Investigating Direct and Indirect Grounding in Concrete and Abstract Concepts

Victoria Filippatou
LT2318 course Cognitive AI Systems
University of Gothenburg

Abstract

This small scope study aims to investigate the role of direct and indirect grounding in predicting concrete and abstract concepts respectively, with a focus on nouns, but also the way different language models perform without being specifically fine-tuned for such a task. Assuming that textual and visual modalities offer different kinds of possibly uncomplementary information, two BERT language models were tested for predicting concrete and abstract nouns based only on textual information and on textual + visual information.

1 Introduction

Grounding cognition in perception has evolved over the years and has been the topic of focus for several researchers in cognitive science, psycholinguistics, psychology, and Natural Language Processing (NLP) for a very long time. Cognitive linguists like (Fauconnier, 1985) and (Langacker, 1987) have proposed a number of cognitive theories about mental spaces and cognitive grammars that are grounded in human experience, which relates to how humans perceive things, language, and thoughts. Moreover, Lakoff and Johnson have proposed a metaphorical way in which humans organise their thoughts which is grounded in embodied experience. For example, humans can understand orientational metaphors like *"I'm feeling up today"* because they have a body (Lakoff and Johnson, 1980). Of course, cultural experience plays a role in making associations between up and happy and down and sad.

Given what is discussed above, how simple is it to make computers understand and perceive language the way humans do? Meaning to make the right associations between words and the real world, understand multiword expressions (MWEs) as having a single and not necessarily compositional meaning, and ultimately abstract concepts, i.e., values, emotions, etc.

When it comes to symbolic AI systems Searle (1980) has posed the symbol grounding problem as a problem of intrinsic meaning and gives the example that if a computer could respond appropriately with Chinese symbols to a Chinese input then the computer would be able to understand these symbols in the same way a native English speaker understands the meaning of English words. Basically, there are no concrete insights into how exactly amodal symbols are mapped to perception and entities in the actual world.

(Harnad, 1990) has addressed this problem and proposes an indirect grounding of words and concepts from already grounded ones. He gives the example of the word "*zebra*" being grounded in the categories that constitute its meaning, "*horse*" and "*stripes*". He also argues that if someone had never seen a zebra before, they could potentially identify it if they could identify horses and stripes. This small scope study will venture to explore indirect grounding, which suggests that abstract concepts are also grounded in physical or sensorimotor experiences, but not directly like concrete concepts. Rather, they are learned by the embodied representations activated by language which exist independently of it (Dove, 2014; Günther et al., 2020). Starting from the hypothesis that abstract words are related to concrete words and can get grounded by them, it will try to answer the following questions:

1. Can indirect grounding of abstract concepts to visual images of the concrete concepts they semantically relate to help in their prediction by language models?

2. How do different modality BERT models (textual and visual) perform in predicting them without being specifically fine-tuned for

this task?

The next section describes methods and studies related to this topic. Afterwards, materials and methods used in this project are going to be described, their results will be analysed and discussed, and finally conclusions and future work will be proposed.

2 Related Work

Grounding words into images to distinguish between abstract and concrete concepts is not a new approach to computational semantics. (Bhaskar et al., 2017) explored the performance of various multi-modal models in distinguishing between concrete and abstract nouns. More specifically, they compared count vectors and embeddings from the *word2vec* model and extracted visual features from Google images. They then proceeded to concatenate the vectors acquired and classified them using a binary classifier and a regression model. Their quantitative analysis showed that the multimodal classification only marginally outperformed the single modality ones and the qualitative analysis revealed that concrete words are classified better when relying only on textual features but also when relying on visual rather than textual features. However, they conclude that all variants of the modalities used can similarly well represent abstract and concrete nouns and that both text and images seem to provide sufficient and non complementary information in representing them.

Furthermore, research has also been done in identifying abstract metaphorical concepts by grounding them to visual information. (Shutova et al., 2016) have pioneered by proposing a model that is able to identify metaphors by drawing knowledge from linguistic and visual data. The authors acquired textual and visual embeddings and experimented with different methods in order to construct the textual representation of a word or phrase, as well as different fusion methods to combine linguistic and visual information. They found that extracting visual information is a suitable way to capture concreteness as well as relevant perceptual properties of concepts and their proposed multi-modal model outperformed language only models, suggesting that visual information plays an important role in metaphor identification.

(Kennington and Natouf, 2022) argue the importance and the ongoing presence of the symbol grounding problem and the inability of existing models to break from it. They reframe it as a problem of concreteness-abstractness and stress that distributional models handle abstractness better than concreteness, possibly due to the fact that the first is captured more efficiently by context information. They take inspiration from child development and they propose approaching modeling concrete and abstract words by training different logistic regression classifiers for each word using images. They also provide negative examples of each word and test the resulting binary classifiers on their ability to fit test images. For example, a trained classifier for the color red will probably return a higher probability if presented with an image where red is the dominant color. Abstract word classifiers are trained by grouping together corresponding concrete words for the abstract category, i.e., *chair* and *table* for the ***furniture*** category. They conclude by arguing that it could be possible to build meanings of words that are in the middle of the concrete-abstract spectrum in this way, and that more abstract words could be better learned distributionally, whereas concrete ones are learned directly from symbol grounding.

(Utsumi, 2022) also explore the role of indirect grounding of abstract concepts and they devise a model that incorporates this view into multimodal distributional semantics. The author follows an approach similar to (Kennington and Natouf, 2022). For concrete words, the visual vector is directly computed from images that include the tag of this word in the Flickr database, whereas for abstract words this vector is computed from the visual images of their semantic neighbors. Their model achieved great results and pointed to the fact that the indirect grounding hypothesis stands on some solid ground.

This paper leverages some of the methods discussed above trying to answer the questions stated in the previous section.

3 Materials and Methods

3.1 Dataset

The dataset created for this study was based on the concreteness-abstractness ratings by (Brysbaert et al., 2014). This dataset contains ratings for 25.000 English words, each of which was rated on a scale from 1 (very abstract) to 5 (very concrete) by 25 participants. In order to make this study feasible for a short course project, have a balanced dataset, and given that the words in the dataset were not tagged for POS, only 100 random nouns were extracted using Spacy and grouped into 3 categories: abstract (1-2), middle (2-4), concrete (4-5). The ratings in each category were converted into 3 distinct labels for a classification task.

After that, for each concrete word 12 images were retrieved using the API flickr.photos.search with the relevance sort argument. In order to ground middle and abstract words, hypothesising that they are not directly grounded to visual information, it was first required to find their most relevant concrete semantic neighbors. In order to do that, all the words were encoded by a sentence transformer model and in order to only find concrete words relating to them, all concrete words were extracted from the (Brysbaert et al., 2014) dataset. The semantic neighbors were found in there using the Faiss library (Johnson et al., 2019) with the Euclidean distance metric. 3 semantic neighbors were found for each noun in the middle and abstract categories and 4 images were retrieved for each semantic neighbor using the same API configuration.

However, it is important to mention that not all words have the same amount of corresponding images, since the API did not retrieve the specified amount for all of them. Moreover, some images do not really relate to the search text even if the relevance sorting method was used. Hence, the dataset is balanced for number of instances in each category but not for corresponding images.

3.2 Models

The first model tested relies only on textual information and is a distilbert-base-uncased model (Sanh et al., 2019). The model is used as is, with no special fine-tuning other than a linear classification layer that is added on top of it. The last hidden state of the model's output representations is fed to the linear classifier to make predictions.

For testing both textual and visual modalities Visual BERT was used (Li et al., 2019). Just like the previous model described, a linear classification layer is added on top of the model. Visual BERT needs both textual and visual input and the visual embeddings were acquired using the ResNet50 model. Visual BERT concatenates the textual and visual embeddings and the last hidden layer of this output is passed to the classifier for classification.

3.3 Evaluation

Both of the models described above were evaluated on the performance metrics of accuracy, precision, recall, and F1 score for the classification task. Based on the literature discussed before, an expected result could be that the concrete class might achieve higher overall scores. However, it is important to keep in mind that the models explored here, are not fine-tuned for the specific task, so the results are expected to be poor nonetheless.

4 Results

4.1 Textual BERT Model

After running the model several times, it is noticed that its predictions are random and inconsistent.

Class	Precision	Recall	F1 Score
concrete	0.00	0.00	0.00
middle	0.00	0.00	0.00
abstract	0.33	1.00	0.50

Table 1: First run of Textual BERT Model.

Class	Precision	Recall	F1 Score
concrete	0.42	0.34	0.38
middle	0.00	0.00	0.00
abstract	0.36	0.80	0.50

Table 2: Another run of Textual BERT Model.

As can be observed in Table 1 precision for the abstract class is 0.33, recall is 1.00, and f1 score is 0.50, whereas the other classes have 0.00 scores for these metrics. Since the dataset is balanced for category instances, these results indicate that the model is predicting only one class. In Table 2 the predictions of the model are divided between the concrete and the abstract class with no results for the middle category.

4.2 Visual BERT

For Visual BERT the results run along the same lines as the textual model.

Class	Precision	Recall	F1 Score
concrete	0.00	0.00	0.00
middle	0.33	1.00	0.50
abstract	0.00	0.00	0.00

Table 3: First run of Visual BERT.

Class	Precision	Recall	F1 Score
concrete	0.33	1.00	0.50
middle	0.00	0.00	0.00
abstract	0.00	0.00	0.00

Table 4: Another run of Visual BERT.

As can be seen in Tables 3 and 4 the model is randomly predicting one category in both cases, the concrete one in the first run and the middle category in a later run.

4.3 Training the models

In order to explore exactly the capabilities of these models in classifying words as abstract or concrete, they were trained and tested again on the specific dataset collected for this task. After splitting the dataset into training and testing, the models were trained on various hyperparameters in order to achieve consistent scores. However, the results did not come back very promising.

Class	Precision	Recall	F1 Score
concrete	0.33	1.00	0.50
middle	0.00	0.00	0.00
abstract	0.00	0.00	0.00

Table 5: Textual BERT model after training.

Class	Precision	Recall	F1 Score
concrete	0.79	0.99	0.88
middle	0.93	0.51	0.66
abstract	0.80	0.96	0.87

Table 6: Visual BERT model after training.

As seen in Table 5, Textual BERT does not seem to have improved after training. Its predictions still seem to be random and it predicts only one category. After running it for several times, experimenting with different batch sizes, more epochs, and learning rates this behavior does not seem to change. Moreover, increasing the dataset did not help the model learn to distinguish between concrete and abstract instances.

As for the Visual BERT model, it seems to have improved after training. Table 6 indicates a significant increase in performance compared to Tables 3 and 4. The model seems to have predicted correctly mostly the middle category, with still high precision scores on the other two classes, whereas it captured almost all the concrete and abstract instances in the dataset, but not the middle ones. This means that for the middle category, the model reliably predicts those instances, but it cannot always detect them. The F1 scores indicate that there is an overall balance between precision and recall at least for the concrete and abstract classes.

However, after training the model several times the highest scoring class seems to change, indicating that there is still some randomness in predictions.

5 Discussion

The results discussed in the previous section are not consistent for making reliable conclusions about the models' performance and the role of indirect grounding in predicting abstractness. Training the visual model on the downstream task seems to have superficially improved its performance, however the highest scoring class is different after each training, indicating that the dataset and the training process should be further investigated. The same applies to the Textual BERT model, where investigating the data and the training instances more closely could shed light to its behavior.

Furthermore, it cannot be reliably deducted either whether and to what extent grounding really improves the classification of concrete and abstract concepts, since the results of the two models are not comparable. Even comparing between classes in the individual models' predictions cannot lead to concrete conclusions about the best performing category since that changes after each training.

The studies discussed already, have shown that indirect and normal grounding help the models make better predictions, especially for the concrete category, which shows that concreteness is more tied to symbols and that abstractness can benefit from indirect grounding. (Cerini et al., 2022), have also shown that when researching human abstract-noun-image associations, the visual scene strengthens the linguistic associations which they find to be correlated with image-abstract noun ratings. They even find that the visual context plays a major role in the associations that arise. This means that language is not only situated in words but is a bigger part of human perception and cognition, and modeling abstract concepts might be more complicated than the way it was approached in this study.

6 Conclusions and Future Work

This study attempted to explore different modality models' performance and the role of direct and indirect grounding in predicting concrete and abstract concepts respectively. For this project, a dataset of words and their corresponding images was collected and two BERT models were trained and tested on it. However, the results discussed previously are not very promising and the models' behavior should be investigated further.

Future work should further investigate the associations and dynamic interplay between linguistic and visual context and how can they be modeled in order to approximate how humans use language and establish connections between words and objects in the world. One experiment that might be worth exploring would be to investigate the regions in an image that are activated or associated a single noun (region to word alignment), in order to find the specific areas that contribute the most to its understanding. Understanding how the models connect visual and linguistic elements can give insights into their behavior and ultimately how to model linguistic information in a more optimal way.

References

S. A. Bhaskar, M. Köper, S. Schulte Im Walde, and D. Frassinelli. 2017. Exploring multimodal Text+Image models to distinguish between abstract and concrete nouns. In N. Asher, J. Hunter, & A. Lascarides (Eds.), *Proceedings of the IWCS Workshop on Foundations of Situated and Multimodal Communication*. URL: https://aclanthology.org/W17-7101.

M. Brysbaert, A. B. Warriner, and V. Kuperman. 2014. Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior Research Methods*, 46(3), 904–911. `https://doi.org/10.3758/s13428-013-0403-5`.

Ludovica Cerini, Eliana Di Palma, and Alessandro Lenci. 2022. From Speed to Car and Back: An Exploratory Study about Associations between Abstract Nouns and Images. In *Proceedings of the 2022 CLASP Conference on (Dis)embodiment*, Simon Dobnik, Julian Grove, and Asad Sayeed (Eds.), Gothenburg, Sweden, Association for Computational Linguistics, Sep, 80–88. `https://aclanthology.org/2022.clasp-1.9`

G. Dove. 2014. Thinking in words: language as an embodied medium of thought. *Topics in Cognitive Science*, 6, 371–389. doi: 10.1111/tops.12102.

G. Fauconnier. 1985. *Mental Spaces*. Cambridge, MA: MIT Press.

F. Günther, T. Nguyen, L. Chen, C. Dudschig, B. Kaup, and A. M. Glenberg. 2020. Immediate sensorimotor grounding of novel concepts learned from language alone. *Journal of Memory and Language*, 115:104172. doi: 10.1016/j.jml.2020.104172.

Stevan Harnad. 1990. The symbol grounding problem. *Physica D*, 42(1-3):335–346.

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 [cs.CV]*.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, 7(3):535–547, 2019.

C. Kennington and O. Natouf. 2022. The symbol grounding problem re-framed as concreteness-abstractness learned through spoken interaction. In: *Proceedings of the 26th Workshop on the Semantics and Pragmatics of Dialogue - Full Papers*, Aug. 2022.

George Lakoff and Mark Johnson. 1980. *Metaphors We Live By*. Chicago: University of Chicago Press.

R. W. Langacker. 1987. *Foundations of Cognitive Grammar: Vol. 1. Theoretical Prerequisites*. Stanford, CA: Stanford University Press.

Victor Sanh, Lysandre Debut, Julien Chaumond, Thomas Wolf. 2019. *DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter*. ArXiv preprint, arXiv:1910.01108.

E. Shutova, D. Kiela, J. Maillard. 2016. Black holes and white rabbits: metaphor identification with visual features. In: *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 160–170.

A. Utsumi. 2022. A test of indirect grounding of abstract concepts using multimodal distributional semantics. *Frontiers in Psychology*, 13.