Focusing Image Captioning Models on Action-Effect Relations

Thomas Gregart
tgregart

Ethan Chen
ethandc

Lance Bassett
lanbas

1 Introduction

Despite recent advances in knowledge representation, automated reasoning, and machine learning, artificial agents still lack the ability to understand basic action-effect relations regarding the physical world. For example, slicing an apple will most likely lead to a state of the world where an apple has been separated into smaller pieces. This type of understanding is imperative to these artificial agents helping us in joint tasks, so they can reason about the changing state of the world and plan out its actions. With this setup in mind, it is important for these agents to understand the state of the world through images and reason about them through actions expressed in language. There is a current CNN approach to this problem, but there haven't been many experiments surrounding the architecture this model. Additionally, data is especially important because there are so many possible actions to assess in the world, so experiments that can work around this scarcity would be vastly more beneficial.

This report presents an image captioning based approach using the One-For-All (OFA) model [1] shown in Figure 3, and attempts to translate a captioning model's understanding of the world into reasoning about action-effect causality. We look at the baseline understanding of pretraining captioning models, attempt to improve this with further training, and also explore sources of error present in our dataset.

2 Related Work

Below we have linked the research papers that we will be building upon:

1. OFA paper [1]

2. Action-Effect paper [2]

2.1 Physical Causality Reasoning

Previous work poses physical causality reasoning as a classification problem. A study from the SLED Lab at Umich [2] uses a CNN on effect images to predict the most likely action-effect label in the dataset. As mentioned in [2], there has been previous work on establishing cause and effect relationships from text [3], but these are mostly high level reasoning from information, rather than understand the physical reality of a cause and effect. They also introduce a dataset of 140 verb-noun pairs and associated images that display the result of the action performed on the world e.g. "chop wood" or "boil egg." Examples images are shown in Figure 1.

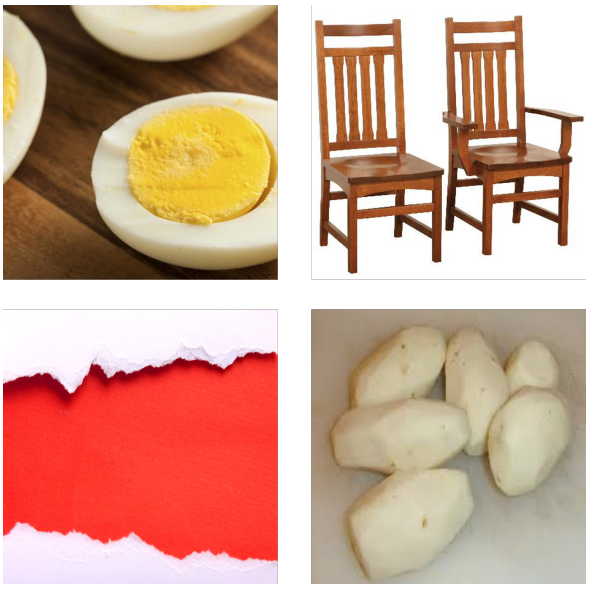

Figure 1: Diagram of our matching process. We obtain a caption, process it, then match to the closest noun-verb pair we find in the GloVe embedding space

2.2 Image Captioning

Image captioning is a multi-modal natural language task used to describe what is happening in an image. This task generally requires understanding of both images and text, as well as the relationship between

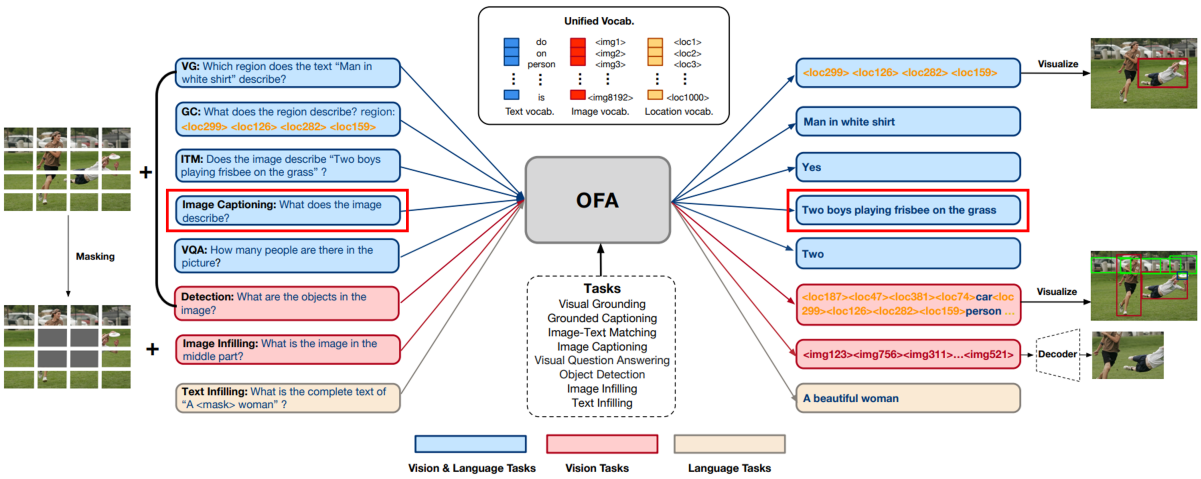

Figure 2: Diagram of the OFA architecture. The image captioning sections relevant to us are circled in red. They encode image captioning as a visual question answering task for the query "What does the image describe?".

the two. For many years, these understandings have been achieved through convolutional neural network feature extractors combined with language encoders and decoders [4] [5]. The state of the art has since shifted towards RNNs, then LSTMs [6], and now Transformers [1] [7] [8]. Transformers alleviate shortcomings in training efficiency and expressibility compared to previous methods. We plan to harness transformer based image captioning models' understanding of scenes to extract action-effect relations in images.

2.3 Unified Frameworks

OFA is not the first attempt at a unified, one-for-all framework. Perceiver io [9] initially found success with combination of architecture from different domains (convolutions, attention, gated layers, etc.). They suggested a move away from domain and task assumptions and proposed a general multi-modality architecture. OFA is inspired by these previous works and builds on them with a multi-modality, task-agnostic, Transformer-based approach. With a uniform byte-sequence representation, OFA makes it easier than ever to unify tasks of different modalities. It's success comes from designing various task-specific layers, but this universality has its drawbacks. The result of the layer blending is performance degradation in downstream tasks.

3 Methodology: Our Approaches

We have approached this problem from the perspective of image captioning, with the idea that models capable of captioning a scene accurately would already have an inherent understanding of the world and actions performed on it. Below we detail our methods of image captioning with OFA, using these captions to extract verb-noun pairs, and our attempts to eliminate error in our dataset.

3.1 Image captioning with OFA

Though image captioning does not inherently understand action-effect relationships, a model with a comprehensive understanding should be able to understand the state of an image well enough to infer action. This assumption is the basis of leverage for our research. To test this assumption, we found the number one image captioning model for the MSCOCO dataset, OFA. With the diversity of MSCOCO's 328,000 images of everyday objects and people, OFA is able to grasp a better understanding of various states of the world. With this understanding, we hope to help OFA understand the relationships between actions and their results on the state-of-the-world. To test this theory, we utilized the relationships built into the action-effect dataset. First, we fed the effect images into a pre-trained OFA-Base model with 180 million parameters. The resulting captions were then used to try to predict a corresponding action-effect label. Second, we fine-tuned the OFA model on the dataset to see if there was any improvement in its understanding in action-effect relationships. If OFA is able to generate a caption detailed enough to incorporate this relationship, the novel field of physical causality research would be able to leverage the work done on a task with much more work devoted to it.

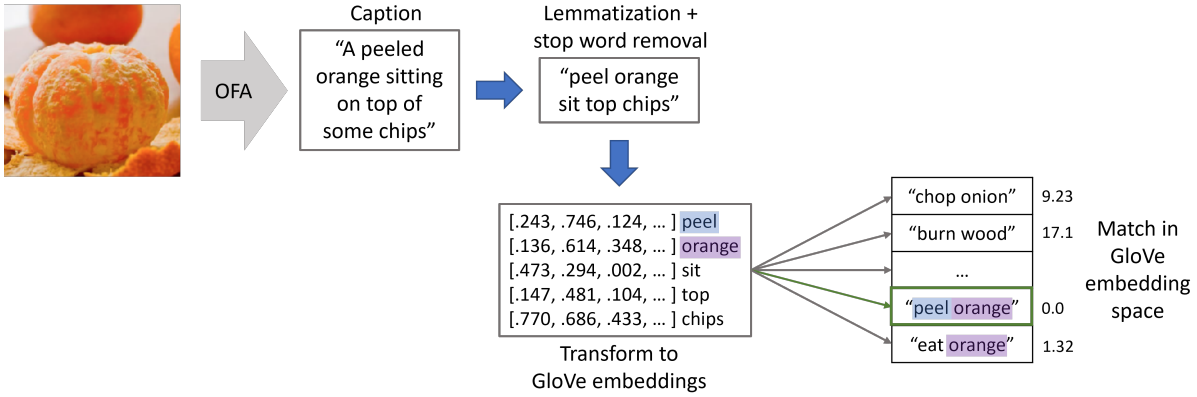

Figure 3: Diagram of our matching process. We obtain a caption, process it, then match to the closest noun-verb pair we find in the GloVe embedding space

3.2 Verb-Noun extraction from image captioning

Similar to [2], we frame our action-effect prediction as a 140-way classification problem, with one class for each verb-noun pair in the dataset. Rather than directly predict the verb-noun class, once we obtain an image's caption from OFA, we employ a series of steps to process and match the caption to it's correct class. First, we obtain the caption from OFA, then we perform lemmatization and stop word removal to improve processing accuracy and avoid noise in our matching step. These steps turn "a peeled orange sitting on top of some chips" into "peel orange sit top chips". We then obtain GloVe [10] embeddings for all remaining words in the caption and all of the verb-noun pairs in the dataset. For each verb-noun pair in the dataset, we find the word in the caption with the smallest euclidean distance d_v to the verb in the GloVe embedding space, then do the same for the noun and obtain d_n. The score $s(P, c)$ for a noun-verb pair P given a caption c is:

$$s(P, c) = \frac{d_v + d_n}{2}$$

, and the prediction for the caption we generate is the noun-verb pair with the lowest score. A visual demonstration of this process is shown in Figure 2.

We perform this process on captions obtained from an OFA model pretrained on MSCOCO, as well as one finetuned on the action-effect dataset. We split the action effect dataset into 80% train, 20% validation and finetune for 5 epochs with a learning rate of 1e-5 and dropout probability 0.1.

3.3 Dataset Cleaning

The action-effect dataset is certainly a small one on the scale of deep learning. First, 40 nouns were selected that represent everyday life objects, most from the COCO dataset (e.g. apple, door, window, etc.). These objects represent a wide range of everyday objects from food, kitchenware, furniture, indoor objects and outdoor objects. Then, analysis was done on the top 3000 most frequently used verbs from Google Syntactic N-gram dataset. From there, the top frequent verb noun pairs containing one of the 3000 verbs and one of the 40 nouns holding a direct object dependency relation were selected. This resulted in thousands of candidate examples, but verb-noun pairs were manually selected from this batch. The selection was based on changes to the objects that were clear results from the verb, as well as changes the reflect a particular dimension that can't be generalized. For example, "cook a meal" can be decomposed to more basic actions like "cut onion." The end result was 140 verb-noun pairs.

While the intention of the dataset construction was to decompose to the most basic actions, we found that these enforcements were not held for the entirety of images selected. After confusing initial results, we investigated the dataset and realized much of the data was difficult even for us to label correctly, not to mention our model. A few such examples are shown in Figure 4, and these difficulties fell into three categories.

The first is what we named the "potato precedent." This problem got its name because we noticed that there were up to 9 verbs associated with one noun, and one such noun was "potato." More specifically, 5 of these verbs had to do with varia-

tions of preparation: [bake, cook, boil, fry, mash] potato. While boil, fry, mash, and bake are different actions that we want to differentiate between, “cook” envelopes all of these under its definition, and the images within “cook” often display multiple variations of cooking that overlap with the other forms of preparation. We expect this can introduce error into our predictions because captions for “cook potato” are likely to match to a more specific variation of cooking depending on the image. We can apply this same reasoning to our pruning of the “meat” and “egg” categories.

Second, we found verbs that were synonymous, or extremely close to each other. For example, "burn wood" and "ignite wood" were separate labels in the dataset. While there are subtle differences to these verbs, these differences weren’t captured in the images under each label. In fact, we found several duplicate images between labels of similar categories. In total, there were 31 images duplicated across different labels. These, of course, were removed because it’s impossible to expect a model to predict different labels for the same image. To further alleviate synonymous verbs, we also removed verb-noun pairs that we thought were equivalent e.g. "close door" and "shut door", or "burn paper" and "ignite paper".

Third, we noticed that while there may have been labels with different nouns and verbs, there still was a possibility that the content of the images made it futile to distinguish between labels. One example of this problem was between "break window" and "crack glass." At first look, these actions seem very distinguishable, but upon a scan of the images we saw many images in the "break window" category with the structure of the window still intact thus seeming like cracked glass. Furthermore, some of the images in "break window" did not include the window frame at all and simply looked like broken and cracked glass, so our assessment would have been cracked glass without knowing what label the image belonged to. Another example of this would be between "chop wood" and "pile wood". These verb-noun pairs mean understandably different things, but both contain many images of piles of chopped wood. In our opinion, both labels are valid for these images, so we remove one of the categories to avoid this error. This issue of the noun not being accurately represented in the images was prevalent throughout our dataset, so extensive handpicking was done throughout many of the labels to ensure there would be no confusion about the effect depicted in the image.

As a result of the various issues we encountered with the dataset, we removed 32 labels and 491 images from the original 2100 in the ’positive’ directories of all the original labels. Of course, limiting the training data in a deep learning problem seems like a bad option. We came to the understanding that the model would be able to distinguish between these labels much easier.

Figure 4: Above are examples from the dataset that we found indistinguishable by their assigned labels. We see "burn wood" (left) vs. "ignite wood" (right), "break window" (left) vs. "crack glass" (right), and "bake potato" (left) vs. "cook potato" (right)

4 Evaluation

On top of the several different approaches in our methodology, we took multiple approaches for our evaluation. We compute accuracy, top 5 accuracy, mAP, average distance to match, average score difference for incorrect predictions, and a GloVe synonym accuracy for baseline and finetuned models on three different datasets, each with different levels of data pre-processing. The first dataset (Dataset 1) consists of all “positive” images for each verb-noun pair. The second dataset (Dataset

1.5) is identical to Dataset 1 except duplicate images across verb-noun pairs have been removed (31 images). For the third dataset (Dataset 2), we removed verb-noun pairs we found too similar (491 removed, detailed in section 3.3).

Dataset 1 consisted of 2048 images split into 1649 training and 399 validation. Datasets 1.5 and 2 were created from smaller, pruned sets of images as mentioned above, but we split these datasets (which are themselves subsets of Dataset 1) such that their training and validation sets were subsets of Dataset 1. In other words, we did not split Datasets 1.5 and 2 randomly, but instead any image that appeared in the training set of Dataset 1 was training for 1.5 and 2, with the same process used for validation. This allows better performance comparison across the datasets and lets us clearly see if our dataset error analyses were correct. Since images in Dataset 1 were uniformly divided into training and validation, the distribution of Datasets 1.5 and 2 are still very close to 80% training, 20% validation. Dataset 1.5 contains 1622 training and 395 validation. Dataset 2 contains 1239 training and 297 validation.

4.1 Metrics

Accuracy and Top 5 Accuracy: Once we have matched each caption to a noun-verb class from the dataset, we compute a simple accuracy metric *num_correct / total samples*. We additionally compute a top 5 accuracy metric, which is computed exactly the same as accuracy, but a prediction is counted as correct if the correct noun-verb pair from the dataset is one of the top 5 choices from the matching step (even if it is not the first choice).

mAP: mean Average Precision (mAP) is computed as normal. For each noun-verb pair, we compute the precision as

$$\frac{TP}{TP + FP}$$

and take the average across all noun-verb pairs to obtain mAP.

GloVe Synonym Accuracy: Because of our previous difficulties with images that could fit multiple labels, we decided to create our own metric that would take label similarity into account. Since we used GloVe embeddings to be able to relate the generated caption to each of the 140 labels, we are able to measure how different the predicted label and ground truth label are in meaning by taking a distance measurement between their GloVe embeddings. Therefore, we could determine that if the GloVe distance between the prediction and ground truth stayed below a designated distance threshold, the predicted and ground truth labels were close enough in meaning to capture an understanding of the action-effect relationship. After combing through GloVe distances between all labels in our dataset, we discovered that a Pythagorean distance of less than 5.3 for verbs and 4.5 for nouns could determine enough similarity in meaning. If the pythagorean distance fell below these values, we determined that the model understood the action that caused the state of the world in the image. This meant that the model could predict "fry potato" and our new metric would mark this prediction correct, even if the label was "cook potato." This metric is especially important for Datasets 1 and 1.5 because synonymous labels were not removed, and we found that this could be a more efficient way to avoid the need to hand-prune these datasets.

Average Score Difference for Incorrect Predictions: When we predict incorrectly, it means that a verb-noun pair, that was not the ground truth label, was calculated to be closer to our generated caption than the label. This would generally suggest that our caption did not contain information relevant enough to our target verb-noun pair, so it matched with some other verb-noun pair. We report a metric that measures how much further, on average, the ground truth verb-noun pair is from our caption, compared to the one we predicted. So, in the context of being incorrect, when this metric is low we were close to being right, and it suggests that our caption had relevant information, but maybe not quite enough. When it is high we were very far from being right, and it suggests the caption had very little relevant information, so some other label in the dataset, likely not close to our ground truth, matched much better.

To give a more concrete example, if our ground truth verb-noun pair was "bake potato", our caption was "mashed potatoes in a brown bowl", and our prediction was "mash potatoes", this metric would likely be very low because we would have been close to predicting "bake potato." In contrast, if our ground truth verb-noun pair was "bake potato", our caption was "fried steak on a plate", and our prediction was "fry meat", this metric would be very high because we would not have been close to predicting "bake potato."

4.2 Results

For each of the three datasets, we applied the baseline pretrained OFA-Base model to the test data, which was about 390 images for the first two datasets and 290 for Dataset 2. These pretrained weights are what performed number one on the MSCOCO dataset for image captioning, so we felt confident in obtaining relatively accurate captions because the action-effect dataset consists of common objects. However, we continued to look for improvements, so we also finetuned the OFA-Base model on the training data from each dataset. This was 1649 images for Dataset 1, 1622 images for Dataset 1.5, and 1239 images for Dataset 2. The action label, e.g. "chop onion", was used as the label this time, instead of an image caption. Our hope was that tweaking the model's expectations for image labels would broaden its efficacy on physical effect reasoning. These results are captured in Figures 5, 6, and 7 with all of the aforementioned metrics. Overall, we found better results after finetuning the model on our datasets with the largest improvements occurring on the Top One Accuracy metric.

5 Discussion

5.1 Dataset 1 Findings

The model in the original paper found its best results from a combination of bootstrapping, seed images, and Dataset 1. We can see the [2]'s relevant results in Figure 8. Though we didn't experiment with the first two methods, we gathered comparable results with our model in all the reported metrics, except mAP [2]. Though our reported mAP was 0.27 lower, we obtained a top 1 accuracy 0.05 lower and a top 5 accuracy 0.07 higher. These results show that there are certainly improvements to be made with our method, but our hypothesis of leveraging image captioning tools has promise.

5.2 Dataset 1.5 Findings

Because we expected the model to generate the same caption for the same image, we expected to see improvements after removing duplicate images from the dataset. This is because the both images would be classified under one of the labels, so we were guaranteed to get one classification wrong. After running our experiments, our expectations appear to be misplaced. One explanation for the unexpected results is the way that duplicates were removed from the dataset. The removal process considered the first label alphabetically to be the rightful owner of that image, and all other duplicates occurring later in the dataset were removed. It could be the case that the first label with a duplicate was not the label the model was always predicting for that image.

5.3 Dataset 2 Findings

After the intensive hand-pruning of Dataset 2, the expectation was that these results would outperform the other two and the data from previous work [2]. While this was the case in a generally, we saw a few surprises. First, the baseline metrics did not improve and mAP was lower than the other results. A possible explanation for this could be the lack of images in the validation set, or the removal of higher quality images from labels deemed to be distracting. Second, the Top 5 Accuracy was nearly identical to the others. This can be taken as a good sign that our model is close to predicting the correct label despite the distractions in the dataset. Still, these results were our collective best. Compared to the previous work's data, We saw 0.02 increase in Top 1 Accuracy and 0.07 increase in Top 5 Accuracy with only 0.2 decrease in mAP. Results like these are not significant enough to determine that our methods are preferable, but they are comparable enough to warrant more research in overlapping image captioning with physical causality reasoning.

5.4 Findings Across Datasets

We can see the direct improvement that our changes brought to each dataset iteration, but we notice that finetuning OFA image captioning on our action-effect dataset reveals performance improvement across datasets. We believe finetuning was very beneficial it allowed OFA to better understand what our dataset was looking for, which was verb-noun pairs. Image captioning is a relatively unconstrained task, where the model will come up with whatever description it feels best fits the image. This may include incorrect information, as well as information that is unnecessary to our action-effect prediction task e.g. the location of objects in the image, or information about the surroundings. Examples of some poor captions obtained by the baseline are shown in Figure 9. We saw only 1 or 2 word predictions after fine-tuning, and these words were limited to verbs and nouns appearing in our dataset, which helped simplifiy the matching step tremendously. These improved predictions (as well

Model \ Metric	Top 1 Accuracy	Top 5 Accuracy	mAP	GloVe Accuracy	Avg. Incorrect Score Difference
Baseline	0.233	0.830	0.223	0.361	8.97
Finetuned on D1	**0.474**	**0.917**	**0.394**	**0.617**	**3.47**

Figure 5: Above are the results of the several metrics measured on **Dataset 1**. The first row holds metrics from the OFA public pretrained weights, and the second row hold metrics resulting from training the OFA model on the dataset.

Model \ Metric	Top 1 Accuracy	Top 5 Accuracy	mAP	GloVe Accuracy	Avg. Incorrect Score Difference
Baseline	0.233	0.823	0.222	0.360	8.948
Finetuned on D1	**0.473**	**0.917**	**0.393**	**0.615**	**3.510**

Figure 6: Above are the results of the several metrics measured on **Dataset 1.5**. The first row holds metrics from the OFA public pretrained weights, and the second row hold metrics resulting from training the OFA model on the dataset.

Model \ Metric	Top 1 Accuracy	Top 5 Accuracy	mAP	GloVe Accuracy	Avg. Incorrect Score Difference
Baseline	0.266	0.832	0.212	0.370	7.970
Finetuned on D2	**0.539**	**0.916**	**0.450**	**0.596**	**4.105**

Figure 7: Above are the results of the several metrics measured on **Dataset 2**. The first row holds metrics from the OFA public pretrained weights, and the second row hold metrics resulting from training the OFA model on the dataset.

	MAP	Top 1	Top 5	Top 20
BS+Seed+Act+Eff	**0.660**	**0.523**	**0.843**	**0.954**
BS+Seed+Act	0.642	0.508	0.802	0.924
Seed+Act+Eff	0.289	0.176	0.398	0.625
Seed+Act	0.481	0.301	0.724	0.926
Seed	0.634	0.520	0.765	0.892

Figure 8: Result table from [2] showing results for prediction verb-noun labels given images. We compare to their best results because no row's data is directly comparable to our procedures.

as their 1 or 2 word format) are likely why we see a reduction in our average score difference metric for incorrect predictions. The average incorrect score difference lowered from 8.97 to 3.47, 8.948 to 3.51, and 7.97 to 4.105 for datasets 1, 1.5, and 2, respectively. These metrics suggest that when our model was incorrect after finetuning, it was much closer to being right than our baseline would have been, likely because the model's caption predictions were constrained to small captions about objects known to be contained in the dataset.

We would like to note however, that it is unlikely that our model or the model from [2] would be able to correctly generalize to new actions. Both of these are setup as classification problems where the models are trained to predict/learn labels in the action-effect dataset, so expecting it to predict outside of these labels in unrealistic. We would expect, however, that it is able to generalize to new images within the classes the dataset already contains. For example, new images of chopped onions or mashed potatoes could be added to the dataset and we would expect the model to correctly match these. A future direction of research could involve exploring more generalizable approaches to this problem, where the model's understanding of the world is general enough to predict verb-noun pairs for unseen actions, rather than just unseen images.

5.5 GloVe Synonym Accuracy Threshold

Similar to the issues in support vector machines, finding the boundary that separates classes was difficult to universally define. In our case, the boundary between synonyms and non-synonyms could not be clearly defined. In the end, we decided that placing separate boundaries for the verb and the noun was the most accurate way to define synonyms. Our final verb threshold was 5.3, and our final noun threshold was 4.5. These numbers are arbitrary with analysis of the embeddings in the dataset. Therefore, these numbers were created by analyzing hand selected synonyms like "ignite" and "burn." While we found a boundary that we were confident in, we understand that this boundary is not absolute. For example, the GloVe distance between "throw" and "break" was below the 5.3 mark even though these are clearly not synonyms. For that reason, we understand this metric has its flaws, and that is the reason we see increases in accuracy in Figure 7 though the point of Dataset 2 was to not require this metric. In sum, GloVe distance analysis highlighted a lot of key issues with the dataset and allows the model to be correct without being exact, but could use some further tuning.

Caption: "a whisk in a stainless steel bowl of pumpkin pie batter"

Caption: "stock image of close up of sliced ham on a white plate"

Figure 9: Example of poor results obtained by OFA pretrained captioning. The top example showcases incorrect information and the bottom example showcases unecessary, noisy details.

6 Conclusion

As the field of AI continues its incredible pace of achievements, we will continue to see robots operate in the physical world. Of course, they will need to perceive their environment which has been a prominent field of research. We see reliable understanding through image captioning tasks in large-scale multi-modal models like OFA. However, an arguably more important task for robots is to understand the consequences of their actions in their environment. Agents with understanding in both of these tasks will be able to understand the state of their environment and map out the effects of the available actions so that they can achieve their goals. While the first part of this task has been researched much more thoroughly, we contend that this research is not mutually exclusive from the second part of this task.

A model like OFA is the perfect candidate for utilizing state-of-the-art research in image-to-text tasks. With its top-of-the-line performance in understanding the setting of an image and its emphasis on multi-modality, this model was the best contender to demonstrate a possibility for an agent to have complete physical causality understanding through one architecture.

This paper presents an initial exploration of how this can be possible. There were many challenges and possible explanations for some of the shortcomings of our research. Still, the results and insights gained from this paper warrant more research in this area and in the idea of leveraging extensive research in related tasks. Further focus in these areas help work toward a world with agents who can perceive their environment and accurately map their actions to the best state of the physical world.

7 References

[1] P. Wang, A. Yang, R. Men, *et al.*, "Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework," *CoRR*, vol. abs/2202.03052, 2022. arXiv: 2202 . 03052. [Online]. Available: https://arxiv.org/abs/2202.03052.

[2] Q. Gao, S. Yang, J. Chai, and L. Vanderwende, "What action causes this? towards naive physical action-effect prediction," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Melbourne,

Australia: Association for Computational Linguistics, Jul. 2018, pp. 934–945. DOI: 10.18653/v1/P18-1086. [Online]. Available: https://aclanthology.org/P18-1086.

[3] L. Kaiser, A. N. Gomez, N. Shazeer, *et al.*, "One model to learn them all," *CoRR*, vol. abs/1706.05137, 2017. arXiv: 1706.05137. [Online]. Available: http://arxiv.org/abs/1706.05137.

[4] R. Kiros, R. Salakhutdinov, and R. Zemel, "Multimodal neural language models," in *Proceedings of the 31st International Conference on Machine Learning*, E. P. Xing and T. Jebara, Eds., ser. Proceedings of Machine Learning Research, vol. 32, Bejing, China: PMLR, 22–24 Jun 2014, pp. 595–603. [Online]. Available: https://proceedings.mlr.press/v32/kiros14.html.

[5] A. Karpathy, A. Joulin, and L. Fei-Fei, *Deep fragment embeddings for bidirectional image sentence mapping*, 2014. DOI: 10.48550/ARXIV.1406.5679. [Online]. Available: https://arxiv.org/abs/1406.5679.

[6] R. Kiros, R. Salakhutdinov, and R. S. Zemel, *Unifying visual-semantic embeddings with multimodal neural language models*, 2014. DOI: 10.48550/ARXIV.1411.2539. [Online]. Available: https://arxiv.org/abs/1411.2539.

[7] Y. Wang, J. Xu, and Y. Sun, *End-to-end transformer based model for image captioning*, 2022. DOI: 10.48550/ARXIV.2203.15350. [Online]. Available: https://arxiv.org/abs/2203.15350.

[8] Y. Luo, J. Ji, X. Sun, *et al.*, *Dual-level collaborative transformer for image captioning*, 2021. DOI: 10.48550/ARXIV.2101.06462. [Online]. Available: https://arxiv.org/abs/2101.06462.

[9] A. Jaegle, S. Borgeaud, J.-B. Alayrac, *et al.*, *Perceiver io: A general architecture for structured inputs amp; outputs*, 2021. DOI: 10.48550/ARXIV.2107.14795. [Online]. Available: https://arxiv.org/abs/2107.14795.

[10] J. Pennington, R. Socher, and C. Manning, "GloVe: Global vectors for word representation," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Doha, Qatar: Association for Computational Linguistics, Oct. 2014, pp. 1532–1543. DOI: 10.3115/v1/D14-1162. [Online]. Available: https://aclanthology.org/D14-1162.