Intelligent Human Systems Integration (IHSI 2023), Vol. 69, 2023, 359–366
https://doi.org/10.54941/ahfe1002855

Conceptual Exploration of Contextual Information for Situational Understanding

Stratis O. Aloimonos and Adrienne J. Raglin

U.S. Army Research Laboratory (ARL), Adelphi, MD 20783, USA

ABSTRACT

The Army is often required to deploy soldiers into dangerous situations to offer assistance and relief. When deployed, these soldiers need to be aware of the potential dangers, properly assess the level of possible threats, and make the best choices to respond. One solution for this problem space is to have an intelligent system that recognizes scenes which may contain danger, regardless of the type or timeframe associated with that danger. This type of system would help make decisions about what to do in situations where danger may be prevalent. Thus, creating an intelligent system that could identify the scene and contextual information, for example, potential dangers, would provide greater situational understanding and support autonomous systems and solider interactions. As a proxy for representing scenes that may be similar to those encountered by soldiers, a set of images of natural or manmade disasters were selected and used to identify strengths and weaknesses in existing models for this type of intelligent system. In this work, images from CRISISMMD, a dataset of natural disasters tweets, as well as other images of disasters in the public domain which do not belong to any particular dataset, are used. For the initial phase of the work this dataset was used to determine and showcase the strengths and weaknesses of existing object recognition and visual question answering systems that when combined would create a prototype intelligent system. Specifically, YOLO (You Only Look Once), augmented with Word2Vec (a natural language processing (NLP) system which finds the similarities of different words in a very large corpus) was selected for performing the object recognition (Bochkovskiy et al. 2020). This system was selected to identify objects further based on the presence of other, similar objects using the similarities between their names. Also, CLIP (Contrastive Language Image Pretraining), which identifies the probabilities of scenes based on a certain number of possibilities and BLIP (Bootstrapping Language Image Pretraining) (Li et al. 2022), an advanced visual question answering system which is also capable of generating captions for images were explored. In addition, a concept of an intelligent system where contextual information is identified and utilized can be used to support situational understanding.

Keywords: Scene understanding, Situational understanding, Object recognition

INTRODUCTION

Soldiers are often deployed into dangerous situations and environments. These situations and environments may be in a tactical setting or a humanitarian setting. To inform these soldiers of the challenges they may

face in these settings, images and videos would be useful. For images of the scenes require analysis whether that is done by a system, by a person, or a combination. Ideally, the analysis would uncover signs of hidden dangers or indications of damage or dangers. There are existing detection systems, as well as visual question answering (VQA) systems, which can be trained to recognize certain objects, or scenes from a library of many possibilities. For example, systems such as YOLO, an object detection system that performs with high accuracy across various images, CLIP, which can be used to recognize scenes, in addition to objects, and BLIP, which can be used for VQA, to answer questions about the scene depicted in an image, as well as create captions which describe that image. These methods all have their own strengths and weaknesses when applied to specific images that may be more relevant to specified tasks. The point of this research was to investigate these methods, and identify these strengths and weaknesses, to see if they, or versions or combinations of them, would be good to analyze images that are proxies for scenes soldiers may encounter.

Methodology

Images from the dataset CRISISMMD were selected. CRISISMMD is a multimodal dataset of images of disasters, or the damage caused by disasters, and accompanying tweets (CRISISMMD: Multimodal Crisis Dataset). In addition to these, several images of disasters from the internet which do not belong to any particular dataset were also used. These images were analyzed with YOLO, CLIP and BLIP, in various ways. YOLO was augmented with an NLP system called Word2Vec (Rong 2016). Whenever YOLO returned more than one probability for a detected object, Word2Vec was used to compare this object to all other objects in the image to help identify the additional objects that are likely to be in similar images. CLIP was used to compare several images of disasters to each other, to determine what kind of disaster they were and how severe the disaster. BLIP was used to ask questions related to disasters, as well as identify if the image was categorized as a disaster.

Experiments and Results

For the object detection system YOLO, the results were bounding boxes, as well as the probability of the label given to the object for each image analyzed. If there were any objects which YOLO assigned multiple labels to, or assigned labels with a significantly low probability, Word2Vec was used to compute the similarities. Specifically, Word2Vec finds the similarities by calculating cosine similarity in the vector space. The vector space represents various words. In this case, it found the similarity between each label of the category of potential objects (any object which could have been more than one label with probability greater than 0.00) with the label of every object the detector assigned only one probability to. If the image contained more than one object which the detector gave more than one label and associated probability for, then the similarity of each possible label of every potential object with every possible label of every other potential object was also calculated.

Once these similarities were found, the probabilities for all possible combinations of categories were then calculated. This was done by multiplying the probabilities of each known object occurring with the products of probabilities of each potential object, and then multiplying this product with the product of similarities between different potential objects. Once all the products of probabilities were found, the largest combined product, corresponding to the most likely combination of objects, was selected as the correct one. Then used to assign categories to the potential objects which the system did not identify previously.

Thus, PX represents the probability of a detected object, X. When an object is assigned more than one label and associated probability, the probability is represented by PX_i. For example, in an image two objects were assigned one label with a single associated probability, P1 and P2. Also, two objects were assigned multiple labels with associated probabilities, $P3_1$, $P3_2$ and $P4_1$ and $P4_2$.

Each object has a similarity to every other object. Thus, SXY represents the similarity between detected objects X and Y. If one object is assigned multiple labels with associated probabilities, SXY_j represents the similarity between object X and label j for object Y. If both objects were assigned multiple labels with associated probabilities, SX_iY_j represents the similarity between label i for object X and label j for object Y.

For example, for the objects in the previous example, the similarity between the object whose probability is P1 and the label of the object represented by $P3_1$ is $S13_1$. The similarity between the object whose probability is P1 and the label of the object represented by $P3_2$ is $S13_2$. The other similarities between the objects with only one assigned label and probability to those with multiple labels and probabilities are $S14_1$, $S14_2$, $S23_1$, $S23_2$, $S24_1$ and $S24_2$. The similarities between the objects with more than one label and assigned probability are $S3_14_1$, $S3_14_2$, $S3_24_1$ and $S3_24_2$.

From these probabilities and similarities, the choices for a probability that each object the detector is unsure of is calculated as follows:

(1) For this option, $P3_1$ and $P4_1$ are selected:
$$P(P3_1 \text{ and } P4_1) = P1*P2*S13_1*S23_1*P3_1*S14_1*S24_1*P4_1 * S3_14_1$$
(2) For this option, $P3_1$ and $P4_2$ are selected:
$$P(P3_1 \text{ and } P4_2) = P1*P2*S13_1*S23_1*P3_1*S14_2*S24_2*P4_2*S3_14_2$$
(3) For this option, $P3_2$ and $P4_1$ are selected:
$$P\left(P3_2 \text{ and } P4_1\right) = P1*P2*S13_2*S23_2*P3_2*S14_1*S24_1*P4_1*S3_24_1$$
(4) For this option, $P3_2$ and $P4_2$ are selected:
$$P\left(P3_2 \text{ and } P4_2\right) = P1*P2*S13_2*S23_2*P3_2*S14_2*S24_2*P4_2*S3_24_2$$

When all these products of probabilities are calculated, the highest product is selected, and the labels corresponding to the choices of object for the objects the detector was unsure of are selected as the true labels of these objects. Two examples of this are shown in Figure 1 below.

For the scene identification system CLIP, several images of disasters were uploaded into the system, and compared with four categories, specifically earthquakes, fires, floods, or anything else (to create a group for outliers). In general CLIP did very well in recognizing which of these categories were

Figure 1: Two images put through YOLO, where the detector is not sure about some of the objects. The first is a few emergency trucks driving through a street to offer assistance, where the detector is confused about the identity of the final two trucks. The second image is police cars at a roadside at night, where the detector is confused about the identity of the last car, closest to the right side of the image. Taken from internet.

represented by the images. These images were then grouped by different categories related to the level of danger. These categories were major damage, moderate damage, minor damage or other to capture outliers. When examined, the system did reasonably well, although some categories were different from those anticipated. For instance, below, in Figure 2, one car accident is believed by the system to be an earthquake, when it should be "other". In another image, a different car accident is believed to be a fire, likely due to the presence of an emergency vehicle in the image. In addition, it is unclear how the system "defines" what constitutes, a lot, a medium amount, or very little damage than a human might, as shown in Figure 3.

For the VQA and captioning system BLIP, several images were analyzed based on the answers to pre-selected questions. The questions were *(1) What is this? (2) Is this a disaster? (3) What type of disaster is this? (4) Is this dangerous? (5) How dangerous is this? (6) Has the danger passed? (7) Does help need to be sent in? (8) Are there emergency vehicles in this image? (9) Should there be emergency vehicles in this image?* The responses to the questions for each image were manually recorded. In addition, BLIP was also used to create captions for these images, i.e. a short sentence describing what was in the image. This was done three times for each image, since BLIP gave a different caption for each request.

For the image in Figure 4 the results were as follows:

Captions:

People are walking down a flooded street carrying a boat, the water is moving quickly down this flooded street, people are wading across a flooded street.

Answers to Questions:

(1) **What is this?** Answer: Flood
(2) **Is this a disaster?** Answer: Yes
(3) **What type of disaster is this?** Answer: Flood
(4) **Is this dangerous?** Answer: No
(5) **How dangerous is this?** Answer: Not very
(6) **Has the danger passed?** Answer: Yes

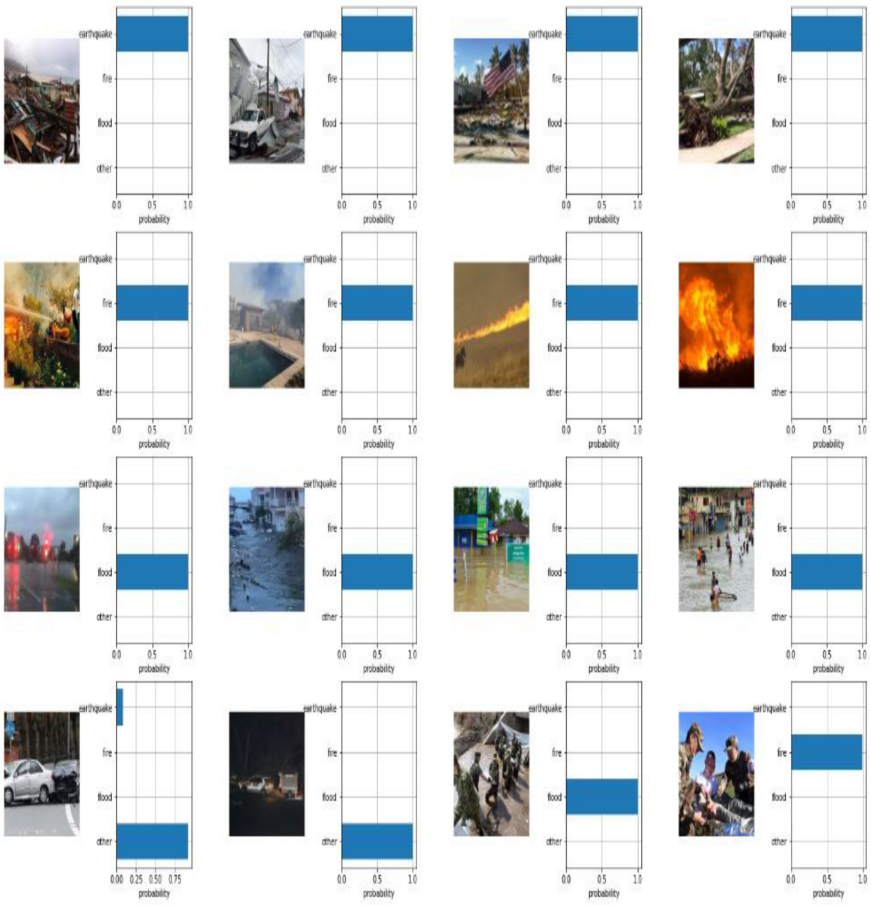

Figure 2: Sixteen images put through CLIP, compared between earthquake, fire, flood and other. The images that would be considered "other" are images of car crashes. These images are taken from internet and CRISISMMD.

(7) **Does help need to be sent in?** Answer: Yes
(8) **Are there emergency vehicles in this image?** Answer: No
(9) **Should there be emergency vehicles in this image?** Answer: Yes

CONCLUSION

The results for each system highlighted their strengths and weaknesses.

For YOLO augmented with Word2Vec, the system was good at recognizing objects. However, the system did not usually identify indications of danger that was shown in the background. Since background information can be helpful for identifying the context of the scene and associated objects, this would be a helpful capability for systems.

In addition, the system had trouble identifying the objects that were either partially obscured by darkness in the image and smaller objects in the background.

Also, systems such as Word2Vec are trained on a corpus that is not tailored to specific jobs as those found in the military, proxies were manually

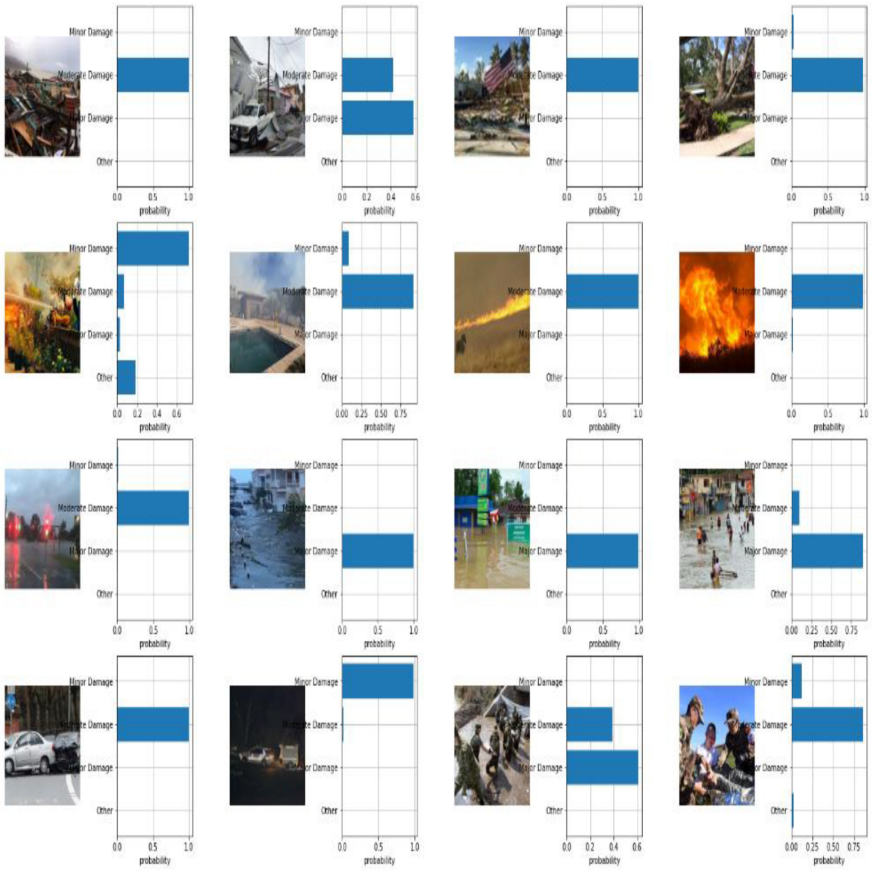

Figure 3: The same sixteen images put through CLIP, now compared between major damage, moderate damage, minor damage and other. These images are taken from internet and CRISISMMD.

Figure 4: An image of a flooded street in India. Taken from CRISISMMD Dataset.

generated by a person. The ability for systems to adapt to these types of specific requirements would be another useful capability.

For CLIP, while the system worked well in identifying the overall scene, it did not identify the dangers within the scene. Having a system that could provide multiple levels of information would be helpful.

In addition, the system only compared a few categories, limiting the labeling for the images and thus reducing the number of possible scenes the image could be identified as. With more possible identifiers for scenes, the image identification could be more accurate.

For BLIP, while it was able to answer the questions, and give useful information relevant to the images, it did have some limitations. For some images, to get an answer which related to the question asked, one had to phrase the question in a specific way. Also, different answers were given when the question was phrased differently. For example, for certain images, when the question "*For what reason is this dangerous?"* was asked, the response was *"fire"*, This was different than the anticipated answer, which was expected to be something relating to either an earthquake or hurricane. When the question "*Why is this dangerous?"* was asked of the same image, the answer was *"it's dangerous"*.

For another example, both the questions above were asked of another image of hurricane damage. The answer for both was *"flooding"* despite barely any water being in the image.

For another, which was assumed to be damage after a storm, the question *"Why is this dangerous?"* returned as an answer *"falling down"* while the question *"For what reason is this dangerous?"* returned *"flooding"*.

In addition, the captioning system gave a different response to the same image for multiple runs. In general, these captions were similar. For example, one of the flood images was stated to be *"People are wading across a flooded street"*, while another one was *"The water is moving quickly down this flooded street"*. For another, one of the fire images was stated to be *"A fire burns near the back of a home"*, while another caption of this image was *"The house on fire has been burnt by smoke"*.

Future Directions

Based on the outcome of this initial analysis key features from each system would be helpful for a future system that combined these capabilities, particularly to identify the dangers within already identified scenes. Perhaps if these systems were run in series, the information gained from the combination would be much more useful than each one run separately. As this work continues, additional systems will be investigated. One system might be MMF (multimodal framework), a captioning system originally developed for Facebook, that contains several vision and language models. In addition, a system from the University of Maryland, which employs the dataset RIVAL10, may also be investigated to generate and visualize representations of images with saliency alignment (Feizi et. al. 2022).

REFERENCES

Announcing MMF: A framework for multimodal AI models, *MetaAI*. https://ai.facebook.com/blog/announcing-mmf-a-framework-for-multimodal-ai-models/. Accessed Sept. 19, 2022.

Bochkovskiy, Alexey; Wang, Chien-Yao.; and Liao, Hong-Yuan Mark. 2020. *YOLOv4: Optimal Speed and Accuracy of Object Detection*.

CRISISMMD: Multimodal Crisis Dataset. CRISISNLP, https://crisisnlp.qcri.org/crisismmd. Accessed Sept. 19, 2022.

Feizi, Soheil; Moayeri, Mazda; and Banihashem, Kiarash. 2022. *Explicit Tradeoffs between Adversarial and Natural Distributional Robustness*. NeurIPS 2022.

Ferda Ofli, Firoj Alam, and Muhammad Imran, Analysis of Social Media Data using Multimodal Deep Learning for Disaster Response, In Proceedings of the 17th International Conference on Information Systems for Crisis Response and Management (ISCRAM), 2020, USA.

Firoj Alam, Ferda Ofli, and Muhammad Imran, CrisisMMD: Multimodal Twitter Datasets from Natural Disasters, In Proceedings of the 12th International AAAI Conference on Web and Social Media (ICWSM), 2018, Stanford, California, USA.

Li, Junnan.; Li, Dongxu.; Xiong, Caiming.; and Hoi, Steven. 2022. *BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation*.

Rong, Xin; 2016. *Word2Vec Parameter Learning Explained*.