Evaluating real world safety and robustness of deep learning models

Authors: K. Fallon, A. Lorenz, J. Loucks-Tavitas, S. Tsiorintsoa, and B. Warren
Pacific Northwest National Laboratory Mentors: C. Godfrey and H. Kvinge

EXECUTIVE SUMMARY

Deep learning and AI have the capability to positively impact businesses in myriad ways. Their ability to replicate or even outperform humans in many seemingly non-computational tasks has been making headlines for the past couple of years now. However, their knack for "hallucination" and confidently proclaiming incorrect outputs can be worrisome. In light of this, our task was to evaluate two deep learning models, a language model to be used as a scientific knowledge engine and a segmentation model to be used for person detection, from the perspective of overall robustness. In our context, we interpret robustness as sensitivity to perturbations in the input. Our main takeaway is that both models, especially the language model, are quite sensitive to input perturbations. Moreover, the nature of the models' responses to input perturbations is unpredictable. We recommend accepting these models for their proposed uses under fairly narrow guidelines: in the case of the language model, the user should be provided with a template of how to input their prompt and the model should be provided with a detailed pre-prompt explaining its task and giving examples. In the case of the segmentation model, it should be verified that any image being segmented is sufficiently clear, especially with respect to blurriness and pixelation. Moreover, we do not recommend the use of this model for person detection in high stakes contexts as the level of error and unpredictability is too high. Finally, it is important that the user interpret the output of both of these models not as ground truth, but as an approximation thereof. If the guidelines are followed, the approximations should be sufficiently accurate for the proposed uses.

RECOMMENDATIONS AND LIMITATIONS

We recommend using the models for their proposed uses under the following guidelines.
For BLOOM:

- Provide the user with a template of how to input their prompt.
- Provide the model with a detailed pre-prompt explaining its task and giving examples.

and for SAM:

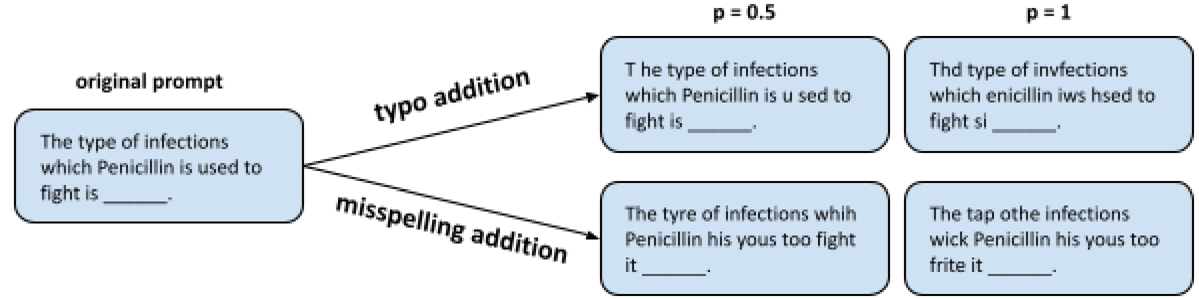
- Verify that any image being segmented is sufficiently clear, especially with respect to blurriness and pixelation.
- Do not use this model for person detection in high stakes contexts.

For both: it is important that the user interpret the output of both of these models not as ground truth, but as an approximation thereof. With that being said, we experienced a few limitations in our research which, if addressed, could potentially lead to less restrictive guidelines in using these models. One limitation is that the dataset we were using to study SAM consisted of only 22 images. With such a small dataset, large scale trends are very difficult to identify and accurately characterize. Moreover, even with just 22 images, it took quite some time to run all the tests we ran; some experiments took multiple hours. In a three week long project, this posed a serious limitation on the amount of experimentation we could do and thereby the amount of data we could collect.

METHODS

The models that we researched in this project were the BLOOM language model and the Segment Anything Model (SAM) by Meta. Preliminary experiments made it clear that a variance in input data corresponded to a variance in the achievement of a desired result. For example, asking BLOOM to complete a sentence was much more successful than asking BLOOM to answer a question. Thus our primary measure of robustness in both models began to center on the perturbation of inputs and the resulting effect on outputs. We fed each model corrupted input data and compared the outputs to ground truth values, resulting in a measure of accuracy which we then committed to the visualizations later in this paper.

Our work with the BLOOM model began with improving the output accuracy via prompt engineering. We included a fixed pre-prompt in each query with the goal of uniformity and accuracy of responses. We varied what we included in these pre-prompts, from instructions to context to examples, and measured the results of each.



The type of infections which Penicillin is used to fight is ______.

T he type of infections which Penicillin is u sed to fight is ______.

Thd type of invfections which enicillin iws hsed to fight si ______.

The tyre of infections whih Penicillin his yous too fight it ______.

The tap othe infections wick Penicillin his yous too frite it ______.

*Figure 1: Adding typos or misspellings to the prompt at probability **p***

The other change in input modeled user errors in prompting (see Figure 1). In both cases of altering the query input, accuracy was measured by a word embedding: Representing two phrases as normalized vectors in a high-dimensional vector space and measuring the angle between them gives a strong metric of semantic similarity. This metric, called “cosine similarity,” ranges from -1 (least similar) to 1 (equal in meaning).

With SAM, our perturbations focused on simulating in an elementary way the physical conditions that a camera may encounter. These input changes include blurring an image to model a camera going out of focus, making changes to the contrast and brightness to simulate diverse lighting conditions, and pixelating an image to simulate processing error. Our dataset was a set of images from PascalVOC that had been filtered down to 22 images that contained exactly one person in each image. The SAM model was run on the altered versions of these images with the prompt to segment the person in the image. The model returns three “masks”, or segments, which are the best attempt by the model to segment the person in the image.

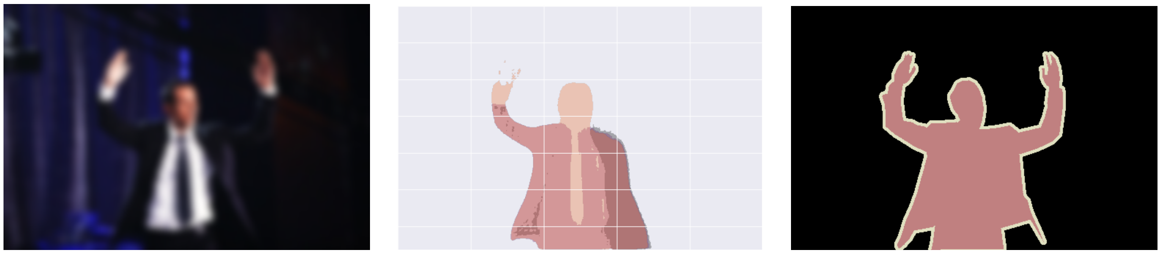


Figure 2. From left to right: A blurred image, a segmentation of the blurred image by SAM, and the ground truth segmentation of the original image.

We added these masks together to form the model’s prediction of the person and compared this to the ground truth image, which contains the true segmentation. Accuracy was measured using the “Jaccard score”, or intersection over union (IoU). Each pixel in both images is assigned “person” or “not a person”, then IoU counts the number of pixels that agree with each other divided by the total number of unique “person” labeled pixels between the two. This metric results in a score from zero to one, zero being the worst and one being the best.

RESULTS

As alluded to previously, both rephrasing the questions and including a pre-prompt of examples when submitting queries to BLOOM resulted in higher accuracy of the model output. As seen in Figure 3, the response accuracy (measured by average cosine similarity between the output and correct answer) nearly doubled after having added one example, but the increase in accuracy is negligible after more examples are added. In both the cases of simulating typos and misspellings in the prompts, we observed a substantial decline in accuracy of the model's responses. The results can be seen in Figure 4, in which we plot the change in correct responses as the probability of misspelling a word increases from 0 to 1 (at increments of 0.2). The left-hand graph indicates a steady decrease in the percentage of correct answers that BLOOM replied with as the probability of an error increased. We see that BLOOM responded correctly around 76% of the time when no errors were introduced, while introducing typos and misspellings at probability 1 produced only 51% and 32% correct answers, respectively (see Appendix for typo graphics). The number of "very incorrect" (accuracy score of less than 0.5) answers also increased significantly, which is visible in the right-hand graph.

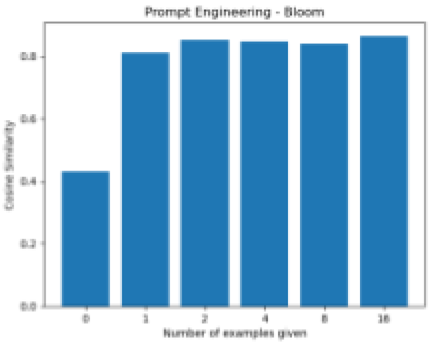


Fig 3: Benefits of adding more examples in the pre-prompt

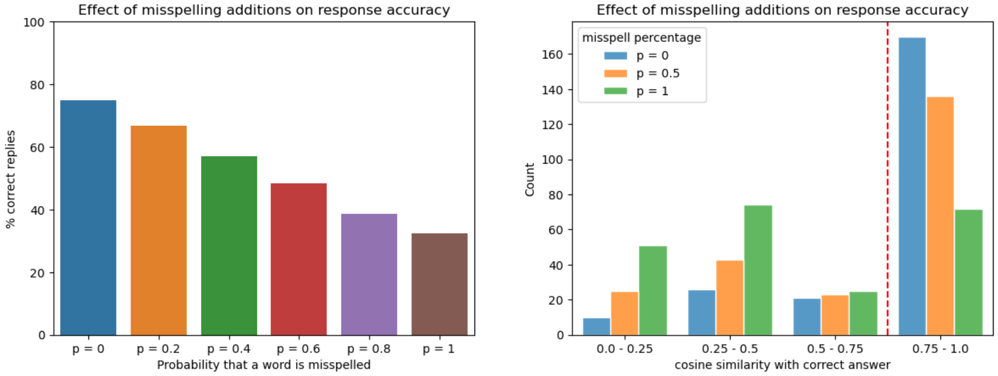


Figure 4: Results from the BLOOM language model testing

The first round of testing done on SAM adjusted the contrast, darkness, contrast and darkness combined, and blur of the image before passing it through the model. The collection of graphs below illustrate the behavior and accuracy of the SAM model on these adjusted images.

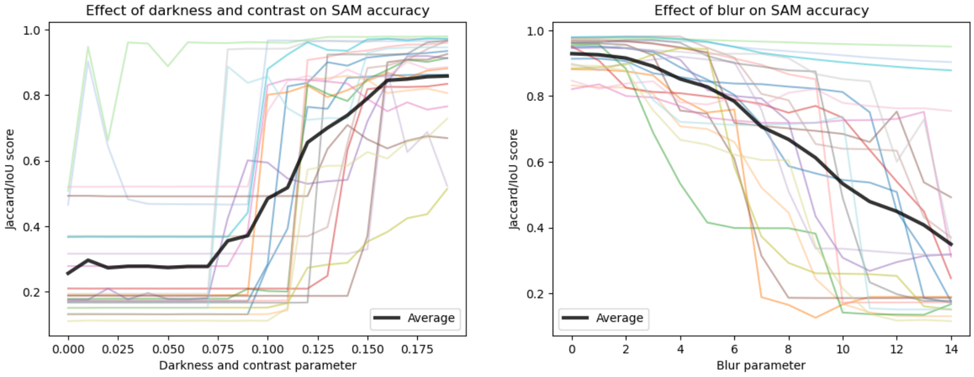


Figure 5. Results from SAM testing using Pillow (2 graphs)

Each colored line in Figure 5 represents a single image. The values along the vertical axes in each graph are the Jaccard scores, while the values along the horizontal axes are the different parameters. For darkness and contrast, a parameter of 0 corresponds to a black or gray image, respectively, while a parameter of 1 returns the original image in both cases. The darkness and contrast combined, the parameter number "x" is interpreted as applying parameter "x" in the contrast function, then taking the resulting image and applying “x” once again in the darkness function. The blur parameter returns the original image at 0 and increases blur as the number grows. The bold black line in each graph is the curve that models the average Jaccard score amongst all images with the given parameter.

Of particular note in the visualizations above is the “thresholding” behavior of each image that is disguised in the average curve. That is to say, most images exhibit a tipping point when things go from bad to good (or vice versa). In our observations, this aligned with the first point in which the model began to incorporate the background into its segmentation, resulting in a large difference from the ground truth. We believe this to be natural, as this would be the first point when the model can’t tell apart the person from the background, i.e. cannot accurately detect the person. However, we also believe that part of this is due to our data collection methods, which will be discussed later in this paper.

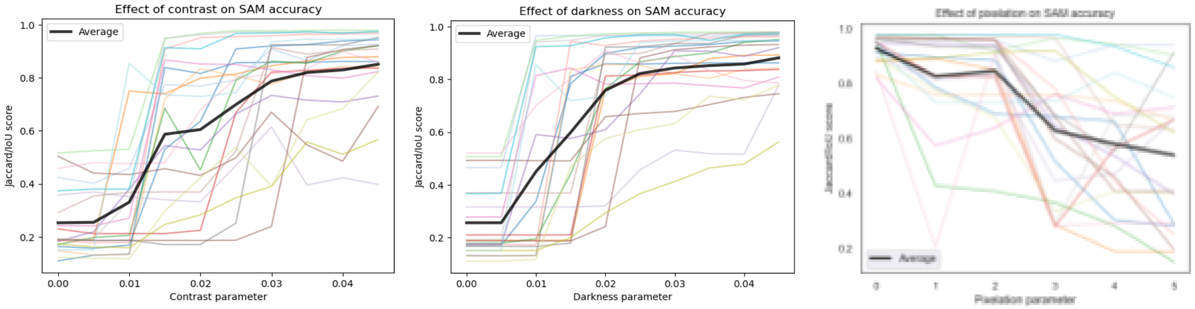
As for pixelation, our collected data indicates that it does negatively impact SAM’s ability to properly segment images. In particular, we note the general downward trend in the plot on the bottom of Figure A.1 as depicted by the average line. However, there are easily identifiable examples where the IoU score counterintuitively increases with increased pixelation.

ACKNOWLEDGEMENTS

We are grateful to Lambda Labs for the provision of GPU hours to run the SA Model. We also thank Hugging Face for the API access to run the language models. Special thanks goes to our project mentors, Charlie Godfrey and Henry Kvinge, as well as PNNL as a whole, for their excellent mentorship and feedback throughout our project. We are deeply thankful for the invaluable guidance and instruction from our IMA Math-to-Industry Boot Camp mentors Thomas Höft and Daniel Spirn and our Bootcamp instructors. Finally, we would like to thank the other members of our cohort whose discussions, camaraderie, and intellectual engagement enriched our experience during the Boot Camp.

APPENDIX

We include some more plots of our data. Explanations can be found in the methods and results sections.



.

Figure A.1. Results from SAM testing using Pillow (3 graphs)

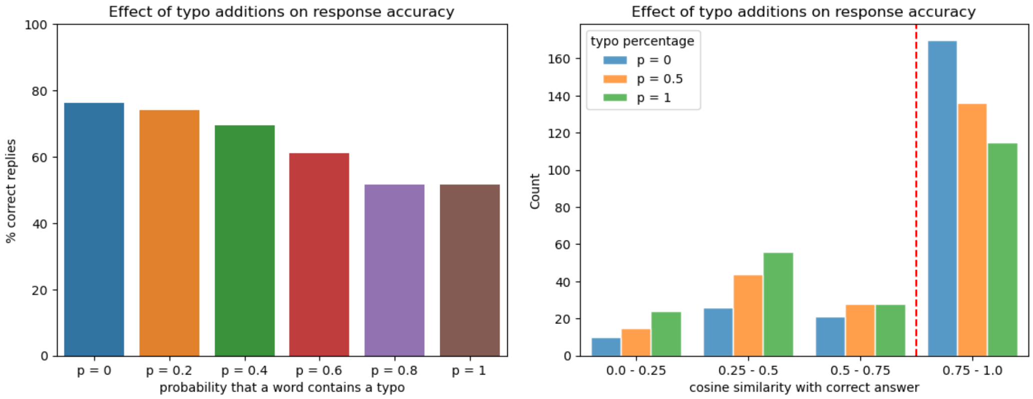


Figure A.2. Results from the BLOOM language model testing (2 graphs)

REFERENCES

Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., … Oliphant, T. E. (2020). Array programming with NumPy. Nature, 585, 357–362. https://doi.org/10.1038/s41586-020-2649-2

McKinney, W., & others. (2010). Data structures for statistical computing in Python. In Proceedings of the 9th Python in Science Conference (Vol. 445, pp. 51–56).

Python Package Index - PyPI. (n.d.). Python Software Foundation. Retrieved from https://pypi.org/

all-MiniLM-L6-V2 sentence transformer. Hugging Face. (n.d.). https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2

Mitton, R. (n.d.). Corpora of misspellings for download. Roger Mitton's Home Page. https://www.dcs.bbk.ac.uk/~ROGER/corpora.html

Newman, S. (2021, October 28). 220+ science trivia questions and answers. Thought Catalog. https://thoughtcatalog.com/samantha-newman/2020/04/science-trivia-questions/

Everingham, M., Van Gool, L., Williams, C. K. I., Winn, J. and Zisserman, A. *International Journal of Computer Vision, 88(2), 303-338, 2010.* http://host.robots.ox.ac.uk/pascal/VOC/

Segment Anything. (n.d.). https://segment-anything.com/