# On the Effects of Video Grounding on Language Models

**Ehsan Doostmohammadi** and **Marco Kuhlmann**
Linköping University, Linköping, Sweden
{ehsan.doostmohammadi, marco.kuhlmann}@liu.se

## Abstract

Transformer-based models trained on text and vision modalities try to improve the performance on multimodal downstream tasks or tackle the problem of lack of grounding, e.g., addressing issues like models' insufficient commonsense knowledge. While it is more straightforward to evaluate the effects of such models on multimodal tasks, such as visual question answering or image captioning, it is not as well-understood how these tasks affect the model itself, and its internal linguistic representations. In this work, we experiment with language models grounded in videos and measure the models' performance on predicting masked words chosen based on their *imageability*. The results show that the smaller model benefits from video grounding in predicting highly imageable words, while the results for the larger model seem harder to interpret.

## 1 Introduction

A traditional language model is only exposed to textual data. While ample information exists in the form of text, some text-external knowledge might be missing, such as commonsense knowledge about the physical world, how objects look like, relate to each other, and how we interact with them. There is an abundance of work on trying to expose language models to other information sources and modalities, or in other words, grounding them; however, it is not clear how that would affect a language model in general. One promising modality to ground language models in is vision. Previous work has studied the grounding of language models in visual input and how this affects their performance on downstream multimodal tasks, such as visual question answering and image retrieval (Touvron et al., 2021; Li et al., 2020b; Lu et al., 2019; Su et al., 2019), and on models' "understanding" of the world and their grasp of commonsense knowledge (Sileo, 2021; Hendricks and Nematzadeh, 2021; Norlund et al., 2021).

Our aim with this work is to see whether *grounding in videos* affects the performance of transformer-based language models on masked lnaguage modeling. Masked language modeling is the task of predicting one or more masked tokens, given other tokens in the sentence. Evaluating a model's performance on such cloze-test-style fill-in-the-blank tasks is simple to implement and does not require expensive annotated data. Still, it can provide us with helpful intuition about how models work. This method also makes it easy to compare language models grounded in different modalities without further fine-tuning them. We choose to experiment with videos rather than images because they contain more information about the physical world, and may be more useful for the development of spatial, temporal and causal reasoning. Videos are also less studied in the literature.

The masked words that we want the model to predict are chosen based on their *imageability*. Imageability is a well-established notion from the field of psychology, defined as "the ease with which a word gives rise to a sensory mental image" (Paivio et al., 1968). For instance, words like "to prance" and "oven" are considered highly imageable, while words like "to consider" and "problem" are not. Imageability is highly correlated with concreteness, but the class of imageable words also includes abstract words, e.g. emotion words such as "anger". At the same time, this class does not include less experienced, yet concrete, words such as "armadillo" (Paivio et al., 1968). We use a dataset consisting of 2,645 words annotated with their imageability scores (Bird et al., 2001) to experiment with different types of models and investigate whether there is a performance difference between grounded and not grounded language models when predicting low-imageability versus high-imageability words. The words in our dataset are labeled with their parts-of-speech, which we will use in our experiments and analysis of the results.

*Proceedings of the 29th International Conference on Computational Linguistics*, pages 1–6
October 12–17, 2022.

We continue the paper with explaining the models' architecture in §2 and the data sets used in §3. The experimental settings and the results are described in §4, where we also analyze the results and try to interpret them. In §5 we briefly discuss some related work.

## 2 Model

We mainly follow the data preprocessing steps, the architecture, and the training regime of VideoBERT model (Sun et al., 2019). We experiment with a pre-trained BERT-base model (Vaswani et al., 2017) and DistilBERT (Sanh et al., 2019). BERT is essentially a transformer-based model (Vaswani et al., 2017) pretrained with masked language modeling and next sentence prediction objectives, and DistilBERT is the distilled version of the BERT model, which has half the number of layers as BERT-base. Both language models are pretrained on the same data.

As for the video features, we use the I3D model pre-trained on the Kinetics dataset (Carreira and Zisserman, 2017) to encode video clips that are sampled at 20 fps and are 1.5 seconds long. We then apply hierarchical $k$-means clustering to the video features, setting the number of hierarchy levels to 4 and the number of clusters per level $k$ to 12, which results in $12^4 = 20{,}736$ clusters. Henceforth, we use the closest cluster centroids as video tokens instead of continuous video features. As the output of the I3D model is of size 600, we use a fully connected layer to map to the embedding size of the respective model.

We further train the pre-trained language models with a masked language modeling training objective with a masking probability of 0.15 for each modality. The embeddings of the word tokens ($w_i$) and the video tokens ($v_j$) are concatenated with a new special token [>] as the text–video separator. This results in an input $I$ of the form

$$I = ([\text{CLS}], w_1, ..., w_n, [>], v_1, ..., v_m, [\text{SEP}])$$

The [CLS] and [SEP] tokens are the models' special tokens for classification and separation of sentences, respectively. The embedding weights and video features are frozen during training. The input $I$ is then fed to the model to get the output $O$, which is mapped to the vocabulary space by means of a projection layer consisting of two fully connected layers ($FC$) and layer normalization:

$$\hat{y} = FC_2(LN(FC_1(O)))$$

All the new layers and embedding weights are initialized randomly from a uniform distribution (He et al., 2015). The final objective is to maximize the log-likelihood $\sum_{l=1}^{L} \log p(\hat{y}_l \mid x_{\setminus l}; \theta)$, where $l$ is the masked token, and the $x$s are the input tokens, text or video, without the $l$th token. Special tokens are never masked.

We train two models with almost the same architecture, as described above, once only with textual input, and once with text and video input. The only difference in the structures is that the text model lacks the projection layer, which makes comparison between the models possible. The random seed is the same for both models all the time and changes by epoch. The models are trained using the Adam optimizer with a learning rate of $10^{-5}$ and batch size of $2^{10}$. We stop the training when the model's loss and accuracy plateaus on the validation set.

The implementations and the pretrained weights of the Hugging Face Transformer library (Wolf et al., 2019) are used in these experiments.

## 3 Data

To get imageability scores for nouns and verbs, we use Bird's dataset, in which words with different parts-of-speech, are rated with imageability scores from 100 to 700. For training and testing the models, HowTo100M (Miech et al., 2019) dataset is used, which is a collection of 1.2M narrated English YouTube videos from various categories. We randomly choose 55K videos from the dataset and split these into 45K videos for training, 5K for development, and 5K for testing, or in other words 4.7M samples for training and ~500K for the other sets. To get some idea on how the HowTo100M data looks, we measure the mean imageability score on a random set of 300K tokens from the dataset, which was 454 on type level, and 366 on token level, which shows a high frequency of low imageability words in the dataset.

There are a total of 892 verb types and 1,304 noun types in the Bird dataset. We split the words in the Bird dataset into low imageability ($\leq 300$) and high imageability ($\geq 500$) ones. This results in 114 low imageability and 511 high imageability types, or 67K and 92K tokens, respectively. The type-token ratio for low imageability words is $17 \times 10^{-4}$, while being $55 \times 10^{-4}$ for highly imageable ones.

| | Train | Test | Imageability | Accuracy (Δ) | N. Acc. (Δ) | V. Acc. (Δ) |
|---|---|---|---|---|---|---|
| DistilBERT | Baseline | | Low | 22.1 | 22.8 | 22.1 |
| | | | High | 10.1 | 10.5 | 9.4 |
| | T | T | Low | 34.3 | 24.0 | 35.7 |
| | | | High | 16.8 | 16.6 | 17.5 |
| | TV | TV | Low | 33.7 (-0.6) | 23.7 (-0.3) | 35.0 (-0.7) |
| | | | High | 17.7 (0.9) | 17.2 (0.6) | 18.9 (1.4) |
| | TV | T | Low | 34.1 (-0.2) | 24.0 (0.0) | 35.5 (0.2) |
| | | | High | 17.1 (0.3) | 16.7 (0.1) | 18.0 (0.5) |
| BERT | Baseline | | Low | 24.4 | 23.7 | 24.4 |
| | | | High | 10.8 | 12.4 | 7.5 |
| | T | T | Low | 38.6 | 32.8 | 38.6 |
| | | | High | 21.4 | 21.5 | 21.2 |
| | TV | TV | Low | 39.1 (0.5) | 32.9 (0.1) | 39.6 (1.0) |
| | | | High | 21.9 (0.5) | 21.8 (0.3) | 22.0 (0.8) |
| | TV | T | Low | 39.3 (0.7) | 33.0 (0.2) | 39.7 (1.2) |
| | | | High | 21.0 (-0.4) | 21.1 (-0.4) | 20.7 (-0.5) |

Table 1: Accuracy on low and high imageability words for the DistilBERT and BERT models. The results columns are for the overall accuracy (noun and verb), the noun accuracy, and the verb accuracy, respectively. The Δ is the difference between that result and the corresponding result (in terms of imageability) of the T-T model. The baseline is the model with pre-trained weights, but not fine-tuned on this dataset. For more details about the T and TV abbreviations refer to the text.

## 4 Results and Analysis

The model is fed with those sentences from the HowTo100M dataset that contains at least one word from Bird's dataset. For each sample, we only mask one noun or verb at a time to make the analysis simple. The experiments are done on two models and in three different settings:

(1) only textual input to the text-only model (T-T),

(2) text and video input to text and video model (TV-TV), and

(3) only text input to text and video model (TV-T).

The same settings are repeated for both DistilBERT and BERT-base models.

Table 1 contains the main token-level results of masked word prediction accuracy of the aforementioned three different scenarios on low and high imageability nouns and verbs. The overall accuracy is simply a weighted sum of the noun and verb accuracy. For DistilBERT, which is the smaller of the two models, the results show an increase in performance on high imageability when the model is grounded in videos (TV-TV). For the same scenario, but with low imageability words, we see some decrease in performance, which might be due to the model treating the video signal as noise. The performance goes up when removing the video from the input of the same model (TV-T). For high imageability words in the same TV-T setting, the results show some increase compared to the T-T setting, which might be due to the model learning information from the video input which is useful to masked word prediction task, even in the absence of the video signal.

On the other hand, for BERT, numbers are harder to interpret. We still see some increase for high imageability words, and more for verbs compared to nouns, but we see more or less the same amount of increase for low imageability words. It is hard to say why this is happening only for the BERT model, but one reason might be that the model receives more learning signals during training when the sequences are longer (TV), hence the higher number of masked tokens. Removing the video input from the input (TV-T) hurts the high imageability words the most, which shows the dependence of the model on the video signal. These results are not consistent with the DistilBERT model.

One should bear in mind that the relative increase in accuracy for high imageability words, e.g., between T-T and TV-TV, is higher than for low imageability ones, as the accuracy for low imageability words is always considerably higher than that of the high imageability ones. For example, an increase of $1.4\%$ in high imageability verb prediction accuracy in the DistilBERT TV-TV model is a $7.4\%$ relative increase, while $1.0\%$ for BERT TV-TV low imageability verbs is only a $2.5\%$. One should also consider the fact that low imageability words have a much higher frequency in the data, which means the model has seen them more often. While the average imageability score in the Bird dataset is around 460, the average token-based imageability score is around 360 for Howto100M and some other datasets, including Violin (Liu et al., 2020), and TVQA subtitles (Lei et al., 2018).

The differences between different models' performances are not large, however, considering the size of the test set, they are quite significant. Additionally, a bootstrap test always shows a p-value of smaller than $3.9e-5$, which indicates a very high significance for all the results. We ran the bootstrap test as described in Berg-Kirkpatrick et al. (2012): a sample $x^{(i)}$ of the same size as the test set is drawn with replacement for $b = 10^6$ times, and p-value is calculated as $s/b$, where $s$ is the number of times where $\delta(x^{(i)}) > 2\delta(x)$ holds. $\delta$ is the performance difference of systems $A$ and $B$, and $x$ is the original test set.

Comparing DistilBERT T-T and TV-TV shows that the words that benefit from the video signal are predominantly highly imageable ones, e.g., *add, cook, plant, hair, bottom, turn, pour, house, remove,* and *ground*, while low imageability words, such as *see, want, go, way, take,* and *like*, see a reduction in prediction accuracy. *Go* is a special verb in the sense that it typically appears as an auxiliary verb to indicate the future tense, which is low in imageability. When removing the video signal in BERT (TV-T), high imageability words see a reduction in accuracy, while it is the opposite for the low imageability ones. Interestingly, the top 30 words that benefit the most from the video signal in DistilBERT (TV-TV) have a 63% overlap with the ones that see the most reduction when removing the signal in BERT (TV-T). BERT (T-T) is already good at predicting the words (high or low in imageability), and does not benefit from the video signal as DistilBERT. However, training it on video signals apparently makes it more dependent on them for predicting high imageability words, so that removing the signal hurts the performance.

## 5 Related Work

Recent work on visual grounding has explored the effects of joint modeling of paired textual and visual modalities, with a focus on neural models based on the Transformer architecture (Frank et al., 2021; Li et al., 2020b; Chen et al., 2020; Huang et al., 2020; Lu et al., 2019). There is also some work that goes deeper into the problem, such as Sileo (2021), who studies the effects of visual grounding on text processing abilities of a language model using transferred and associative grounding, and how they improve text-only baselines, such as commonsense-related downstream tasks.

Another work is Hendricks and Nematzadeh (2021), who study how text-image pre-trained transformer models perform in situations that require "noun or verb understanding". According to them, such models perform poorly when evaluated on verbs compared to other parts of speech. Ebert and Pavlick (2020) experiment with an interactive simulated kitchen environment and conclude that certain machine learning models predict verbs less accurately than nouns, given a scene. They are motivated by work in psychology showing that predicting actions (verbs) is much harder than predicting objects (nouns) for people, given a video scene and the linguistic context of the word (Gillette et al., 1999).

In this work, we mainly followed VideoBERT (Sun et al., 2019), but there are other methods of integrating text and video as well. One other work is HERO (Li et al., 2020a), which does not use discretized video features, but continuous features with a regression loss. One other interesting work is ClipBERT (Lei et al., 2021), which tries to utilize sparse sampling to use fewer video frames to improve the text-video downstream tasks. There are also some work on joint representation of text and video, such as ActBERT (Zhu and Yang, 2020) and MIL-NCE (Miech et al., 2020).

## 6 Conclusion and Future Work

Although it is hard to draw strong conclusions based on these results, it might be that smaller models benefit more from video grounding than larger ones in the task of masked token prediction. The results are in line with the recent work on image grounding (Iki and Aizawa, 2021; Li et al., 2021), which suggests that the visual input might not be exploited by the model to the fullest. While the results are not strongly indicative, these models are relatively small and training data size is also minimal. The data size is chosen based on the results in Sun et al. (2019), who show that this much data should be enough to see some improvement. Increasing the data and model size could be a direction for future work. Another interesting research question that was not addressed in this paper is whether we really need to ground in videos for the model to gain the relevant knowledge, or can get the same results by using images or sampled key frame(s).

## Acknowledgements

This work was partially supported by the Wallenberg AI, Autonomous Systems and Software Program (WASP) funded by the Knut and Alice Wallenberg Foundation. The computations and data handling were enabled by resources provided by the Swedish National Infrastructure for Computing (SNIC), partially funded by the Swedish Research Council through grant agreements no. 2021/7-111 and 2021/23-556, and by the Berzelius resources provided by the Knut and Alice Wallenberg Foundation at the National Supercomputer Centre.

## References

Taylor Berg-Kirkpatrick, David Burkett, and Dan Klein. 2012. An empirical investigation of statistical significance in nlp. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 995–1005.

Helen Bird, Sue Franklin, and David Howard. 2001. Age of acquisition and imageability ratings for a large set of words, including verbs and function words. *Behavior Research Methods, Instruments, & Computers*, 33(1):73–79.

Joao Carreira and Andrew Zisserman. 2017. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308.

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. Uniter: Universal image-text representation learning. In *European conference on computer vision*, pages 104–120. Springer.

Dylan Ebert and Ellie Pavlick. 2020. A visuospatial dataset for naturalistic verb learning. In *Proceedings of the Ninth Joint Conference on Lexical and Computational Semantics*, pages 143–153, Barcelona, Spain (Online). Association for Computational Linguistics.

Stella Frank, Emanuele Bugliarello, and Desmond Elliott. 2021. Vision-and-language or vision-for-language? on cross-modal influence in multimodal transformers. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9847–9857.

Jane Gillette, Henry Gleitman, Lila Gleitman, and Anne Lederer. 1999. Human simulations of vocabulary learning. *Cognition*, 73(2):135–176.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*, pages 1026–1034.

Lisa Anne Hendricks and Aida Nematzadeh. 2021. Probing image-language transformers for verb understanding. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3635–3644.

Zhicheng Huang, Zhaoyang Zeng, Bei Liu, Dongmei Fu, and Jianlong Fu. 2020. Pixel-bert: Aligning image pixels with text by deep multi-modal transformers. *arXiv preprint arXiv:2004.00849*.

Taichi Iki and Akiko Aizawa. 2021. 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*, pages 2189–2196.

Jie Lei, Linjie Li, Luowei Zhou, Zhe Gan, Tamara L Berg, Mohit Bansal, and Jingjing Liu. 2021. Less is more: Clipbert for video-and-language learning via sparse sampling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7331–7341.

Jie Lei, Licheng Yu, Mohit Bansal, and Tamara Berg. 2018. TVQA: Localized, compositional video question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1369–1379, Brussels, Belgium. Association for Computational Linguistics.

Jiaoda Li, Duygu Ataman, and Rico Sennrich. 2021. Vision matters when it should: Sanity checking multimodal machine translation models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8556–8562.

Linjie Li, Yen-Chun Chen, Yu Cheng, Zhe Gan, Licheng Yu, and Jingjing Liu. 2020a. HERO: Hierarchical encoder for Video+Language omni-representation pre-training. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2046–2065, Online. Association for Computational Linguistics.

Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. 2020b. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *European Conference on Computer Vision*, pages 121–137. Springer.

Jingzhou Liu, Wenhu Chen, Yu Cheng, Zhe Gan, Licheng Yu, Yiming Yang, and Jingjing Liu. 2020. Violin: A large-scale dataset for video-and-language inference. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10900–10910.

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.

Antoine Miech, Jean-Baptiste Alayrac, Lucas Smaira, Ivan Laptev, Josef Sivic, and Andrew Zisserman. 2020. End-to-End Learning of Visual Representations from Uncurated Instructional Videos. In *CVPR*.

Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. 2019. HowTo100M: Learning a Text-Video Embedding by Watching Hundred Million Narrated Video Clips. In *ICCV*.

Tobias Norlund, Lovisa Hagström, and Richard Johansson. 2021. Transferring knowledge from vision to language: How to achieve it and how to measure it? In *Proceedings of the Fourth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 149–162.

Allan Paivio, John C Yuille, and Stephen A Madigan. 1968. Concreteness, imagery, and meaningfulness values for 925 nouns. *Journal of experimental psychology*, 76(1p2):1.

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

Damien Sileo. 2021. Visual grounding strategies for text-only natural language processing. In *Proceedings of the Third Workshop on Beyond Vision and LANguage: inTEgrating Real-world kNowledge (LANTERN)*, pages 19–29.

Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. 2019. Vl-bert: Pre-training of generic visual-linguistic representations. *arXiv preprint arXiv:1908.08530*.

Chen Sun, Austin Myers, Carl Vondrick, Kevin Murphy, and Cordelia Schmid. 2019. Videobert: A joint model for video and language representation learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7464–7473.

Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. 2021. Training data-efficient image transformers & distillation through attention. In *International Conference on Machine Learning*, pages 10347–10357. PMLR.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

Linchao Zhu and Yi Yang. 2020. Actbert: Learning global-local video-text representations. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8746–8755.