019). The task of the probe is to correctly map representations of objects with representations of affordances. By comparing the performance of the probes, we can see to what extent knowledge about object affordances is encoded in the representations of the two models and more specifically if having trained on images result in more suitable representations in this regard.

2 Materials and methods

The data used in this experiment consists of names of the 62 objects from Zhu et al. (2014), annotated with 15 affordances. In Zhu et al.'s experiment (2014), the objects are divided into train and test set considering their affordances. E.g., 'guitar' is part of their trainset and 'banjo', with identical affordances, appear in the test set so that their hypothesis can be tested. In this experiment however, the division into train, validation and test set is made randomly. This resulted in three object pairs with identical affordances with one of the objects in the training set and the other in the test set (see 3.3).

The objects and affordances are tokenized and given to the two language models. The penultimate hidden layer is used to represent them, and the representations are multiplied and assigned a truth value from the annotations. The multiplied representations are passed to the probe which consists of a linear layer and a Sigmoid function.

As in Ilharco et al. (2020), the probe has a simple design since its purpose is to investigate the usefulness of the model representations for mapping objects to their affordances. In other word, the goal is not excellent performance of the probe but rather to examine if there are advantages in the representations of any of the models.

3 Results

3.1 Evaluation metrics

The probe performs similarly on both models with comparable accuracy and F1-score. The probe trained on BERT representations obtain a higher score for recall at the expense of precision while it is the opposite for the VisualBERT probe.

	BERT	VisualBERT
Accuracy	86,67%	87,33%
Precision	76,36%	84,09%
Recall	85,71%	75,51%
F1	80,77%	79,57%

Table 1. Evaluation metrics for the BERT and VisualBERT probes.

The accuracy of the probe is higher than the baseline of 72.26% which would be obtained by always predicting 0. However, the affordances are not equally common. 90.23% of the objects in the total dataset has the affordance 'push' while ‘row’ only applies to 3.23% of the objects (see 3.2). Considering this, an accuracy of 83,23% would be obtained by always guessing 1 on the common affordances and 0 on the rare ones. The probe only performs slightly better than this baseline.

The imbalanced and small dataset, which consists of only 930 object and affordance pairs, causes the probe to struggle with convergence. The difficulty of assigning truth values to products of word representations with such a simple probe is another possible explanation.

Figure 1 and 2 shows the plotted curves of training and evaluation accuracy for the BERT and VisualBERT probes. The diagrams visualize the difficulty for the model to converge as accuracy jumps up and down between epochs. The figures also show that the models start to overfit to the training data after around 2000 epochs. Therefore, the model with the highest validation accuracy before reaching 2000 epochs is saved.

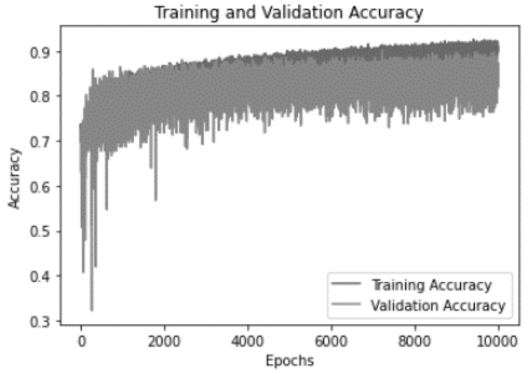

Figure 1. Training and validation accuracy for BERT probe.

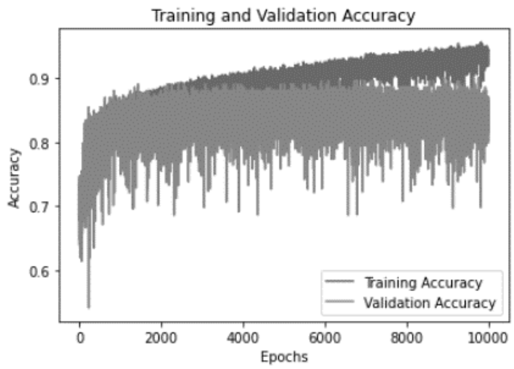

Figure 2. Training and validation accuracy for VisualBERT probe.

Figure 3 shows the evaluation metrics for the probes trained using ten different manual seed. As visualized in the diagram, the scores differ remarkably between runs. The results and analysis presented in this report are based on the best performing probe among these ten examples.

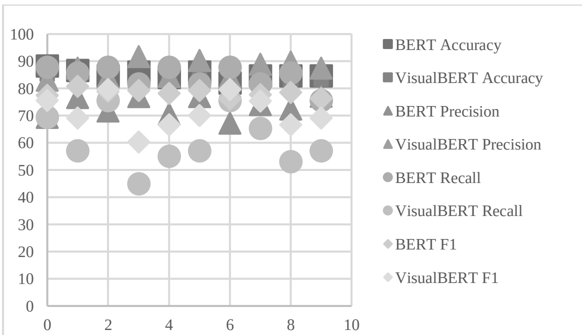

Figure 3. Evaluation metrics for BERT and VisualBERT probes using different manual seed.

3.2 Accuracy per object and per affordance

While the evaluation metrics above imply a better performance than the baseline, this is not the case when considering the per-affordance accuracy since the affordances are not equally common. Table 2 below shows the per-affordance accuracy of the two probes on the test set. The baseline column displays the percentage of the objects that do or do not have that affordance in the total dataset while baseline test set shows the percentage based on the objects in the test set.

	BERT	VisualBERT	Baseline	Baseline testset
grasp	90.0 %	80.0 %	59.68 %	90.0 %
lift	90.0 %	80.0 %	82.26 %	90.0 %
throw	70.0 %	70.0 %	50.0 %	80.0 %
push	100.0 %	100.0 %	90.32 %	100.0 %
fix	70.0 %	50.0 %	59.68 %	60.0 %
ride	80.0 %	90.0 %	80.65 %	90.0 %
play	80.0 %	80.0 %	95.16 %	80.0 %
watch	60.0 %	90.0 %	93.55 %	90.0 %
sit on	80.0 %	90.0 %	74.19 %	90.0 %
feed	100.0 %	100.0 %	90.32 %	100.0 %
row	90.0 %	90.0 %	96.77 %	90.0 %
pour from	100.0 %	100.0 %	90.32 %	100.0 %
look through	100.0 %	100.0 %	95.16 %	100.0 %
write with	100.0 %	100.0 %	95.16 %	100.0 %
type on	90.0 %	90.0 %	95.16 %	90.0 %

Table 2. Per-affordance accuracy compared with two baselines.

An accuracy of 90% for the affordance 'row' might seem good, but since the baseline of 'row' calculated on the objects in the test set is also 90% it means that the model fails to assign 'row' to the only object in the test set with that affordance. Only for the affordance 'fix', the BERT probe performs better than the test set baseline. VisualBERT never performs better than the test set baseline.

	BERT	VisualBERT	Baseline
carving knife	86.67 %	93.33 %	73.33 %
dustcloth	93.33 %	93.33 %	73.33 %
guitar	80.0 %	93.33 %	60.0 %
handset	100.0 %	100.0 %	66.67 %
laptop	80.0 %	86.67 %	53.33 %
power saw	93.33 %	93.33 %	73.33 %
violin	93.33 %	86.67 %	60.0 %
bowl	100.0 %	93.33 %	73.33 %
kayak	53.33 %	60.0 %	73.33 %
walkie-talkie	86.67 %	73.33 %	66.67 %

Table 3. Per-object accuracy compared with baseline.

The per-object accuracy presented in Table 3 shows that both probes perform better than the baseline. However, this baseline does not consider that some affordances are more probable than other given their distribution in the dataset.

3.3 Seen and unseen objects with identical affordances

There are three pairs of objects with identical affordances where one is seen during training and the other during testing. These pairs are 'guitar' and 'banjo', 'carving knife' and 'sickle', and 'kayak' and 'small boat'. They are particularly interesting since they have the potential to tell whether the models' representations of the objects have similarities that facilitate the task of assigning them similar affordances and if this differs between the multi-modal and text-only model.

The predictions of the VisualBERT probe for 'kayak' and 'small boat' are presented in Table 4. The model incorrectly assigns the most common affordances 'grasp' and 'lift' to the objects and fails to assign the less common affordances 'ride', 'sit on' and 'row'. This is the case for both the seen and the unseen object.

	affordance	small boat	kayak	target
0	grasp	1	1	0
1	lift	1	1	0
2	throw	1	0	0
3	push	1	1	1
4	fix	1	1	0
5	ride	0	0	1
6	play	0	0	0
7	watch	0	0	0
8	sit on	0	0	1
9	feed	0	0	0
10	row	0	0	1
11	pour from	0	0	0
12	look through	0	0	0
13	write with	0	0	0
14	type on	0	0	0

Table 4. Predictions of the VisualBERT probe on seen and unseen objects with identical affordances.

4 Discussion

Despite the small amount of data and the simple architecture of the probe, the model learns to map representations of objects with representations of affordances. Nevertheless, the probe achieves an accuracy only slightly higher than what it would get by always assigning the common affordances and never the uncommon ones to each object. It is reasonable to assume that this is because of the limited training data.

A way to overcome this difficulty is to use images of objects instead of word representations and map them to representations of affordances. By using e.g., 100 images of each object instead of just one representation, the dataset would expand from 930 to 93000 pairs of objects and affordances. This would allow the model to see combinations of objects and affordances more than once.

Another advantage of training on images of objects instead of their representations is that the visual features of the objects are made explicit to the probe. This might help in the task of assigning affordance labels, especially to novel objects.

While the probing method has potential for examining the knowledge of object affordances encoded in representations from large models, the architecture of the probe in this experiment might be too limited. Even though the idea is to use a simple probe, this needs to be balanced with the difficulty of the task. Perhaps a linear layer is not enough to generalize about products of embeddings, and it might be more appropriate to use an LSTM, as in Ilharco et al. (2020).

5 Conclusions and further work

While multi-modal VisualBERT has no advantage over BERT in this experiment, no conclusions about the encoded knowledge in text-only and multi-modal models can be drawn due to the poor performance of the probe. The results of this experiment are not enough to make assumptions about the ability of these models to hypothesize about object affordances and whether it helps to see images during training in this regard.

It is still an interesting question whether the knowledge encoded in multi-modal and text-only models are different in terms of object affordances, and whether the results are in line with the findings of Ilharco et al. (2020). Future work consists of expanding the dataset with images of objects and training an LSTM probe to map them with representations of affordances.

References

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. 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.* Association for Computational Linguistics.

Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2019. Visualbert: A simple and performant baseline for vision and language. arXiv:1908.03557

Ilharco, Gabriel & Zellers, Rowan & Farhadi, Ali & Hajishirzi, Hannaneh. 2020. Probing Contextual Language Models for Common Ground with Visual Representations.

Yuke Zhu, Alireza Fathi, and Li Fei-Fei. 2014. Reasoning about Object Affordances in a Knowledge Base Representation. In: *Fleet D., Pajdla T., Schiele B., Tuytelaars T. (eds) Computer Vision – ECCV 2014. ECCV 2014. Lecture Notes in Computer Science, vol 8690.* Springer, Cham. https://doi.org/10.1007/978-3-319-10605-2_27