DOES VISUAL INFORMATION BENEFIT PHYSICAL COMMONSENSE UNDERSTANDING? AN ANALYSIS OF LARGE LANGUAGE MODELS

Griffin Tarpenning & Rafael Esteves
CS197
Stanford University
{gritarp, resteves}@cs.stanford.edu

ABSTRACT

Recent approaches to analyzing and understanding multimodal models have mostly focused on how text affects the performance of the visual encoder, disregarding the impact visual stimuli has on the text encoder itself. A domain that might be most impacted by the inclusion of visual data is physical commonsense reasoning, where models answer seemingly trivial questions about the world, like: *Is an elephant heavier than a dog?* The detached language portions of four multimodal models (CLIP, LXMERT, UNITER, VisualBERT) are compared against three language-only models (RoBERTa-Large, RoBERTa-Small, T5) on three separate physical reasoning datasets (DOQ, VERB, PROST). This paper identifies that both architectures are able to achieve reasonable accuracy in the physical reasoning space, but the inclusion of visual data does not provide significant positive performance benefits.

1 INTRODUCTION

Natural Language Processing (NLP) is generally focused on the extraction and synthesis of information from text, and central to that task is developing understandings of the built environment. Using touch, humans recognize that a tree branch is rough while a table is smooth; through sight, a worm is small, a cow is big, and a whale is even bigger. While most humans would consider these associations trivial to identify, and thus have developed intuitive spacial understandings of the world, have the latest NLP models developed that same intuition? Language models have received considerable attention as potential solutions to learning knowledge that is assumed to be widely known, often and henceforth referred to as "common sense," (Feldman et al., 2019; Sap et al., 2020). Current language-only models like BERT and ELMO have a strong understanding of context, yet they severely under perform when it comes to common sense reasoning (Jastrzebski et al., 2018; Aroca-Ouellette et al., 2021; Li et al., 2021).

One potential reason for this could be that language datasets spend relatively little time describing things a human might consider trivial (Jastrzebski et al., 2018). One might describe a banana as red or an abnormal carrot as purple because they are extraordinary. When processing such textual data, a model might associate "green" with "banana" or "purple" with "carrot" when you might associate "yellow" with the former and "orange" with the latter. The solution to encoding better common sense understanding could lie in multimodal models, which combine NLP model structures with state of the art visual data encoding schemes, and thus have access to significantly more intuitive visual information. The combination of vision and text might allow the model to learn grounded representation of languages, which corresponds to objects in the real world (Bender & Koller, 2020).

Recent approaches to analyzing and understanding multimodal models have mostly focused on how text data affects the performance of the visual encoder. A few prominent works have found that training in this way can achieve state of the art results and improve robustness in computer vision tasks, but have not investigated the effects on the language encoder (Radford et al., 2021; Tan & Bansal, 2019b). One domain that provides a basis for this approach is language translation, where the inclusion of visual data has been shown to have a semantic grounding effect, positively affecting machine translation (Su et al., 2021; Zhou et al., 2021; Jia et al., 2021).

In this paper, we focus on the textual component of pretrained visual + language models, and compare them to pretrained text-only models by assessing their performance on common sense tasks. Many unimodal models have already been tested on their ability to answer common sense questions (Zhang et al., 2020a; Singh et al., 2021). Additionally, many multimodal models have been tested on rigorous common sense understanding tasks that require visual data at test time (Radford et al., 2021; Chen et al., 2020). However, in this study we analyze the performance difference of these two classes of models to detect if visual training helps language grounding or, more generally, improves performance in some way.

We use the word and sentence embeddings from various multimodal and unimodal models to encode information about the physical world. We choose four prominent multimodal models, CLIP, ViLBERT, LXMERT and UNITER, to compare against the following text-only models: RoBERTa, RoBERTa small, and T5. Then, sentences from three datasets are encoded and used as input to a logistic classifier.

The results of our quantitative evaluations suggest that there is no measurable relationship between multimodal language training and common sense understanding of the physical world. We measure each model's ability to understand physical attributes by comparing their performance on a collection of datasets (Yanai Elazar, 2019; Forbes & Choi, 2017; Aroca-Ouellette et al., 2021). We compare the performance of CLIP, LXMERT, ViLBERT with the previous results on these dataset and find that all models achieve very similar results. Qualitatively, we find some interesting correlations that do not present across datasets. These results allow us to better understand the effect of multimodal training on text encoders.

Our work contributes to the understanding of multimodal models and their effect on a computer's understanding important aspects of the physical world. Our experiments show that multimodal training does not lead to the language model understanding physical reasoning information. Improving an AI's commonsense knowledge has implications in creating software that better serves humans, who often do not explicitly state this type of information, as well as improving text and image retrieval tasks and search engines.

2 Related Works

2.1 The Development of Multimodality Techniques

Many current approaches to NLP use the language supervised, predictive training procedure. However, there has been some interest in using multimodal training techniques, particularly for improving computer vision (CV) based classification. Convolutional Neural Networks (CNN) trained on Flickr images drawn from (image, text) pairs predict the images corresponding captions and learn useful image features, achieving a similar performance on downstream tasks compared to image net pretrained models (Joulin et al., 2015). Radford et al continued this work by developing a contrastive multimodal learning technique in which the model tries to match (image, text) pairs (Radford et al., 2021). The zero shot CLIP model was the state of the art in many of the downstream tasks it was evaluated on and had particularly impressive performance in measures of robustness. Zhang et al finds that this approach can also perform well when trained on a specific dataset for a medical imaging task (Zhang et al., 2020b).

While an important goal of many machine learning methods is to be successful in a task-agnostic setting, the current reality is that most multimodal SOTA solutions to tasks are the product of specific fine-tuning (Yun et al., 2021). Prominent multimodal models like CLIP fail to outperform SOTA for tasks with well-defined datasets or in a many-shot learning environment. While much of the focus on multimodal techniques has been for CV tasks (Tan & Bansal, 2019a; Lu et al., 2019; Li et al., 2020), there are a few areas of NLP that multimodal models have immediately shown significant promise. Here, we don't focus on areas that *require* the combination of vision and language, like the new BD2BB challenge (Pezzelle et al., 2020), but rather on natural language tasks enhanced with the inclusion of vision.

2.2 Advantages of Multimodality

A major area where multimodal training enhances textual understanding is language translation. Trained on image data captioned with two languages, a multimodal model combining output from a CNN and bidirectional parsing of the text lead to SOTA language translation performance (Su et al., 2021). After being introduced to image data, the NLP encoder placed higher weights on different parts of the sentence, corresponding to actions that were more clearly understood in the image. Similarly, UC2 is a multimodal language translation model that beat SOTA in 5 of 7 languages tested, which the authors attribute to the integrated learning process between image and text (Zhou et al., 2021). Moreover, even multi-modal models pre-trained for a wide variety of tasks, like ALIGN, can achieve SOTA for language translation, even beating UC2 in French translation (Jia et al., 2021). Language translation appears to be an excellent example of the synergy between image and language.

Moreover, using specific scene-based images with captions, multimodal models have also been shown to successfully outperform pure-text or pure-image based similarity metrics (de Lacalle et al., 2020; Cafagna et al., 2021). Multimodal models thus have shown strength at inferring the meaning behind (perhaps ambiguous) text with the help of images, a crucial unique advantage of having both image and text. Do these improvements in language recognition get transferred to the language model even when images are not present?

Using multimodal models to train more than semantic understandings of language have been shown to improve language concepts in the past. Piglet highlights multimodality as a uniquely powerful tool to perform logically sound next-action prediction by learning to simulate possible next actions (Zellers et al., 2021). Real world interactions clearly benefit here, as neural architectures trained on a combination of visual and text data have also been shown to bridge the gap between text input and robotic action output (Giorgi et al., 2021). The language portion of these models learn more than the literal meaning of a given word, but also how the object that is associated with that word interacts with other objects. In summary, prior work has shown much interest in multimodal models and how well they improve state-of-the-art computer vision tasks. This interest largely stems from the lack of generalization that the language only models offer and the promising performance of multimodal models like CLIP. Prior work largely examines the performance of multimodal models in computer vision tasks, next action prediction as well as language processing tasks like language translation. Instead of largely focusing on multimodal models' performances in computer vision tasks and image-to-text or text-to-image mapping, our work shifts the focus to how multimodal models understand language compared to their counterparts that have only been trained on text. Though prior work has shown that multimodal models prove effective in understanding text meaning and relation between languages as well as understanding semantics, shifting the focus to natural language will help improve our understanding of how additional sensory inputs influence a model's understanding of language.

3 Evaluation

3.1 Models

We explore the benefits of multimodal training on language models by evaluating a collection of models, both unimodal and multimodal, on a set of common sense tasks. The unimodal models were chosen to cover a variety of important aspects of text models. Included in our tests are a large, well-cited and robust transformer-based model, a model with a similarly sized language encoder to the other multimodal models, and an alternative state of the art transformer model; in order, they are RoBERTa, RoBERTa-small, and t5 (Liu et al., 2019). The multimodal models, VisualBERT, LXMERT, UNITER, and CLIP, were chosen to represent different designs in multimodal training.

3.1.1 VisualBERT

The VisualBERT model works by projecting the text, through the embedding layer, and images, through a convolutional network, to the embedding space. These embeddings are then jointly passed through a BERT transformer which outputs their transformed embeddings. It is pretrained on two task agnostic objectives: learning to predict masked tokens in an image caption and deciding if a caption and image belong together

3.1.2 LXMERT

The LXMERT model has a similar architecture to VisualBERT. It projects images and text to the transformer embedding space then passes them through a transformer. One significant difference is that while they are initially passed through the transformer together such that the self attention is computed between image and text, the second part of the transformer performs the self attention on image and text separately. The model is pretrained on two task agnostic objectives: masked language modeling on image captions and image QA.

3.1.3 UNITER

UNITER has a similar model architecture to LXMERT and VisualBERT, but is trained with a few key differences. Unlike previous models, UNITER is pretrained with a conditional masking regime where one modality is masked while the other remains un-obscured (Chen et al., 2020). This could prove a key difference than the default randomized masking, which occurs on both modalities at the same time. It is possible that isolating one modality at a time proves crucial when the word embeddings are removed in experiment 2.

3.1.4 CLIP

Finally, the CLIP model uses a fundamentally different architecture than the early fusion models described above. CLIP separately projects the image and text through a visual transformer and a language transformer. It is pre-trained by contrastive learning, matching a batch of images and text together by their cosine similarity.

3.1.5 Baselines

In addition to the three language only models, a standard BERT model was initialized with random weights and used to train a logistic classifier. This serves to isolate the significance of the pretraining from the training step, as we are only interested in the performance differences between models pretrained on different kinds of data.

3.2 Datasets

The 3 datasets we explore are Verb Physics, Distribution Over Quantiles (DOQ), and Physical Reasoning about Objects through Space and Time (PROST).

Dataset	Train Size	Test Size	Ave. Tokenization Size
DOQ	1101	1101	7.1
Verb Physics	732	1828	7.1
PROST	2000	7500	62

Table 1: Dataset test/train size with average tokenization size of examples across models.

3.2.1 Verb

The Verb Physics dataset has tasks in which the model compares two objects and determines which of them has more of a specific physical characteristic. For example, “an elephant is larger than a house,” where the correct answer would be False. The Verb Physics dataset compares objects using the following five key attributes: weight, speed, size, strength, and rigidness.

3.2.2 DOQ

The DOQ dataset has a similar task with similar physical characteristics, however, combines three of its attributes into one task: weight, speed, and size. An example from the DOQ dataset is ”a kangaroo is heavier than a hummingbird”, which should be labelled as a true statement.

3.2.3 PROST

The PROST dataset again asks questions about common sense topics, but is framed as a multiple choice question between four sentences. Each text example is encoded as two sentences separated by a special token. The first is a declarative sentence, containing information about the world, which is different for each of the four different options. Only one of the four declarative sentences is correct, which corresponds to a True label. The second sentence is an action that explicitly uses the context. Importantly, while random chance for the first two datasets would return 50% accuracy, PROST is twice as hard, with only one-in-four answers corresponding to a True label. The attributes tested in the PROST dataset are directions, mass, height, circumference, stackable, rollable, graspable, breakable, slidable, and bounceable. An example input from the PROST dataset is ”the [MASK] is most likely to break. [SEP] a person drops a glass, a pillow, a coin, and a pen from a balcony.” where each of the four objects would be placed into the [MASK] token separately.

3.3 Experiments

Quantitative evaluation between unimodal and multimodals is inherently impeded by the difference in input data; some multimodal models require both image and text input. Our focus on the language portion of multimodal models can conflict with this input requirement. With only text data, we focus on the language portion of our models. For the language-only subset of the models (RoBERTa, T5), this takes the form of the entire model. For VisualBERT and CLIP, the language encoder can be completely detached and used on our text input. Finally, for UNITER and LXMERT, the only text-only portion of the model is the word embedding layer itself. This highlights the inextricably linked nature of visual and text data in a meshed transformer model, take UNITER for example, where language and visual data are encoded into a “common embedding space” (Chen et al., 2020).

Thus, we propose the following two experiments: using the complete model (for only the relevant subset of model), and using just the word embedding layer.

3.3.1 Complete Models

Our first experiment involves evaluating the complete models of those that accept exclusively text data as a valid input on our three datasets. As a baseline, both RoBERTa models are used, along with the multimodal models with separate encoding spaces for text: CLIP and VisualBERT. On each dataset, text is tokenized, padded with zeros from the right, and then passed through the model. To extract features from the model, the last hidden layer is used, which should represent the most distilled version of the data before any classification heads. Finally, this featureized data is fed into a logistic classifier, and then the logistic classifier is tested against reserved data.

An important note here is that because we are grabbing a hidden layer in the model, this shape is not consistent across models. For RoBERTa (both small and large) as well as CLIP and VisualBERT, the hidden layer size is 768. However, for T5 the hidden layer size is 512.

3.3.2 Word Embeddings

The second experiment is only slightly different in structure than the first, but allows comparison across all models by utilizing word embeddings. Through the process of training, each model develops a way of storing a vector representation for every given word in its allowed vocabulary. These can be thought of as word embeddings, and can be used a proxy for how the model thinks about each word. Extracting the word embedding layer from each model, all seven models can encode tokenized text to their own feature spaces, and be used to train and test a logistic classifier. Justification of this method follows in the results section, after the analysis of experiment 1 results.

4 Results

4.1 Quantitative Methods

First, we will discuss the results of the word embedding experiments found in table 4.1; notice the distribution of performance across the first two datasets. Immediately we can see that the results

Model	DOQ	Verb	PROST
RoBERTa-small	59.27	64.70	**61.95**
RoBERTa-large	**65.0**	**78.02**	31.96
T5	62.27	76.77	35.85
CLIP	63.27	76.68	30.89
LXMERT	64.45	77.16	31.59
VisualBERT	63.64	77.02	31.85
UNITER	62.64	72.58	32.34

Table 2: Experiment 1: Word-Embedding based featurization with text-only input.

on DOQ are extremely uniform, with only a 3.1 percentage point range in accuracy. That range is capped by RoBERTa on the top end, and its significantly smaller counterpart, RoBERTa-small, on the low end. Furthermore, while there is somewhat less uniformity for the Verb dataset, RoBERTa large is still the best performer. This shows that multimodal models did not significantly outperform unimodal models in this experiment.

However, not all of the metrics are as uniform. RoBERTa-small performance is consistently poor, significantly worse across all metrics other than PROST. The fact that RoBERTa-small is a huge over-performer on the PROST metric, more than a standard deviation above the mean, is grounds for further analysis. Does the smaller model force more semantic information to be stored in the word embeddings? It is possible that in RoBERTa-large, the significantly increased number of parameters decreases the reliance on the word embeddings, and thus performs worse than the smaller model in this experiment. To emphasize our results, the top two performers on PROST were unimodal trained models.

These results contradicts our earlier proposition that multimodal training might help the language model learn common sense reasoning.

4.2 Word Embedding Justification

Model	DOQ	Verb	PROST
RoBERTa-small	+0.26	+3.25	+20.12
RoBERTa-large	-1.0	+0.38	+58.95
T5	+0.0	-2.06	+51.65
CLIP	+3.45	-3.06	+60.78
VisualBERT	+0.17	+2.15	+58.11

Table 3: Difference in accuracy between Experiments 1 & 2

The word embedding method narrows the focus to only the learned semantic understandings of each word, providing the level of fidelity on the word-level required to investigate our question. We highlight the difference in performance between complete encoders and just the word embeddings, with RoBERTa-small and T5 as unimodal baselines, and with CLIP and VisualBERT as multimodal representatives in Table 3.

Interestingly, the increased parameters do not dramatically improve performance over the word embedding layer on the DOQ and Verb datasets. There are a few possible explanations for this. First, it is possible that there are some number of questions in the dataset that are much more difficult than the others. This would prove a soft upper bound on accuracy, and limit the differentiation between experimental results. A second explanation could be that the word-embedding layer simply encodes nearly all of the information required to compare objects.

One important limitation of using only word embeddings is that it relies entirely on the information gained during pretraining to make its way all the way back through the model to the very top layer. We have shown that this does, in fact, happen, through the small difference in performance between the word embedding and full language encoder models, despite a massive increase in feature space. However, one potential downside to this method is a decrease in uniqueness or expressiveness of the model. Evidence for this exists in the correlation in predicted output across all seven models, visualized in **Figure 1**. Here, the raw probability output from the logistic classifier on the test sets of all three datasets were used. The only obvious finding is that RoBERTa-small is lacking when compared to any of the other models, which is expected. While all six remaining models share a relatively high correlation, there is no obvious distinction between the language and the visual + language models. Had visual information proved a unique and vital source of semantic grounding, one might have expected the visual models to perform more similar to each other. However, the two models that answer most similarly are VisualBERT and RoBERTa, representing the top performers of both architectures.

Now, we discuss a theoretical explanation for why word embeddings might contain the common sense meaning of a word. If a model does learn common sense reasoning, we would expect that it would be propagated back through the network towards the word embeddings. For the sake of contradiction, lets assume that the model does learn information about physical reasoning yet this information is not contained in the word embeddings. This would require that the interaction between word embeddings, computed with several self attention layers, would have to generate the common sense information. Take the interaction between two objects, glass and concrete. We are currently assuming that the model has zero knowledge of the rigidness of either, yet when connected by two layers, it possesses the ability identify which might be more likely to break. This seems intuitively suspect, and indeed it is not possible to generate entirely new knowledge by the interaction of variables that did not somewhat contain that knowledge when combined. We have reached a contradiction, so either the model does not learn common sense information or it is in part lies within the word embeddings.

The exception to this observation is the PROST dataset. Clip, RoBERTa-small, and VisualBERT all perform demonstrably better in the encoder feature space, with an average improvement across models of 34 percentage points better. We suspect that PROST significantly improves with the full transformer language model compared to just the word embeddings because the knowledge required for this dataset requires the computation of more word interactions than a simple linear probe can compute. While the other two datasets ask questions in the form of comparing two physical objects, PROST requires that the model compare four objects. Further more this dataset asks questions that are very dependent on context. As seen in the example of dropping an object from a balcony that context is essential in answering the question. For both of these reasons, the linear probe simply does not have the representational capacity to represent all these interactions while several self attention layers do. If we look to Table 1, we see that for PROST, the average tokenized shape is 62. However, CLIP, RoBERTa and VisualBERT models all have more than 10 times as many features in the last hidden layer, allowing for dramatically more information to be outputted by the model.

Model	DOQ	Verb	PROST
Random LogReg	54.88	52.18	26.61
RoBERTa-small	59.55	67.95	82.07
RoBERTa-large	64.0	78.40	90.91
T5	62.27	74.71	87.5
CLIP	**66.72**	73.62	**91.67**
VisualBERT	62.81	**79.17**	89.96

Table 4: Experiment 2: Accuracy for Full-Model featurization with text-only input.

While using word embedding with PROST might not be the most representative method, our results from DOQ and Verb support our earlier rejection of our initial hypothesis. However, looking at quantitative metrics across a suite of models and data only provides evidence for a surface-level comparison between types of models.

4.3 Qualitative Methods

The above qualitative results reinforce the question, *are there any substantive differences between questions that the language models get correct, and the visual + language*? To answer this, we dive into a qualitative analysis, identifying the most frequent missed questions, and attempt to map them to a subject matter domain. There are two possible implications to patterns being found during error analysis across our seven models; either differing training data or differing model architecture leads to better accuracy.

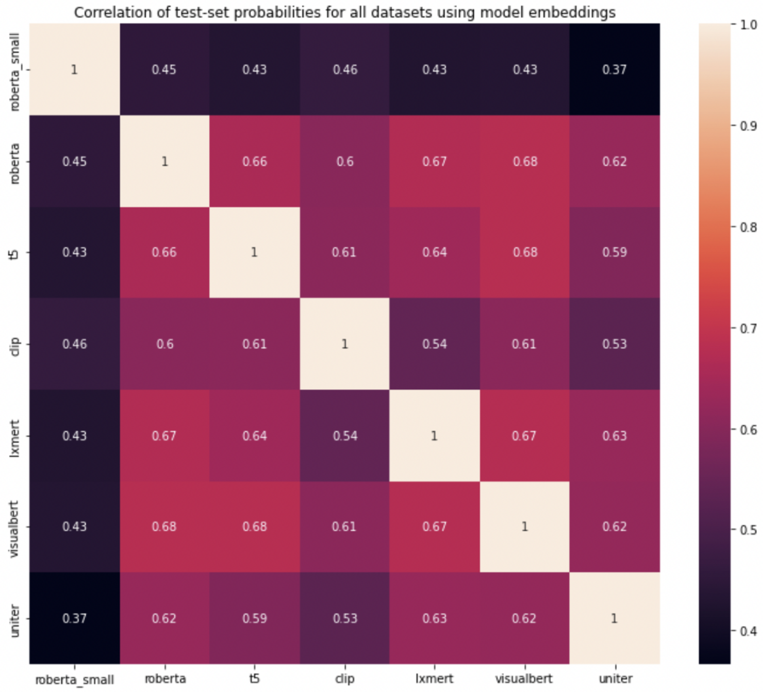

Figure 1: Covariance between model word-embedding accuracy.

Sentence from DOQ	**Label**
A scissors is heavier than a hand	False
A shoulder is heavier than a door	False
A messenger is heavier than a king	Ambigious
A person is heavier than a brother	Ambigious
A person is heavier than a coach	Ambigious
A throat is heavier than a hand	False
A coach is heavier than a friend	Ambigious
A person is heavier than a lady	Ambigious
A body is heavier than a friend	Ambigious
A town is heavier than a store	True

Table 5: All sentences that no multimodal model is able to correctly classify in the Verb-WEIGHT dataset, that are also answered correctly by RoBERTa.

Just analyzing the word embedding results, so that all models can be included, we find that one of the key advantages RoBERTa has over the multimodal models is its sensitivity to ambiguous answers. To illustrate this, consider the 10 sentences in the Verb dataset in the weight category that no multimodal model is able to answer, but RoBERTa gets correct **Table 5**. While ”Ambigious” as a label is present in less than 5% of Verb-weight, it makes up 6/10 of the sentences that RoBERTa can answer that multimodal models cant.

Another pertinent observation relates to the subject matter in these questions. The 6 questions that are ”Ambigious” all relate to the human body, with the following words: ”messenger, king, person, brother, coach, friend, person, and lady.” None of these words are especially obscure, and can all reasonably be assumed to appear in visual + language datasets that the multimodal models were trained on. However, disentangling the relationship between these objects is non-trivial, as they all refer to the weight of the same fundamental unit, a person. Perhaps, the robust language-processing unit of RoBERTa is able to understand these relationships more efficiently. On the other hand, it is also possible that the billions more words seen during pre-training allow better connections between like-concepts, and thus RoBERTa can identify the ambiguous case more accurately.

5 Conclusion

Prior research into multimodal models has primarily focused on the importance of textual data on the visual encoder, which appears to be a more rewarding direction to explore. By focusing on the domain of common sense reasoning, we believed that the inclusion of visual data during pretraining would provide improved semantic understanding of physical objects. While initial results from running text data through large multimodal and unimodal models showed promise, comparing across 7 models showed no significant improvement. It is possible that the reliance on word embeddings hinders visual + language models more than language only models, but the evidence as to why that might be is elusive.

One significant finding is how much semantic physical reasoning information is stored in the word embeddings of large language models. It is not immediately clear why performance of RoBERTa-large, when the full model is used on the DOQ dataset, actually decreases. This confounding result highlights the importance of further research into where exactly semantic understanding of commonsense physical reasoning is being encoded across different model architectures.

REFERENCES

Stéphane Aroca-Ouellette, Cory Paik, Alessandro Roncone, and Katharina Kann. Prost: Physical reasoning of objects through space and time, 2021.

Emily M. Bender and Alexander Koller. Climbing towards NLU: On meaning, form, and understanding in the age of data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 5185–5198, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.463. URL `https://aclanthology.org/2020.acl-main.463`.

Michele Cafagna, Kees van Deemter, and Albert Gatt. What vision-language models 'see' when they see scenes, 2021.

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Universal image-text representation learning, 2020.

Oier Lopez de Lacalle, Ander Salaberria, Aitor Soroa, Gorka Azkune, and Eneko Agirre. Evaluating multimodal representations on visual semantic textual similarity, 2020.

Joshua Feldman, Joe Davison, and Alexander M. Rush. Commonsense knowledge mining from pretrained models, 2019.

Maxwell Forbes and Yejin Choi. Verb physics: Relative physical knowledge of actions and objects, 2017.

Ioanna Giorgi, Angelo Cangelosi, and Giovanni Masala. Learning actions from natural language instructions using an on-world embodied cognitive architecture. *Frontiers in Neurorobotics*, 15, 05 2021. doi: 10.3389/fnbot.2021.626380.

Stanisław Jastrzebski, Dzmitry Bahdanau, Seyedarian Hosseini, Michael Noukhovitch, Yoshua Bengio, and Jackie Chi Kit Cheung. Commonsense mining as knowledge base completion? a study on the impact of novelty, 2018.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V. Le, Yunhsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision, 2021.

Armand Joulin, Laurens van der Maaten, Allan Jabri, and Nicolas Vasilache. Learning visual features from large weakly supervised data, 2015.

Xiang Lorraine Li, Adhi Kuncoro, Cyprien de Masson d'Autume, Phil Blunsom, and Aida Nematzadeh. A systematic investigation of commonsense understanding in large language models, 2021.

Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, Yejin Choi, and Jianfeng Gao. Oscar: Object-semantics aligned pretraining for vision-language tasks, 2020.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692, 2019. URL `http://arxiv.org/abs/1907.11692`.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks, 2019.

Sandro Pezzelle, Claudio Greco, Greta Gandolfi, Eleonora Gualdoni, and Raffaella Bernardi. Be Different to Be Better! A Benchmark to Leverage the Complementarity of Language and Vision. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 2751–2767, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.findings-emnlp.248. URL `https://aclanthology.org/2020.findings-emnlp.248`.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021.

Maarten Sap, Vered Shwartz, Antoine Bosselut, Yejin Choi, and Dan Roth. Commonsense reasoning for natural language processing. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts*, pp. 27–33, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-tutorials.7. URL `https://aclanthology.org/2020.acl-tutorials.7`.

Shikhar Singh, Nuan Wen, Yu Hou, Pegah Alipoormolabashi, Te-Lin Wu, Xuezhe Ma, and Nanyun Peng. Com2sense: A commonsense reasoning benchmark with complementary sentences, 2021.

Jinsong Su, Jinchang Chen, Hui Jiang, Chulun Zhou, Huan Lin, Yubin Ge, Qingqiang Wu, and Yongxuan Lai. Multi-modal neural machine translation with deep semantic interactions. *Information Sciences*, 554:47–60, 2021. ISSN 0020-0255. doi: https://doi.org/10.1016/j.ins.2020.11.024. URL `https://www.sciencedirect.com/science/article/pii/S0020025520311105`.

Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers, 2019a.

Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers, 2019b.

Deepak Ramachandran Tania Bedrax-Weiss Dan Roth Yanai Elazar, Abhijit Mahabal. How large are lions? inducing distributions over quantitative attributes. In *Proceedings of ACL 2019*, 2019.

Tian Yun, Chen Sun, and Ellie Pavlick. Does vision-and-language pretraining improve lexical grounding?, 2021.

Rowan Zellers, Ari Holtzman, Matthew Peters, Roozbeh Mottaghi, Aniruddha Kembhavi, Ali Farhadi, and Yejin Choi. Piglet: Language grounding through neuro-symbolic interaction in a 3d world, 2021.

Xikun Zhang, Deepak Ramachandran, Ian Tenney, Yanai Elazar, and Dan Roth. Do language embeddings capture scales?, 2020a.

Yuhao Zhang, Hang Jiang, Yasuhide Miura, Christopher D. Manning, and Curtis P. Langlotz. Contrastive learning of medical visual representations from paired images and text, 2020b.

Mingyang Zhou, Luowei Zhou, Shuohang Wang, Yu Cheng, Linjie Li, Zhou Yu, and Jingjing Liu. Uc2: Universal cross-lingual cross-modal vision-and-language pre-training, 2021.