CSE 517 Final Report

Johanna Karras, Abhishek Babu, Phuong Nguyen
Paul G. Allen School of Computer Science & Engineering
University of Washington
{jskarras, babua, phdn}@uw.edu

Reproducibility Summary

Scope of Reproducibility

The paper *Does Vision-and-Language Pretraining Improve Lexical Grounding?* [7] claims that vision-language input to transformer models does not give additional supervision compared to language-only input, specifically for the tasks of Physical Commonsense QA [1] and adjective-noun composition [4].

Methodology

Our implementation was based on the authors' original code and pretrained models. We replicated the authors' described experiments using the original datasets. For the first task, we replicated the fine-tuning experiments of the paper using the provided pre-trained models and using two of the CSE Titan Xp GPU's for 8 days. For the second task, we utilized the authors word embeddings of nouns to replicate (1) the Kmeans clustering of adjectives and nouns together and (2) the evaluation of adjectives, given specific nouns.

Results

Our reproduction of the experiments mostly support the authors claims. For the first experiment, we reproduced the experiments' results to within 0.12% of the original paper's reported values, and our results support the authors' claim vision-and-language data does not provide additional supervision for question-answer tasks in terms of accuracy. However, in the second experiment, not all of our qualitative examples of TSNE plots of noun embeddings were as poor in quality as the authors' provided example, leading us to believe that the authors over-generalized their claims about the adjective-noun composition results to a specific example. We were unable to observe significant difference between models trained on only text and on vision-text inputs on the various natural language tasks.

What was Easy

Both tasks were straightforward to implement, as the authors provided the necessary scripts, datasets, and commands for running the experiments. Since much of the model scripts were abstracted behind simple shell script commands, rerunning the experiments was a simple and straightforward task.

What was Difficult

The fine-tuning required for the transformer probing in the first QA task was unexpectedly very resource- and time-consuming. We were unable to run the transformer probing in the given time with the hardware resources available to us. Additionally, one of the tasks described in the paper regarding coreference and semantic roles was implemented by the authors using the research conducted in another paper. The implementation of this experiment was not provided and not clearly explained in the paper, and as such, we could not reproduce this task.

Communication with Original Authors

While we haven't been in contact with the authors during the reproduction of their experiments, we plan on emailing the final report and code repository to the authors of the paper to get any feedback they might have for us.

1 Introduction

In this paper, we reproduce the experiments from the paper *Does Vision-and-Language Pretraining Improve Lexical Grounding?*, published at EMNLP 2021 [7]. The authors compare models for text-based segmentation tasks when pre-trained using text-only models versus vision-and-language (VL) models. Namely, the authors compare pretrained BERT [2], VideoBERT [6], and VisualBERT models [5] for question-answering, semantic representations of verbs, and inference about object properties tasks.

The first task considered is whether VL pretraining produces gains in benchmark NLP tasks that rely on multimodal knowledge using the Physical Commonsense QA [1] as a benchmark for reasoning. The second task considered is whether multimodal pretraining affects the conceptual structure of linguistic elements at the lexical level using the WikiHow [4] and MIT States datasets [3] to investigate adjective-noun compositions under the different pretrained models.

Ultimately, the authors claim that multimodal models have no significant benefit over their counterparts that are trained using only text.

2 Scope of Reproducibility

The paper concludes that that while vision-and-language (VL) pretraining sometimes produces gains in accuracy and F1-score in both question-answering and adjective-noun composition, the margins of improvement are too small to to support any conclusions that VL pretraining in its current state can produce improvements for NLP tasks in general.

On the Physical Commonsense QA reasoning task, both the VisualBERT and VideoBERT models yield marginally higher accuracy when pretrained on VL data than the text-only pretrained BERT model. Similarly, when the VisualBERT and VideoBERT models are pretrained on text-only data, they will yield marginally lower accuracy than their counterparts trained on VL data.

On the task of classifying adjectives that modify nouns in adjective-noun compositions (conceptual structure), both the VisualBERT and VideoBERT models pretrained on text-only data and pretrained on VL data will yield higher accuracy than the text-only pretrained BERT model. However, for each of the VisualBERT and VideoBERT models, training on VL data yields very similar accuracy (neither a consistent increase nor a consistent decrease) compared to training on text-only data.

We will replicate the experiments using the author's provided pre-trained models of BERT, VisualBERT and VideoBERT in an attempt to reproduce the claims from the paper.

2.1 Addressed Claims from the Original Paper

1. Compared to language-only data, vision-and-language data does not provide additional supervision for question-answer tasks in terms of accuracy.
2. Vision-and-language models do not significantly outperform language-only models on adjective-noun composition tasks in terms of qualitative clustering of adjectives and nouns

3 Methodology

3.1 Model Descriptions

The first of three model used in the original paper's experiment is BERT, a language representation model, which stands for **B**idirectional **E**ncoder **R**epresentations from **T**ransformers [2], and is based on the Transformer architecture. The BERT model had two learning objectives. The first objective was a masked language modeling (MLM) objective to predict masked out tokens based on surrounding context, and the second objective was a next sentence prediction (NSP) objective to predict if the second sentence in a pair of sentences input is the subsequent sentence in the document. The BERT model was pre-trained on text-only data. The original paper specifically used the pre-trained $\text{BERT}_{\text{BASE}}$ model that has 12 layers of transformer blocks, 768 hidden nodes, 12 self-attention heads, and 110M parameters.

The VideoBERT model, a joint visual-linguistic model to learn high-level features from video [6], and the VisualBERT model, a joint visual-linguistic model to learn semantics from images and associated text [5], are both built upon BERT, and is based on the Transformer architecture. The learning objectives of both VideoBERT and VisualBERT were similar to that of the $\text{BERT}_{\text{BASE}}$ model, but modified to the multimodal (visual and textual information) setup. The first learning

objective was a masked language modeling (MLM) objective to predict masked out tokens based on surrounding context, where the VideoBERT predicts both visual and text tokens, while VisualBERT predicts only text tokens. The second objective was a visual-language prediction objective to predict whether the visual and language sequences come from the same video/image or not. The VideoBERT and VisualBERT model pretrained on visual-language data (video and associated speech for VideoBERT, and image and associated caption for VisualBERT), are referred to as VideoBERT$_{\mathrm{VL}}$ and VisualBERT$_{\mathrm{VL}}$ models in the original paper. The pretrained text-only VideoBERT$_{\mathrm{text}}$ and VisualBERT$_{\mathrm{text}}$ model are trained using the same text data as their VL model counterpart, but with the visual data removed. Both VideoBERT and VisualBERT model are configured on BERT$_{\mathrm{BASE}}$, which has 12 layers of transformer blocks, 768 hidden nodes, 12 self-attention heads. The VideoBERT model has 125M parameters, and the VisualBERT model has 109M parameters.

3.2 Datasets

We utilized the same dataset used in the original paper to reproduce its experiments.

For the first task that asks whether VL pretraining produces gains over text-only pretraining in benchmark NLP tasks, the paper used the benchmark dataset Physical Interaction: Question Answering (PIQA) [1]. The dataset is designed to investigate the physical knowledge of existing models, where a model is asked to predict the correct solution (out of two choices) that best satisfy an instruction or a goal as the question. The dataset is defined in terms of goal and solution pairs. The goal indicates a post-condition and the solutions indicate the procedure for accomplishing this. An example of a goal/solution pair from the dataset is {"goal": "fire", "sol1": "can melt humans ", "sol2": "can melt water "} and the correct solution is accompanied by the label "0" or "1" to denote whether "sol1" or "sol2" is correct. The dataset contained 16 thousand training data points with associated labels, 3 thousand validation data points with associated labels, and 3 thousand testing data points with associated labels. The dataset is obtained from the PIQA dataset's author's website.

For the second task that asks whether multimodal pretraining affects the conceptual structure of linguistic elements at the lexical level, the paper used the WikiHow dataset [4], which contains WikiHow articles (that contains step-by-step instructions of daily tasks) and their summaries. The dataset is obtained from this download link from the WikiHow dataset GitHub repository. Additionally, the paper used the "visually-groundable" adjectives in MIT States dataset [3] as their adjective filter for this second task. The MIT States dataset contains 63,440 images depicting 245 nouns (objects, scenes, and materials) modified by a total of 115 adjectives (a state) that describes the noun's transformation (a transition from a one adjective to its antonym). The dataset is obtained from MIT States and Transitions website.

3.3 Hyperparameters

For the experiments, we used the same hyperparameters for training and testing as is described in the paper. The models have input size of 768, hidden size of 512, dropout of 0. They are are trained by cross-entropy loss and by using the Adam optimizer, with a batch size of 32, an initial learning rate of 1e-4.

3.4 Implementation

We will use the existing GitHub repository and pre-trained models (available at `https://github.com/tttyuntian/vlm_lexical_grounding`). The code is written in python using the PyTorch package. The provided repository contains documentations necessary to reproduce the paper's experiments, including dependencies, data download instructions, pretrained model download instructions, and bash script commands to precompute sentence embeddings, as well as commands to train and evaluate models to perform the probing experiments.

For our additional experiment, we manually truncated (the first) half of the PIQA dataset, and re-computed the vector embeddings and re-trained and re-evaluated the models using the same commands provided by the original authors.

3.5 Experimental Setup

Our code is available in our GitHub repository (available at `https://github.com/johannakarras/reproducibility-vlm-lexical-grounding`).

For the first task, we fine-tuned the pre-trained models (available on the paper's official Github repo) on Physical Commonsense QA [1] dataset using two of the CSE Titan Xp GPUs for a total of about 8 days.

For the second task, we created a Jupyter notebook file that we used to load the WikiHow [4] and MIT States [3] datasets, run the provided scripts from the authors' GitHub repository, and reproduce the original TSNE plots and accuracy table. This section did not require additional computational resources.

3.6 Computational Requirements

The first set of (question-answering) experiments require fine-tuning pre-trained models to their specific tasks. For the text-only BERT model, the original paper reports using the openly-available pre-trained BERT weights. For pre-training VideoBERT and VisualBERT, the original paper uses GPUs 4 Cloud TPUs for 2 days and 4 TitanV 25 hours, respectively. These models are provided as checkpoints in the GitHub repository. As such, we do not require additional computational resources for pre-training. Then, to fine-tune the models on the specific tasks, and to run the experiments, we use two of the CSE Titan Xp GPUs.

Prior to running the experiment, we estimated that it will take four days to complete each task. So, in total, the compute time will be about eight days.

In actuality, for the first task (question-answering), the precomputing sentence embeddings step took 81 minutes to complete (for 7 different varieties of models used in the experiment). The linear and MLP probing experiments took 64 minutes to complete, where each probing experiment was performed 5 times each (for a total of 70 trials for linear and MLP probing experiments accross 7 model varieties). The number of epochs varied in each run and between different models, but the most common range of epochs ran in a single trial was from 50 to 80 epochs, and takes no longer than a 2 seconds to run each epoch. We attempted to run the transformer probing experiments, but were not able to complete it with our given access to hardware resources. The transformer probing experiment on the base BERT model ran for two days before it was unexpected killed before completion. The same happened with the transformer probing experiment on the $\mathrm{VideoBERT}_{\mathrm{text}}$ model, which was also running for two days before it was unexpected killed before completion. We decided against performing the transformer probing experiments out of respect for other students in the course who also need access to these shared computing resources.

For the second set of (adjective-noun composition) experiments, we do not require extensive computational resources, because they rely on Kmeans clustering of the word embeddings from the pretrained models to find adjective-noun associations. As such, we used Google Colab to run the experiments with a single GPU. The script is located in our Github repo, called `Adjective_Noun_Composition.ipynb`. Each model takes around four hours to compute noun embeddings, reaching 24 hours in total for the runtime.

4 Results

As a whole, our reproduction of the experiments support the authors claims. We were unable to observe significant difference between models trained on only text and on vision-text inputs on the various natural language tasks. Any improvements provided by VL models are marginal at best and not consistent across different types of tasks.

4.1 Result 1: QA Does Not Benefit Significantly from Multimodal Data

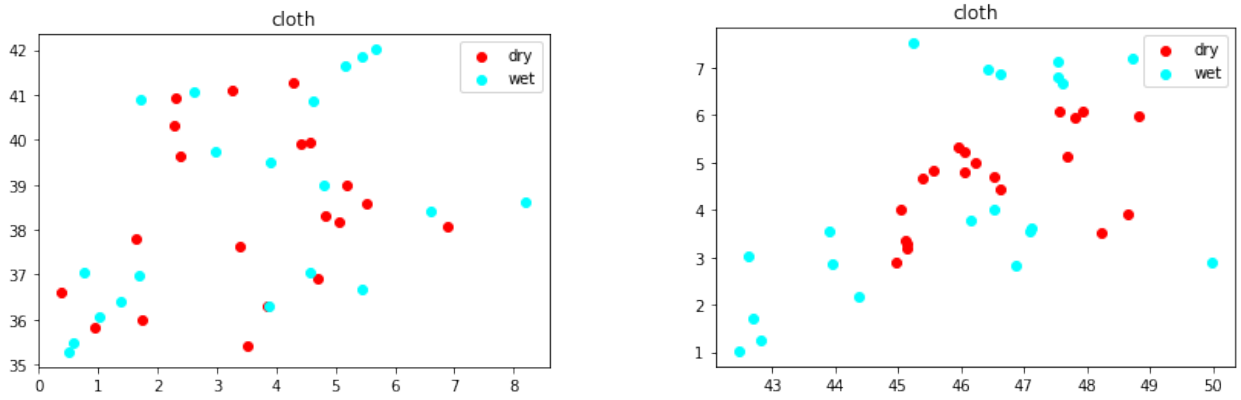
The first experiment indicates that adding vision as a modality for training VideoBERT and VisualBERT models yields marginal improvement over training on text-only data, but the gains are quite small - only a few points improvements even in the best case scenario, and not a significant improvement overall.

Encoder	Linear	MLP	Transformer
$\mathrm{BERT}_{\mathrm{base}}$	55.40 ± 0.35	58.05 ± 0.18	
$\mathrm{VideoBERT}_{\mathrm{text}}$	57.67 ± 0.34	58.99 ± 0.31	
$\mathrm{VideoBERT}_{\mathrm{VL}}$	58.86 ± 0.60	58.84 ± 0.41	
$\mathrm{VisualBERT}_{\mathrm{text}}$	55.05 ± 0.16	56.77 ± 0.37	
$\mathrm{VisualBERT}_{\mathrm{VL}}$	55.70 ± 0.14	59.22 ± 0.14	

Table 1: Accuracy $\pm$ standard deviation of different pretrained representations on the validation split of PIQA. Numbers are averaged over five runs. VL pretraining only brings marginal improvements over text-only pretraining.

We could not verify the original paper's transformer probing results due to the limited computing resources available to us, as discussed previously in the computing requirements section.

We reproduced the accuracy to within 0.12% of the orignal experiment's reported values (excluding transformer probing results), and our findings uphold the paper's conclusion that vision-and-language data does not provide additional supervision for question-answer tasks in terms of accuracy.



$\text{VideoBERT}_{\text{text}}$ $\text{VideoBERT}_{\text{VL}}$

Figure 1: TSNE projections of noun embeddings for "cloth" from $\text{VideoBERT}_{\text{text}}$ and $\text{VideoBERT}_{\text{VL}}$

.

4.2 Result 2: Adjective-Noun Composition Does Not Benefit Significantly from Multimodal Data

The second experiment indicates that adding vision as a modality for training VideoBERT provided no advantage for learning useful noun representations.

Figure 1 qualitatively compares the clustered noun representations of "cloth" for both models, with colors indicating the adjective associated with each occurrence of "glass". Notice that there is no significant difference between "dry" vs. "wet" for either model. This suggests that the learned representations of the word "glass" are not useful for identifying its qualities.

Table 2 compares the homogeneity, completeness, and V-measure scores for the noun clusters generated by BERT, VideoBERT, and VisualBERT, fine-tuned with and without vision input. Each value is averaged over five runs, as in the original paper. Note that according to the paper "homogeneity of 1" means that every point in a cluster belongs to the same class. "Completeness of 1" means that every point belonging to a given class is in the same cluster. Vmeasure is the harmonic mean of the two" [7]. Similar to the results in the paper, the differences in averaged values for vision-text versus text-only models is negligible. In fact, the VideoBERT and VisualBERT models trained with text-only data slightly out-perform their respective models trained with vision and text input.

Encoder	Homo.	Compl.	V-Meas.
$\text{BERT}_{\text{base}}$	0.288 ± 0.054	0.301 ± 0.056	0.293 ± 0.055
$\text{VideoBERT}_{\text{text}}$	0.283 ± 0.0307	0.304 ± 0.034	0.292 ± 0.029
$\text{VideoBERT}_{\text{VL}}$	0.283 ± 0.030	0.303 ± 0.033	0.292 ± 279
$\text{VisualBERT}_{\text{text}}$	0.301 ± 0.027	0.309 ± 0.026	0.304 ± 0.027
$\text{VisualBERT}_{\text{VL}}$	0.285 ± 0.038	0.303 ± 0.049	0.291 ± 0.040

Table 2: Summary metrics for clustering noun embeddings according to their adjective modifiers.

4.3 Additional Results not Present in the Original Paper

We suspect that the text-only training data is large and robust in the cooking setting, meaning that the additional information gained from visual input may be unnecessary for improving model improvement on language tasks. As such, we propose to conduct an additional experiment where we rerun the text-only and VL models using a limited dataset, in order to see if VL models have an advantage by being able to glean missing information from the visual input.

We also hypothesize that the models are over-fitted to the domain of cooking videos, so testing on non-cooking texts is not representative of the actual capability of these models to perform tasks. As such, we conduct an experiment for the task of adjective-noun composition that uses only words that would appear in the cooking domain.

4.3.1 Additional Results for QA Task

We investigated whether or not rerunning the models with half of the original experiment's PIQA training dataset would result in a more significant improvement in accuracy achieved by the VL models compared to the text-only models. Specifically, we want to see whether the models trained on VL data have an advantage in settings with lower amounts of data by being able to able glean missing information from the visual input.

Encoder	Linear	MLP	Transformer
$\text{BERT}_{\text{base}}$	53.70 ± 0.24	57.42 ± 0.11	
$\text{VideoBERT}_{\text{text}}$	56.91 ± 0.27	56.86 ± 0.27	
$\text{VideoBERT}_{\text{VL}}$	58.15 ± 0.56	58.14 ± 0.26	
$\text{VisualBERT}_{\text{text}}$	54.74 ± 0.19	56.12 ± 0.11	
$\text{VisualBERT}_{\text{VL}}$	56.32 ± 0.17	58.18 ± 0.16	

Table 3: Accuracy $\pm$ standard deviation of different pretrained representations on the validation split of PIQA. Numbers are averaged over five runs. VL pretraining only brings marginal improvements over text-only pretraining.

We once again finds that VideoBERT and VisualBERT models yields marginal improvement over training on text-only data, but the gains are quite small - only a few points (2) improvements even in the best case scenario, and are not significant improvement overall.

Additionally, we once again could not run the transformer probing experiments due to limited computing resources as outlined in the computational requirement section.

4.3.2 Additional Results for Adjective-Noun Composition Task

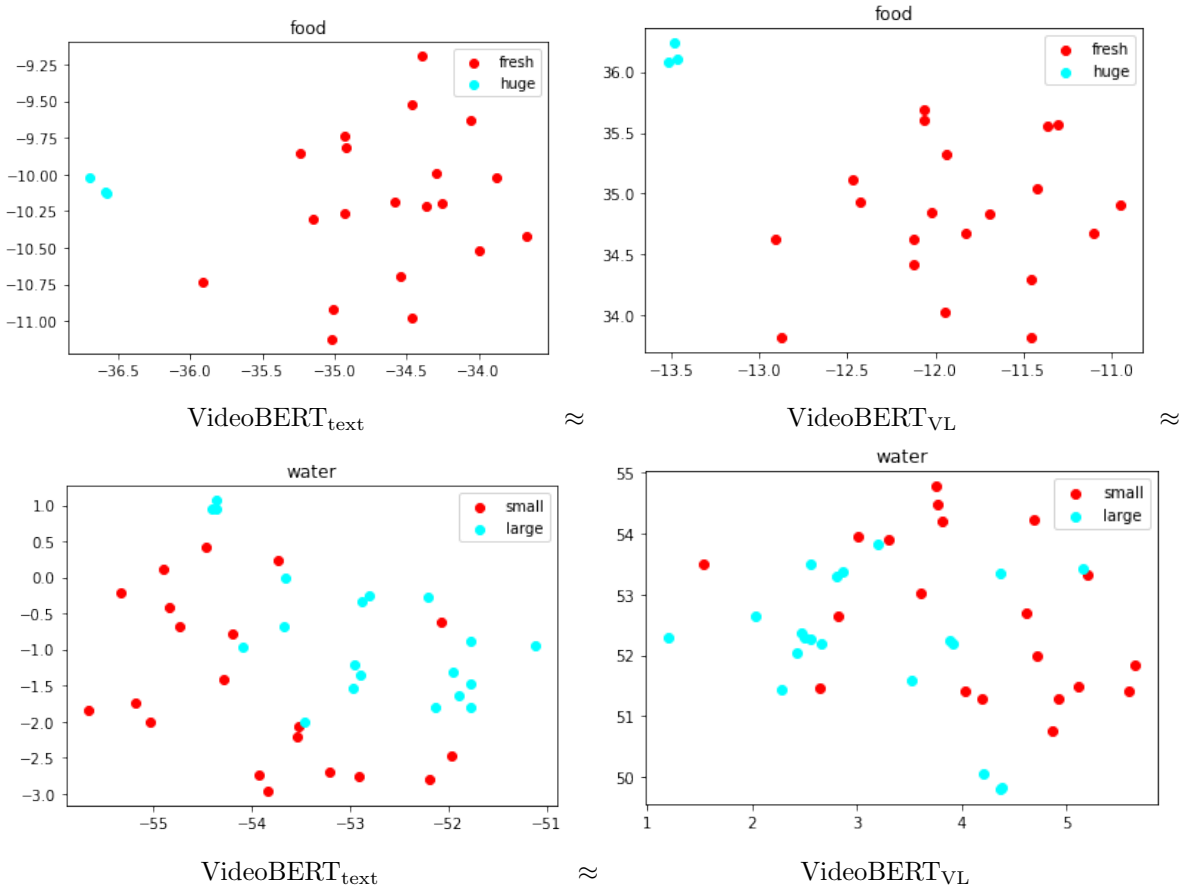


Figure 2: TSNE projections of noun embeddings for "food" and "water" from $\text{VideoBERT}_{\text{text}}$ and $\text{VideoBERT}_{\text{VL}}$

The authors only provide one qualitative example of adjective embedding clustering for the words describing "glass". Not only is this insufficient for generalization about how the models embed adjectives related to all nouns, but is a word that is only present in the training data (cooking videos) in a limited context (a glass holding water or other drink).

We investigated whether or not other noun and adjective embeddings yielded better clusters, particularly those more present in the context of the training dataset. We chose "food" and "water", since these are related to cooking. The results are shown in Figure 2.

Compared to the TSNE projections of noun embeddings for "cloth" shown in Figure 1 or the original example for "glass" in the paper, the clusters for "food" and "water" are not so poor. There is a visible distinction in both $\mathrm{VideoBERT}_{\mathrm{text}}$ and $\mathrm{VideoBERT}_{\mathrm{VL}}$ between "fresh" and "huge" food and between "small" and "large" water. This comes to suggest that perhaps the authors selected a particularly bad example of clustering using the multimodal transformer embeddings or that the models do not generalize well to out-of-distribution text.

5 Discussion

The original paper was mostly reproducible, thanks to the availability of the exact datasets, model configurations, and pre-trained checkpoints used by the authors for the first and third experiments described in the paper. Unfortunately, there was little instruction and no provided resources for reproducing the coreference and semantic role labelling experiment (second experiment in the paper). As such, we did not reproduce that result.

For the experiments we were able to reproduce, the evidence supports the claims of the paper. Yet, our approach faced certain limitations due to time and resources. For example, given the extensive time needed for fine-tuning the first task, we were unable to also re-train the pretrained models.

Furthermore, we believe that doing additional experimentation with more diverse training and testing datasets than in the original paper, smaller dataset sizes, and more expansive sweep of hyperparameter configurations would have yielded more robust and conclusive results.

5.1 What was Easy

Firstly, the paper was well-written and made it easy to understand their experimental setup. Secondly, the code was well-documented and the authors provided explicit instructions for how to train and test the models and methods. Additionally, the authors provided clear instructions and commands to download and extract the different datasets and where to place the data for the scripts to use them for probing. The clear documentation made it a simple and straightforward task to reproduce their experiments.

Since most of the Python scripts were abstracted behind shell scripts, and since the authors made it easy to specify the model type using script flags, reproducing the experiments was made even more simple. Additionally, the extensive logging for each experiment made it very easy to look at runtimes and results since they were all logged and persisted in log files.

5.2 What was Difficult

A challenge is that the original paper describes three tasks, but their GitHub only provides implementations for the first and third. The second task related to coreference and semantic roles, but we were unable to find the dataset publicly available online or any helpful resources for replicating this experiment.

Another challenge was the time and computing resource required to reproduce the authors' results. Factors contributing to this were the large size of the datasets, the complexity of the models chosen, the number of different models to train and test, and also the number of distinct experiments to run.

5.3 Recommendations for Reproducibility

In order to improve reproducibility, we recommend the authors provide links and resources for running the entity coreference and semantic role labeling suite of tasks described in the second experimental section of the paper. Additionally, we recommend providing abridged datasets and their corresponding results on the models, in order to make it more feasible to reproduce the results, especially for the first task which required 8 days and 2 GPU's to reproduce. Lastly, if the authors were to provide a notebook with the inference and training code, that would make reproduction easier.

Communication with Original Authors

While we haven't been in contact with the authors during the reproduction of their experiments, we plan on emailing the final report and code repository to the authors of the paper to get any feedback they might have for us.

References

[1] Yonatan Bisk, Rowan Zellers, Ronan Le Bras, Jianfeng Gao, and Yejin Choi. Piqa: Reasoning about physical commonsense in natural language. In *Thirty-Fourth AAAI Conference on Artificial Intelligence*, 2020.

[2] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[3] Phillip Isola, Joseph J. Lim, and Edward H. Adelson. Discovering states and transformations in image collections. In *CVPR*, 2015.

[4] Mahnaz Koupaee and William Yang Wang. Wikihow: A large scale text summarization dataset, 2018.

[5] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. What does BERT with vision look at? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5265–5275, Online, July 2020. Association for Computational Linguistics.

[6] Chen Sun, Austin Myers, Carl Vondrick, Kevin Murphy, and Cordelia Schmid. Videobert: A joint model for video and language representation learning, 2019.

[7] Tian Yun, Chen Sun, and Ellie Pavlick. Does vision-and-language pretraining improve lexical grounding? In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4357–4366, Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. `https://arxiv.org/pdf/2109.10246.pdf`.