|  |  |
| --- | --- |
| - When commenting on the bias toward the passive voice of verbs at page 32, we read: “CLIP shows a preference for active voice sentences with the verb “hold”. However, in Tab 7.3 we observe that this preference is only within the contrast TA-FP, while an opposite pattern is observed between TA-TP. By observing that, for “hold”, the contrast TA-FA is at chance level while TF-FP achieves 71% accuracy, it rather seems that the “frequency advantage” of seeing “hold” very frequently in the passive voice during pre-training makes the model learn to evaluate the verb when in the passive voice, at least to some extent. Could this be an alternative explanation to this bias? More in general, what could be done to more directly test whether the observed patterns are attributable to the training data?  1.The CLIP’s active voice preference for “hold” is discussed because it is very different from the trend of other cases, where others prefer passive voice sentences even when the comparison is between True Active vs. False Passive (TA-FP). We expect the correct sentence is ranked higher than the incorrect ones, so it is less surprising to see models prefer TP in TP-FA (but it’s very slightly as the number 47.46% is very close to the chance level ). But I do agree that the TA-FA, TP-FP comparisons seem to suggest that models are trained better to distinguish semantic roles when the verb is in passive voice. Without access to the CLIP training dataset, it’s difficult to directly test this. We can continue training the pretrained model with more selected cases (e.g. more samples using active voice or passive voice for selected predicates). After presenting the models with more selected cases, we then again evaluate their performance on our tasks and check whether such trends have changed.  - By looking at Table 7.5 and 7.6, it appears that the only predictor that has a significant, positive role towards higher scores in all three tasks and models is the visual one, i.e., person\_size. Whenever present, attribute\_size plays also a significant role in all tasks and models. How could the relation between model accuracy and object/attribute’s size be better tested and further investigated?  2. Can use the object/attribute’s size to predict model prediction accuracy (1 for correct prediction, 0 for incorrect prediction). This can be done with **Binary Logistic Regression**. In addition, as the object/attribute nouns can refer to multiple entities in the image (e.g. more than one boys present in the image). In such cases, it would be good to consider the size of multiple objects instead of the one used to construct the dataset. Our hypothesis can be model’s accuracy is affected by the object/attribute size of the largest entity (e.g. the size of the largest “boy” entity).  - Do you expect fine-tuning the models on some amount of samples from the datasets would lead to improved performance one the tasks? More in general, could the BLA benchmark be used to push multimodal models pay more attention to the fine-grained semantics of visually-grounded sentences?  3. Yes. As discussed in Section 8.1, Nikolaus and Fourtassi (2021) uses a dataset with similar settings as ours (one image with one correct caption and one distractor that design to test language knowledge) to train V&L models from scratch, which is to simulate children’s cross-situational learning. Their models achieved good performance on language tasks like distinguishing semantic roles though their dataset uses cartoon pictures with simple scenes and their models are not the state-of-the-art V&L models. However, we expect that models would benefit from training or finetuning from the BLA benchmark tasks to gain more semantic knowledge. | - Is the benchmark/dataset going to be publicly released to the research community? I am not sure I have seen that mentioned in your thesis. If it is not there, please include the link to the dataset / instructions on how to use it. This is listed as one of the contributions in your thesis, and you should make it public (or clearly explain that also in the contributions).  1. Yes, it is now available on https://github.com/shin-ee-chen/BLA.  - In page 22, you say you use the FOIL-it dataset to examine "models on their noun understanding", but you do not address any of the criticisms to this dataset in the literature. One example is [1], another one is Parcalabescu et al., 2022, who show the dataset contains distributional biases that make it very easy to predict by resorting to "bad" heuristics. In page 41, you state "FOIL task is almost solved by CLIP, which demonstrates the model’s good understanding of nouns" -> I again suggest you at least contextualise these statements.  [1] Pranava Swaroop Madhyastha, Josiah Wang, and Lucia Specia. 2018. Defoiling Foiled Image Captions. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers), pages 433–438, New Orleans, Louisiana. Association for Computational Linguistics.  2. Thanks for pointing out the criticisms of the FOIL. I agree that the statement should be contextualized regarding to these problems of the FOIL dataset.  - In pages 24-25: "If a correct caption is ranked first or second, or a distractor is ranked third or fourth, the prediction for the sentence is considered CORRECT." This is a bit ambiguous to me. Do you mean that if the two correct captions are ranked first and second, and the two incorrect captions are ranked third and fourth, the prediction is considered correct (and it is considered incorrect otherwise)? Please fix. This is the only place you phrase it like this, everywhere else I read as I wrote (my "correction").  3. Yes, your understanding is correct. This sentence should be rephrased.  - There are a few too many rendering/linking mistakes/typos in the thesis, please fix these. For instance, in page 12, you say "In Figure , the correct caption (...)". In page 16, you state "The sentences discarded are shown in Table ??." In page 33, "Table 7.3: Multimodal pair comparision" -> "Table 7.3: Multimodal pair comparison"; In page 42, "trained from scratches" -> "trained from scratch"; In page 45, "In the Version, (...)" -> Which version?  4. Thanks for pointing out these typos. They will be fixed in the updated version.  - In page 14 you state "And the distractor sentence “A woman is held by a knife” can be easily excluded by models without considering the image since such sentences would not occur in the training data." -> Though that is in principle true and an expected behaviour in trained models, that is not what happens in practice. See the Action piece, Actant swap instrument in Parcalabescu et al., (2022).  5. We noticed that current V&L models are not able to detect such counterfactual sentences. However, we hope that this benchmark can be used for future work, where more powerful models might be able to detect such cases. In addition, our evaluation focuses on basic language abilities and uses similar tasks as children’s language comprehension evaluation. So we do not want to include such cases that can be solved using commonsense knowledge.  - In page 42, "The Children’s book contains tasks regarding 7 phenomena" -> Citation needed.  6. The Children’s book was mentioned in the Introduction with citation, but yes citation should be also added here. |