Investigating Negation in Pre-trained Vision-and-language Models

Radina Dobreva and **Frank Keller**
Institute for Language, Cognition and Computation
School of Informatics
University of Edinburgh
r.dobreva@ed.ac.uk, keller@inf.ed.ac.uk

Abstract

Pre-trained vision-and-language models have achieved impressive results on a variety of tasks, including ones that require complex reasoning beyond object recognition. However, little is known about how they achieve these results or what their limitations are. In this paper, we focus on a particular linguistic capability, namely the understanding of negation. We borrow techniques from the analysis of language models to investigate the ability of pre-trained vision-and-language models to handle negation. We find that these models severely underperform in the presence of negation.

1 Introduction

Vision-and-language models have made a lot of progress on complex tasks, going beyond recognition and towards reasoning over the two modalities (Zellers et al., 2019; Suhr et al., 2019). Following the success of pre-trained language models such as BERT (Devlin et al., 2019) on a range of language tasks, recent advances in vision-and-language have involved the introduction of pre-trained models (e.g., UNITER, Chen et al. 2020, VisualBERT, Li et al. 2019, ViLBERT, Lu et al. 2019, LXMERT, Tan and Bansal 2019). These models achieve impressive results, topping task leaderboards and improving over previous approaches by a large margin. However, as with pre-trained language models, it is not clear how and why these models perform as well as they do, what information they learn and use in their predictions, or what their limitations are. While a large body of research has focused on the interpretation of pre-trained language models (e.g., Clark et al. 2019; Tenney et al. 2019; Rogers et al. 2020; Elazar et al. 2021), such work has been more limited for vision-and-language models.

This paper focuses on a particular linguistic capability, namely the ability to understand negation. Negation is universal across languages (Zeijlstra, 2007) and is very important for interpreting the meaning and determining the truth value of a statement. Vision-and-language models have applications in real-world scenarios where it is crucial to understand negation, as this is a behaviour expected by humans interacting with them.

Figure 1: Examples from the NLVR2 corpus. The input to the model consists of two images and a sentence. The model's task is to predict whether the sentence is true of the images or not. The top example shows an instance where the sentence is true and the bottom one, false. Image from (Suhr et al., 2019).

For the purposes of this analysis, we focus on a particular vision-and-language task, Natural Language Visual Reasoning for Real (NLVR2) (Suhr et al., 2019). For this task, the model is presented with two images and a sentence and has to predict whether that sentence is true of the images or not (see Figure 1). Since this task involves the prediction of a truth value, it lends itself well to the exploration of negation. We consider two popular pre-trained vision-and-language models, finetuned for the task: LXMERT (Tan and Bansal, 2019) and UNITER (Chen et al., 2020). Details on the data and models are in Section 3.

In order to investigate the performance of these

models on examples that contain negation in a controlled manner, we annotated a portion of the original NLVR2 test set to create minimally differing instances. That is, we created pairs of instances where the images remain the same and the sentences only differ in the presence or absence of negation (see Section 3 for details). We show that both models under consideration perform worse on the negated examples, compared to the corresponding non-negated examples (Section 4). We also use causal mediation analysis (Pearl, 2001; Vig et al., 2020) on UNITER to examine the contributions of specific neurons to the final predictions (Section 5). Our results show that the effects of negation are mainly seen in the upper layers of the model. We release the new test set and code for our experiments at https://github.com/radidd/vision-and-language-negation.

2 Background

2.1 Negation

Negation is related to the notion of polarity: a clause such as "It is raining" is said to have positive polarity, whereas a clause such as "It is not raining" has negative polarity (Pullum and Huddleston, 2002). Positive polarity is usually structurally simpler, whereas negative polarity is marked by words or affixes. Pullum and Huddleston (2002) categorise negation based on several different properties, but for our purposes we focus on the distinction between verbal and non-verbal negation. In the former, the negation marker is attached to the verb ("I did not see anything at all"), while in the latter, the negation marker is attached to a dependent of the verb ("I saw nothing at all").

Previous analysis of language models has shown that they are not able to handle negation. Pre-trained language models fail to make correct predictions in the presence of negation or even to distinguish between positive and negative sentences (Ettinger, 2020; Kassner and Schütze, 2020). Hossain et al. (2020) finetune several different pre-trained language models for natural language inference and show that performance on instances containing negation deteriorates.

2.2 Vision-and-language models

Following the success of pre-trained language models (Devlin et al., 2019), multimodal tasks such as visual question answering (Antol et al., 2015) and visual commonsense reasoning (Zellers et al., 2019) have also recently been approached using a pretrain-and-finetune method. Pre-trained vision-and-language models fall into two categories: single-stream models (e.g., Li et al. 2019; Chen et al. 2020) and two-stream models (e.g., Lu et al. 2019; Tan and Bansal 2019). While two-stream models have separate encoders for the visual and textual modalities and then jointly process them using a third encoder, single-stream models process both the textual and visual features together from the start.

While these models have been very successful, pushing state-of-the-art results up across tasks, not much is known about their capabilities and limitations. From that perspective, the closest works to ours are by Li et al. (2020) and Cao et al. (2020). Li et al. (2020) analyse attention heads in the model and show that words in the text are correctly mapped to image regions which correspond to them. Cao et al. (2020) probe pre-trained vision-and-language models and report similar observations. Their work focuses more on the grounding aspect of vision-and-language models, whereas our work focuses more on linguistic and reasoning abilities. Cao et al. (2020) also show that multimodal pre-trained models learn some linguistic knowledge, however, they do not analyse this any deeper than providing results on several probing tasks, none of which involves negation.

2.3 Causal mediation analysis

Language-only BERT has been targeted by a lot of recent analysis work focusing on different capabilities of the model (Clark et al., 2019; Tenney et al., 2019; Rogers et al., 2020; Elazar et al., 2021). A recent proposal is the application of causal mediation analysis (Pearl, 2001) to better understand neural NLP models (Vig et al., 2020).

Causal mediation analysis aims to measure the effect of intermediate variables ("mediators") on a response variable. Pearl (2001) defines a natural direct effect (NDE) and a natural indirect effect (NIE), see Figure 2. NDE refers to the direct effect of a particular value of an input variable X on the value of a response variable Y, without the intervention of a mediator Z. More specifically, we can measure the effect of an intervention, or change, of the input variable ($X = x \rightarrow X = x*$), while keeping Z to its value without the intervention. NIE refers to the indirect effect of the input variable X on Y via the intermediate variable Z.

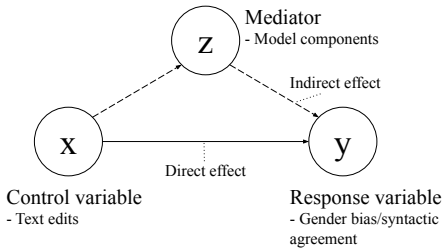


Figure 2: Graphical model of the indirect and direct effect. Figure adapted from Vig et al. (2020).

Specifically, we can fix X to its original value, but change the value of the intermediate variable Z to its value under the intervention $X = x*$.

Vig et al. (2020) apply causal mediation analysis to the study of gender bias in large pre-trained language models. They treat model components – specifically neurons and attention heads – as intermediate variables (Figure 2). They make changes to the input text by switching from gender-ambiguous to gender-unambiguous input and measure the effects of these changes on the amount of bias the model exhibits. This kind of analysis can show whether specific model components are causally responsible for a specific outcome. Indeed, Vig et al. (2020) show that gender bias effects are sparse and concentrated in a handful of attention heads in the middle layers of the model. More recently, Finlayson et al. (2021) apply causal mediation analysis to the problem of subject-verb agreement in language models. They look at neurons in the models and find that some models learn two different mechanisms to resolve agreement for different sentence structures.

3 Data and models

3.1 Models

For our experiments we used two vision-and-language models, both based on the Transformer architecture: LXMERT (Tan and Bansal, 2019) which is a two-stream model and UNITER (Chen et al., 2020), which is a single-stream model. For UNITER, we experiment with two different variants of applying the model to NLVR2: paired with attention and triplet. The paired with attention variant encodes the two images separately and then combines the representations with an additional attention layer, whereas the triplet variant encodes both images together from the start (see Chen et al. (2020) for details).

3.2 NLVR2 dataset

The dataset used for all experiments is NLVR2 (Suhr et al., 2019), which consists of pairs of images, a sentence describing each pair and a True/False tag indicating whether the sentence is true of the image pair. This dataset requires joint reasoning over the two modalities and is more complex than image captioning datasets due to the kind of language it contains – for instance, it requires comparison and counting abilities.

The authors of NLVR2 present statistics of the occurrence of different linguistic phenomena in a portion of their development set. Their analysis shows that 9.6% of the samples they consider contain negation. This statistic is not available for the test and the training set, so we calculated it based on a short list of negation words (see Appendix A). The results are as follows:

- Training set: 7192/86373 samples (8.33%)
- Development set: 630/6982 samples (9.02%)
- Test set: 589/6967 samples (8.45%)

This shows that negation is not very common in the dataset. As a preliminary experiment, we tested the models' performance on the samples identified to contain negation in the original development and test set and compared it to the performance on samples which do not contain negation. Results are shown in Table 1. All three models show a drop in performance on the samples containing negation for both the development and test set, compared to non-negated samples. The drop in performance varies between 1.7 points and 7 points.

There is no more detailed analysis of the types of negation present in the dataset, for example whether it is verbal or non-verbal. This means that there is no way to use the existing data for a more fine-grained analysis of negation. The existing dataset also cannot be used reliably to make performance comparisons between negated and non-negated examples. This is because it is possible that other factors, such as sentence length or complexity of the reasoning required (e.g., counting, comparison between the two images), are influencing performance. Therefore, we manually created a test set of minimally differing pairs by adding negation to the original data.

3.3 Negation test set

In order to investigate whether vision-and-language models perform differently in the presence of negation, we annotated a portion of the NLVR2 test set

	LXMERT		UNITER$_{paired-attn}$		UNITER$_{triplet}$	
	neg.	non-neg.	neg.	non-neg.	neg.	non-neg.
Dev set	71.43	74.92	74.29	77.38	70.00	71.68
Test set	67.74	74.71	74.87	77.89	67.57	73.61

Table 1: Accuracy on samples which contain negation and samples which do not contain negation from the original development and test set.

Negation type	o/n	Example	count
Verbal (content)	o	At least one person is wearing a hat.	91
	n	At least one person is not wearing a hat.	94
Verbal (existential)	o	The left image contains exactly two dogs.	108
	n	The left image does not contain exactly two dogs.	108
NP (nonexistential)	o	All the cars are facing towards the left.	28
	n	Not all the cars are facing towards the left.	29
NP (existential)	o	All the marmots are on rocks.	55
	n	None of the marmots are on rocks.	55
NP (number-to-none)	o	There are a total of four people in the gym.	72
	n	There are no people in the gym.	72
Sentence-wide	o	there are sled dogs moving toward the camera	83
	n	It is not true that there are sled dogs moving toward the camera	83

Table 2: Example negated sentences and counts for each category. "o" stands for original, "n" stands for negated.

to create instances of negation. An important requirement for the annotation was that every negated instance has a flipped label compared to the original.[1] It should also be noted that while negation is a complex phenomenon, here we only consider absolute negators (e.g., *no, not, nobody, nothing*), and we do not consider approximate negators (e.g., *few, little, barely*) or affixal negators (e.g., the prefixes *un-, in-, non-*, see Pullum and Huddleston (2002). The negation categories below are constrained by these two considerations.

First, we identified appropriate samples for negation. We selected the samples in pairs, where both samples belonging to a pair have the same images, but different sentences and True/False labels. This was done so that the labels remain balanced in the new negation test set. See Appendix B for more criteria for sample selection.

Next, we created negated versions of each sample in such a way that the new sample has a flipped label according to the annotator's judgement. It is important to keep in mind that the truth value of an example depends on the images *and* the sentence; if we just look at the sentence on its own, adding negation does not flip the truth value in all cases. The annotator also recorded the type of negation for each example as belonging to one of six categories (see Table 2 and Appendix C for examples):

Verbal negation:

- Content negation: where negation is attached to a verb expressing an action ("is not standing") or a characteristic ("don't have black seats").
- Existential negation: where the negated verb relates to the existence of a predicate ("the image doesn't include . . . ", "there aren't . . . ").

Noun phrase negation:

- Non-existential: the resulting negated NP does not deny the existence of an object, for example "not all birds", "no more than three birds".
- Existential negation: denies the existence of an object in a noun phrase ("no dogs", "neither image").
- Number-to-none negation: similar to existential negation, but the original NP contains a numeral ("at least two dogs", "a total of five birds") and the negated version denies the existence of the object ("no dogs", "no birds").

[1] The flipped label requirement was in place to aid the analysis. It is less straightforward to evaluate pairs with the same label, since we would not be able to tell if the model is doing anything differently with and without the negation. Future work could include analysis of examples where negation does not change the label.

Sentence-wide negation: negating the full sentence by appending "It is not the case that ..." or "It is not true that ..." at the start.

The annotation was done with minimal possible perturbations of the original sentences to allow for a fairer comparison between the original and the negated samples. For some of the original samples it was possible to create several different negated ones. Those were kept to a maximum of three and in practice few samples had more than two possible negations. Table 2 shows the number of samples belonging to each category as well as the number of corresponding original samples. In total there are 441 negated samples created from 400 original samples.

It should be noted that there are some differences between the existing negated examples in the original NLVR2 test set and the examples in our negation test set. Firstly, we had to keep the negated sentences grammatical and minimally different, which restricted the variety of negation structures and types. Second, people likely use negation differently when they are describing images in a natural setting, than when it is added to existing descriptions artificially. See Appendix D for a further discussion of the differences between our test set and negation present in the original NLVR2 test set.

4 Experimental results

4.1 Experimental setup

We followed the instructions in the respective repositories for LXMERT[2] and UNITER[3] to obtain the pre-trained checkpoints and finetune them for the NLVR2 task. For both versions of UNITER we used the UNITER-base pre-trained checkpoint. LXMERT was finetuned for four epochs and achieved an accuracy of 74.61% and 74.12% on the development and test sets respectively. UNITER$_{paired\text{-}attn}$ was finetuned for seven epochs and achieved an accuracy of 77.10% and 77.64% on the development and test sets. UNITER$_{triplet}$ was finetuned for 10 epochs and achieved accuracies of 71.53% on the development and 73.10% on the test set.

4.2 Results on the negation test set

Table 3 shows the results on the negation test set by category. For comparison, we have provided accuracy on the original samples corresponding to the negated samples in each category. As can be seen from the table, all models perform worse on the negation samples, across all categories. The only exception to this is UNITER$_{triplet}$ on the NP (nonexistential) category, where the score for the negative samples is higher than that for the corresponding positive ones. Between the three models, overall the performance of UNITER$_{paired\text{-}attn}$ is highest, followed by UNITER$_{triplet}$ and LXMERT.

Looking at specific categories, LXMERT seems to struggle the most with verbal negation, while both versions of UNITER perform better, achieving scores between 14 and 20 points higher than LXMERT. The scores for the three NP negation types are more varied between models, with UNITER$_{triplet}$ outperforming the others on the number-to-none and nonexistential categories and UNITER$_{paired\text{-}attn}$ outperforming on the existential category. Since the negated forms of the NP (number-to-none) category and the NP (existential) category are very similar, it is surprising that there is such a big gap in performance (around 20 points) for two of the models. All models struggle significantly with sentence-wide negation. One explanation for this could be that the models do not encounter the phrases used to create these samples in the training data and ignore them.

Table 4 shows accuracy of the negated samples, split by whether the model predicts the corresponding original sample correctly or not. A higher score on the originally correct examples indicates that the model is potentially able to handle negation, i.e., it learns that negation inverts the truth value. A higher score on the originally incorrect examples is less clear – since there are only two categories, the model can achieve this by simply ignoring negation and outputting the same prediction as for the original. Looking at the originally correct category, UNITER$_{paired\text{-}attn}$ outperforms both other models across categories, except for the sentence-wide category. For the sentence-wide category, LXMERT outperforms both versions of UNITER by more than 16 points. Turning to the originally incorrect category, here all models perform much better across negation categories, with the exception of NP (existential) negation. The highest score overall is obtained by UNITER$_{triplet}$, followed by LXMERT and UNITER$_{paired\text{-}attn}$.

[2] https://github.com/airsplay/lxmert
[3] https://github.com/ChenRocks/UNITER

	LXMERT		UNITER$_{paired-attn}$		UNITER$_{triplet}$	
	negative	positive	negative	positive	negative	positive
Verbal (content)	28.72	69.23	43.62	73.63	43.62	71.43
Verbal (existential)	30.56	82.41	50.0	77.77	44.44	66.66
NP (nonexistential)	44.83	67.86	48.28	64.29	55.17	50.0
NP (existential)	34.55	80.0	50.91	85.45	32.73	87.27
NP (number-to-none)	54.17	72.22	51.39	77.77	55.56	76.39
Sentence-wide	38.55	66.27	31.33	69.87	38.55	65.06
Overall	36.96	73.5	45.35	76.5	44.22	71.5

Table 3: Accuracy on the negation test set and the corresponding non-negated (positive) examples.

	LXMERT		UNITER$_{paired-attn}$		UNITER$_{triplet}$	
	o. correct	o. incorrect	o. correct	o. incorrect	o. correct	o. incorrect
Verbal (content)	15.38	58.62	40.58	52.0	30.3	75.0
Verbal (existential)	21.35	73.68	46.43	62.5	29.17	75.0
NP (nonexistential)	30.0	77.78	42.11	60.0	40.0	71.43
NP (existential)	36.36	27.27	55.32	25.0	27.08	71.43
NP (number-to-none)	46.15	75.0	51.79	50.0	47.27	82.35
Sentence-wide	25.45	64.29	8.62	84.0	9.26	93.1
Overall	27.38	63.79	40.54	60.19	29.35	79.39

Table 4: Accuracy for the negated examples for whose original (unnegated) version the model makes a correct/incorrect prediction ("o. correct"/"o. incorrect").

5 Causal mediation analysis

We will now take a closer look at the effect of adding negation and the contributions of specific neurons to model predictions using causal mediation analysis. This type of analysis can help us measure the effect of a change in the input (adding negation) on the output beyond looking at accuracy only. We also apply mediation analysis to find out if specific neurons in the model contribute to a change in the output more than others. We continue to work with the negation categories described previously and aim to discover if the results from this analysis correspond to the observed differences in accuracy between the categories and whether neuron effects differ by category.

Similarly to Finlayson et al. (2021), we are working with a task which has correct and incorrect outputs (in their case continuations, in ours, labels). We therefore follow their definitions. We want to measure the ability of the model to predict the correct True/False label given the images and the sentence. Therefore, we define our response variable as:

$$y(u, l) = \frac{p_\theta(l_{incorrect}|u)}{p_\theta(l_{correct}|u)}$$

where u are the images and the text, which can either match (u_{true}) or not (u_{false}), and l is the label. This equation takes the following form depending on whether the gold label is True or False:

$$y(u_{true}, l) = \frac{p_\theta(l_{false}|u_{true})}{p_\theta(l_{true}|u_{true})}$$

$$y(u_{false}, l) = \frac{p_\theta(l_{true}|u_{false})}{p_\theta(l_{false}|u_{false})}$$

The value for y is small ($y < 1$) when the model assigns the correct label, and large ($y > 1$) when the model assigns the incorrect label.

Following Vig et al. (2020) and Finlayson et al. (2021), we define two *do*-operations: (a) `negate`: negate the original sentence so that the truth value changes, and (b) `null`: no change. We also define $y_x(u, l)$ to be the value of y when we apply the operation x to the context u. This takes the following values under the `negate` operation for each of the possible values of u:

$$y_{\texttt{negate}}(u_{true}, l) = \frac{p_\theta(l_{false}|u_{false})}{p_\theta(l_{true}|u_{false})}$$

$$y_{\texttt{negate}}(u_{false}, l) = \frac{p_\theta(l_{true}|u_{true})}{p_\theta(l_{false}|u_{true})}$$

In order to measure the change in the response variable under the intervention (negation), we define the unit-level (per one original/negated pair)

total effect as:

$$TE(\texttt{negate}, \texttt{null}; y, u, l) = \frac{y_{\texttt{negate}}(u,l) - y_{\texttt{null}}(u,l)}{y_{\texttt{null}}(u,l)} = \frac{y_{\texttt{negate}}(u,l)}{y_{\texttt{null}}(u,l)} - 1$$

As Finlayson et al. (2021) state, this quantity measures the margin between the probability of the correct and incorrect answers under an intervention. However, unlike Finlayson et al. (2021), we analyse the originally correct and originally incorrect examples separately. When the model predicts the original example correctly, a larger total effect under the intervention could indicate a better handling of negation, as it suggests a higher probability of the correct label under negation. When the original example is predicted incorrectly, it is less clear what the total effect indicates. While a larger total effect suggests a move towards the correct prediction under negation, this does not necessarily mean negation is handled better. A smaller total effect suggests a move toward the incorrect label, but this means flipping the label, which is desired behaviour.

We calculate the average total effect across all example pairs, for each negation type and for two different sizes of UNITER:

$$\overline{TE}(\texttt{negate}, \texttt{null}; y) = \mathbb{E}_{u,l}\left[\frac{y_{\texttt{negate}}(u,l)}{y_{\texttt{null}}(u,l)} - 1\right]$$

For the mediation analysis, we focus on the effects of individual neurons from the representation of the [CLS] token on the response variable y. In the following definition, z refers to a single neuron. We measure the natural indirect effect (NIE) of a change in the input X on the response variable y, with respect to a mediator z. As mentioned in Section 2.3, this is done by fixing the input to its value without the intervention, but changing the value of the mediator z to its value under the intervention. In this case, we set the input to its value without negation (i.e. the original images-sentence pair which does not contain negation), but set the value of z (the neuron) to the value it would take if the input was the negated version of the images-sentence pair. The population level NIE then is defined as:

$$\overline{NIE}(\texttt{negate}, \texttt{null}; y, z) = \mathbb{E}_{u,l}\left[\frac{y_{\texttt{null},z_{\texttt{negate}}(u,l)}(u,l)}{y_{\texttt{null}}(u,l)} - 1\right]$$

As we are concerned with the effects of specific mediators (the neurons) on the output, we do not calculate the natural direct effect (NDE) which measures the effect of the input without the intervention of a mediator. We also do not consider intervening on attention heads in this work, as the length of the original and the negated sentences is different.

5.1 Results

For this analysis we use the triplet version of UNITER (Chen et al., 2020). We perform the analysis on the base and large versions of the model (see Appendix E for accuracy results of the large model). We use UNITER-base and UNITER-large to refer to the two sizes of the triplet model.

Total effects We report the total effect for each negation type for both model sizes, separately for the originally correct and originally incorrect examples (Figure 3). Looking at the originally correct examples (left side of Figure 3), we observe that the total effects of UNITER-base are in most cases several orders of magnitude larger than those of UNITER-large. This difference between models could mean that the two models assign probabilities differently, i.e., the base model assigns more extreme probabilities to the correct answers, while the large model assigns more moderate probabilities. Both models show similar patterns, with the lowest total effects being observed for the sentence-wide and the verbal (content) negation categories and the highest for the NP (nonexistential) and verbal (existential) categories. This pattern does not show any apparent correlation with the accuracies reported in the previous section – a higher total effect does not necessarily mean a higher accuracy. It is possible that for categories with higher accuracy but comparatively lower total effects the change in probability is moderate, but meaningful – enough to flip the label. At the same time, it could be that the categories with a high total effect contain examples with extreme probabilities which skew the average.

The right part of the figure shows the total effects of examples for which the model incorrectly predicts the original. A larger total effect indicates a

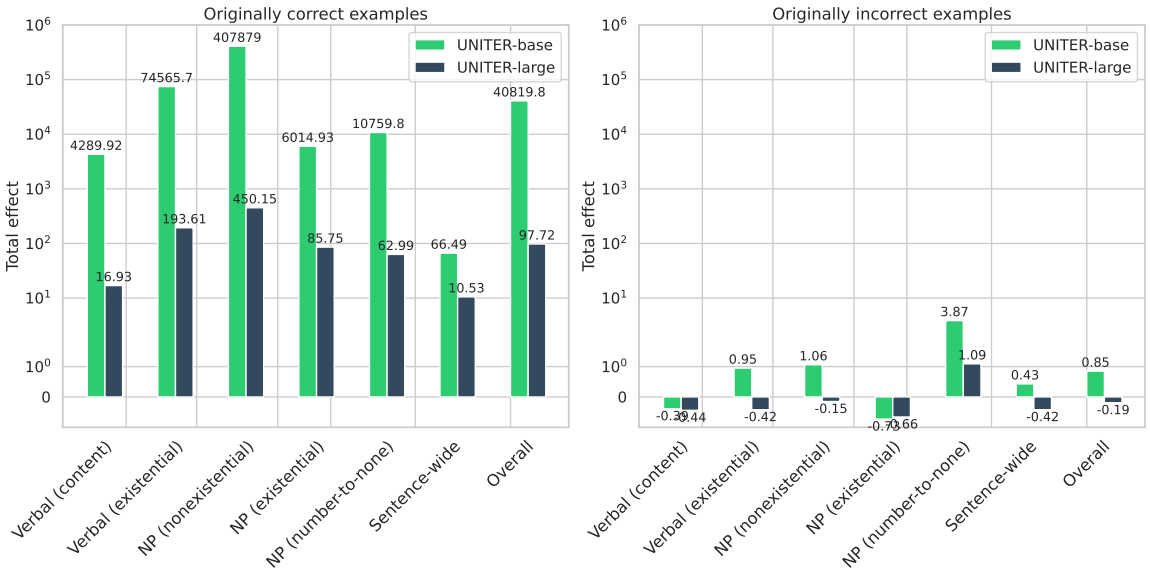


Figure 3: Total effects by correctness of the original (non-negated) example.

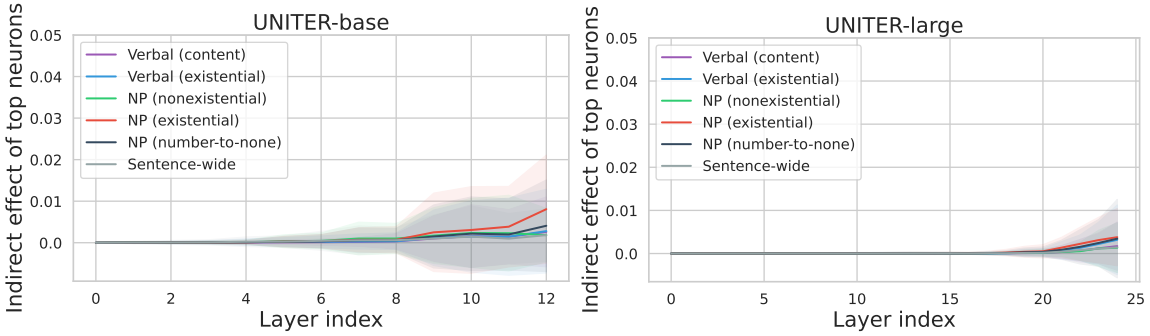


Figure 4: Natural indirect effect of the top 5% of neurons in each layer per negation category. Shaded area represents the standard deviation.

larger change in the probability moving towards the correct prediction for the negated sample, whereas a smaller (negative) total effect indicates a larger change moving in the direction of the incorrect prediction for the negated sample. These changes are more difficult to interpret for the originally incorrect samples. A larger total effect could indicate that the model is handling negation, however, it is not clear why that would be the case if the original sample was incorrectly predicted. A negative total effect indicates that the prediction probability is moving towards the opposite label from the one predicted for the original which could also be interpreted as the model handling the negation, since a change of label is expected. All total effects are smaller than the ones observed for the originally correct examples, which is expected given the observed high accuracies – the model prediction does not change a lot under the intervention and the total effect is closer to zero.

Natural indirect effects Figure 4 shows the NIEs of the top 5% of neurons with the highest NIE in each layer, for each negation category. We observe that for both models, the NIEs are approximately zero in the lower layers and become larger in the upper layers. The NIEs are similar between negation types, however, for both models we observe the highest NIEs for the NP (existential) type. We do not see any differences in patterns based on the negation category, which suggests we do not have evidence for a different treatment of the different categories by the model. We also compared the NIEs of examples split by whether the original/negated one is correctly predicted by the model, however, we do not observe notable differences (Appendix F).

Previous studies on language models have shown

that syntactic information is stored in the middle layers, task-specific information is stored in the upper layers and there are conflicting conclusions regarding semantic information being found in the upper layers, or throughout the model (see Rogers et al. (2020) for an overview). It is unclear whether these patterns hold for vision-and-language models. However, if they do, our results show that the models do not process negation on a syntactic level – at least not in terms of the neurons. The observed effects in the upper layers suggest that negation may be semantically processed in some way. Further investigation is required to draw more definite conclusions.

6 Conclusion

Our work shows that pre-trained vision-and-language models find it difficult to handle negation, which is a finding that is consistent with previous work on pre-trained language models. Using a manually created set of minimally differing pairs we show that two vision-and-language models (LXMERT and UNITER) fail to reach good performance in the presence of negation. We conduct causal mediation analysis on the neurons of one model and find that the main effects of negation are found in the upper layers. However, these effects are small and do not seem to correlate well with model accuracy.

While causal mediation analysis is a useful analysis tool, it is not straightforward to apply it to all models and to the analysis of attention when the input is of different length in the base case and under the intervention. In the future we would like to extend our analysis to other model variations (e.g., the UNITER$_{\textit{paired-attn}}$ model which processes the images separately from each other), and to attention heads.

Acknowledgements

This work was supported in part by the UKRI Centre for Doctoral Training in Natural Language Processing, funded by the UKRI (grant EP/S022481/1) and the University of Edinburgh, School of Informatics and School of Philosophy, Psychology & Language Sciences.

References

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. 2015. VQA: Visual Question Answering. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 2425–2433.

Jize Cao, Zhe Gan, Yu Cheng, Licheng Yu, Yen-Chun Chen, and Jingjing Liu. 2020. Behind the scene: Revealing the secrets of pre-trained vision-and-language models. In *Computer Vision – ECCV 2020*, pages 565–580, Cham. Springer International Publishing.

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.

Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning. 2019. What Does BERT Look at? An Analysis of BERT's Attention. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 276–286, Florence, Italy. Association for Computational Linguistics.

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

Yanai Elazar, Shauli Ravfogel, Alon Jacovi, and Yoav Goldberg. 2021. Amnesic Probing: Behavioral Explanation with Amnesic Counterfactuals. *Transactions of the Association for Computational Linguistics*, 9:160–175.

Allyson Ettinger. 2020. What BERT is Not: Lessons from a New Suite of Psycholinguistic Diagnostics for Language Models. *Transactions of the Association for Computational Linguistics*, 8(0):34–48.

Matthew Finlayson, Aaron Mueller, Sebastian Gehrmann, Stuart Shieber, Tal Linzen, and Yonatan Belinkov. 2021. Causal Analysis of Syntactic Agreement Mechanisms in Neural Language Models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1828–1843, Online. Association for Computational Linguistics.

Md Mosharaf Hossain, Venelin Kovatchev, Pranoy Dutta, Tiffany Kao, Elizabeth Wei, and Eduardo Blanco. 2020. An Analysis of Natural Language Inference Benchmarks through the Lens of Negation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9106–9118, Online. Association for Computational Linguistics.

Nora Kassner and Hinrich Schütze. 2020. Negated and Misprimed Probes for Pretrained Language Models: Birds Can Talk, But Cannot Fly. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7811–7818, Online. Association for Computational Linguistics.

Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2019. VisualBERT: A Simple and Performant Baseline for Vision and Language. In *Arxiv*.

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

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Judea Pearl. 2001. Direct and indirect effects. In *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence*, UAI'01, page 411–420, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Geoffrey K. Pullum and Rodney Huddleston. 2002. Negation. In *The Cambridge Grammar of the English Language*, page 785–850. Cambridge University Press.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2020. A Primer in BERTology: What We Know About How BERT Works. *Transactions of the Association for Computational Linguistics*, 8:842–866.

Alane Suhr, Stephanie Zhou, Ally Zhang, Iris Zhang, Huajun Bai, and Yoav Artzi. 2019. A Corpus for Reasoning about Natural Language Grounded in Photographs. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6418–6428, Florence, Italy. Association for Computational Linguistics.

Hao Tan and Mohit Bansal. 2019. LXMERT: Learning cross-modality encoder representations from transformers. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5100–5111, Hong Kong, China. Association for Computational Linguistics.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019. BERT Rediscovers the Classical NLP Pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4593–4601, Florence, Italy. Association for Computational Linguistics.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber. 2020. Investigating Gender Bias in Language Models Using Causal Mediation Analysis. In *Advances in Neural Information Processing Systems*, volume 33, pages 12388–12401. Curran Associates, Inc.

Hedde Zeijlstra. 2007. Negation in natural language: On the form and meaning of negative elements. *Language and Linguistics Compass*, 1(5):498–518.

Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. From recognition to cognition: Visual commonsense reasoning. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6713–6724.

A Negation word list

not, isn't, aren't, doesn't, don't, can't, cannot, shouldn't, won't, wouldn't, no, none, nobody, nothing, nowhere, neither, nor, never, without

B Sample selection

During the annotation samples with the following properties were discarded:

- Samples which already contain negation.
- Samples with inappropriate or unpleasant images.
- Samples which are too similar to already annotated samples. Because of the way NLVR2 was created, there are sample pairs that are very similar to each other – those were discarded to increase the diversity of the negation test set.
- Samples with sentences containing typos and other errors.
- Samples for which there is no way to negate the sentences and flip the label, and at the same time keep them grammatically correct.

C Negation examples with images

See Figure 5.

D Negation types in original data

We annotated the negation types of 100 examples from the NLVR2 training set which were automatically determined to contain negation, as well as 50 from the development and test sets each. Results are shown in Table 5. We do not distinguish between NP (existential) and NP (number-to-none), since there are no "original" examples to compare with. The "Other" category contains mostly negated adjectives. The original negated examples differ from our annotation in several ways. First, the sentence-wide type is not found in the original data. This was confirmed by a simple search across the whole dataset for the phrases which compose this negation type ("It is not the case that . . . ", "It is not true that . . . "). Second, there are very few instances of verbal (existential) negation in the original data, whereas this type is very over-represented in our annotation. Finally, NP (nonexistential) is very under-represented in our annotation compared to the original data.

	Train	Dev	Test
Verbal (content)	20	28	22
Verbal (existential)	3	2	10
NP (existential)	45	40	46
NP (nonexistential)	31	32	26
Other	2	2	4

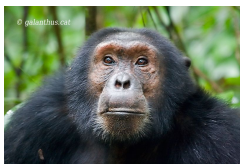
Table 5: Percentage of examples that contain each of the negation types. Note that some examples contain more than one type, so the percentages do not add up to 100.

E UNITER-large triplet results

See Table 6.

F NIE per correctness category

Figure 6 shows the NIEs of each negation type for UNITER-base, comparing between correctness categories. The four correctness categories reflect the correctness of the original sample without the intervention and the correctness of the negated sample. NIE can be negative, which indicates a change towards the incorrect label. Here, we look at both the top 5% neurons and the bottom 5% neurons to find out if some neurons specifically contribute to incorrect predictions. The patterns we see are largely similar between the six negation types. The largest NIEs are observed in cases where both the original and the negated samples are predicted wrong. However, this result may be unreliable due to the small sample size (recall from Section 4 that a large percentage of the originally incorrect samples are predicted correctly when negated, therefore the number of "i-i" examples is small). The other correctness categories all exhibit similar NIEs with the highest values concentrated in the upper layers (8 to 12). We expected to see higher NIE for the "c-c" category, however, that is not the case.

Original: At least one monkey is sitting in a tree in the image on the left. → True

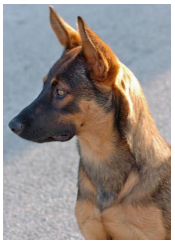
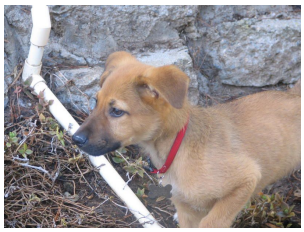
Negated: At least one monkey is not sitting in a tree in the image on the left. → False

(a) Verbal (content)


Original: There are three pandas. → True

Negated: There aren't three pandas. → False

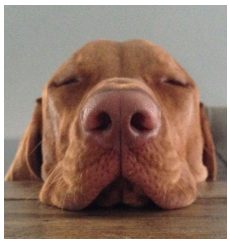
(b) Verbal (existential)

Original: Every dog is wearing a collar. → False

Negated: Not every dog is wearing a collar. → True

(c) NP (nonexistential)




Original: A dog is resting its head on something. → True

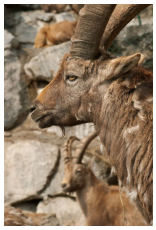
Negated: No dog is resting its head on something. → False

(d) NP (existential)


Original: Four or fewer television screens are visible. → True

Negated: No television screens are visible. → False

(e) NP (number-to-none)

Original: Three or fewer goats are visible. → False

Negated: It is not true that three or fewer goats are visible. → True

(f) Sentence-wide

Figure 5: Examples for each negation type with images and labels.

	UNITER$_{triplet-large}$			
	negative	positive	o.correct	o.incorrect
Verbal (content)	43.62	74.76	36.62	65.22
Verbal (existential)	54.63	68.52	51.35	61.76
NP (existential)	56.36	85.46	59.57	37.5
NP (number-to-none)	58.33	76.38	54.55	70.59
NP (nonexistential)	48.28	60.71	33.33	72.73
Sentence-wide	37.35	72.29	21.67	78.26
Overall	49.43	74.0	43.38	66.38

Table 6: Results on the negation test set.

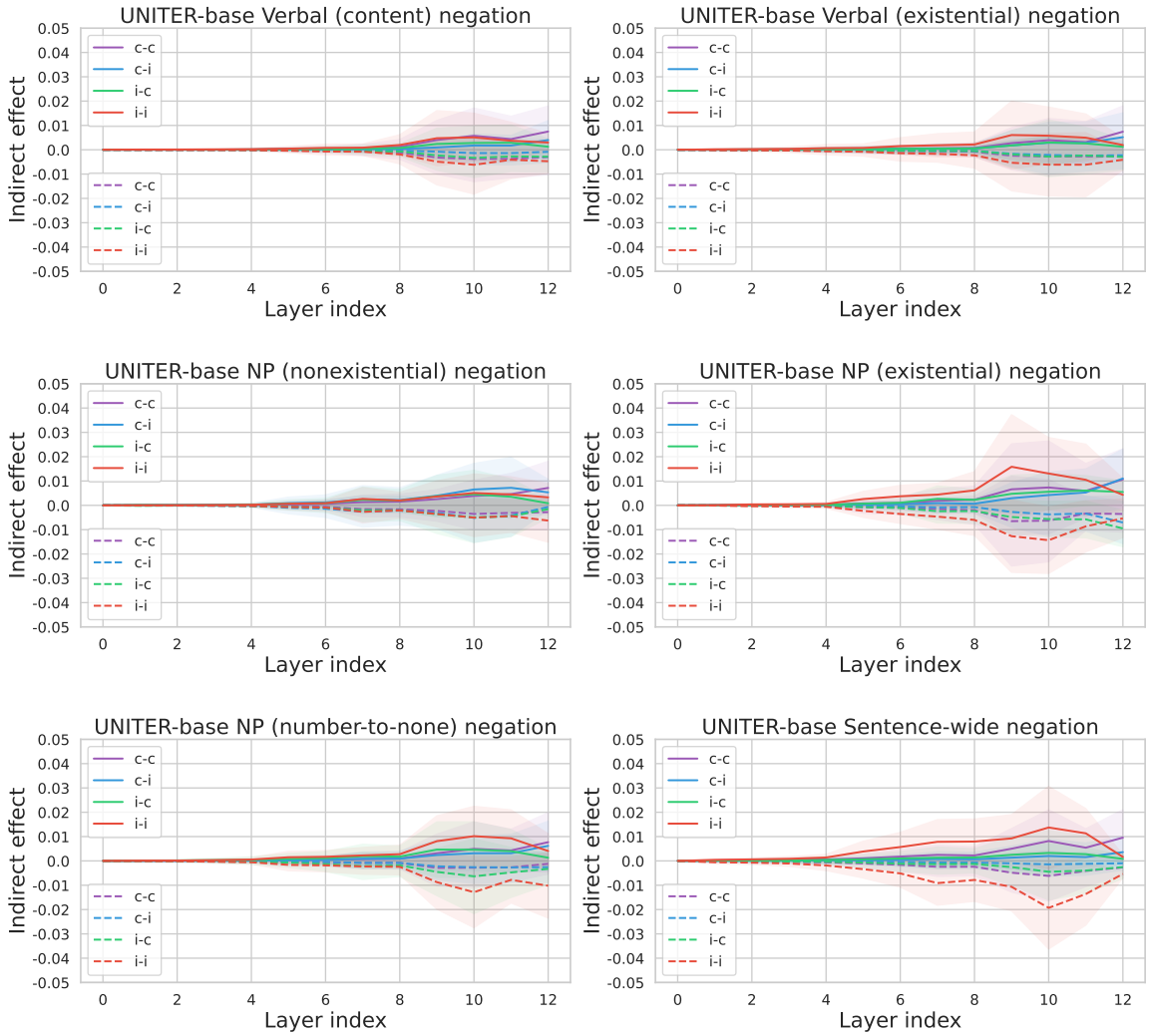


Figure 6: Natural indirect effect of the top (solid line) and bottom (dashed line) 5% of neurons in each layer. The figure shows the NIEs by correctness category: "c-c": both original and negated sample are correctly predicted; "c-i": original is correctly predicted and negated is incorrectly predicted; "i-c": original is incorrectly predicted, negated is correctly predicted; "i-i": both are incorrectly predicted.