

There’s a Time and Place for Reasoning Beyond the Image

Xingyu Fu¹, Ben Zhou^{1*}, Ishaan Preetam Chandratreya^{2*}, Carl Vondrick², Dan Roth¹

¹University of Pennsylvania

²Columbia University

{xingyuf2, xyzhou, danroth}@seas.upenn.edu

{ipc2107, cv2428}@columbia.edu

Abstract

Images are often more significant than only the pixels to human eyes, as we can infer, associate, and reason with contextual information from other sources to establish a more complete picture. For example, in Figure 1, we can find a way to identify the news articles related to the picture through segment-wise understandings of the signs, the buildings, the crowds, and more. This reasoning could provide the time and place the image was taken, which will help us in subsequent tasks, such as automatic storyline construction, correction of image source in intended effect photographs, and upper-stream processing such as image clustering for certain location or time.

In this work, we formulate this problem and introduce TARA: a dataset with 16k images with their associated news, time, and location, automatically extracted from New York Times¹ (NYT), and an additional 61k examples as distant supervision from WIT (Srinivasan et al., 2021). On top of the extractions, we present a crowdsourced subset in which we believe it is possible to find the images’ spatio-temporal information for evaluation purpose. We show that there exists a 70% gap between a state-of-the-art joint model and human performance, which is slightly filled by our proposed model that uses segment-wise reasoning, motivating higher-level vision-language joint models that can conduct open-ended reasoning with world knowledge. The data and code are publicly available at <https://github.com/zeyofu/TARA>.

1 Introduction

Vision and language are two of most important information sources, and the fact that humans reason jointly with both sources has motivated artificial intelligence research to consider visually-grounded



Figure 1: This is an image from the New York Times. Can you tell the time and location when it was taken?

language understanding. Most work in this area has focused on reasoning with local evidence (Suhr et al., 2019; Hudson and Manning, 2019; Lu et al., 2020; Liu et al., 2021), e.g. asking factoid questions such as the colors or shapes of objects and numbers of people, yet very few works (Cui et al., 2021) encourage open-ended reasoning where a model needs to look beyond task inputs. However, humans can relate visual cues to corresponding contextual information that could be multi-modal, and draw on background knowledge when interpreting and grounding images. For example, as Figure 1 shows, people that are familiar with the news can infer that the location is Times Square through the iconic screen panels, and further estimate the period of time by looking at the crowds and the signs. And, this can be done without explicitly including related news pieces as input. In fact, even though some people would not have the prior knowledge to identify the relevant events, it is likely that they would have good estimate of the location and time by interpreting textual evidence in the image, the language in the signs, entity names, building styles, and other details in the input image.

In this work, we identify and formulate this problem, spatio-temporal grounding of images, a

* Both authors contributed equally to this work.

¹<https://developer.nytimes.com/docs/archive-product/1/overview>



Figure 2: What is the time and location for this image?

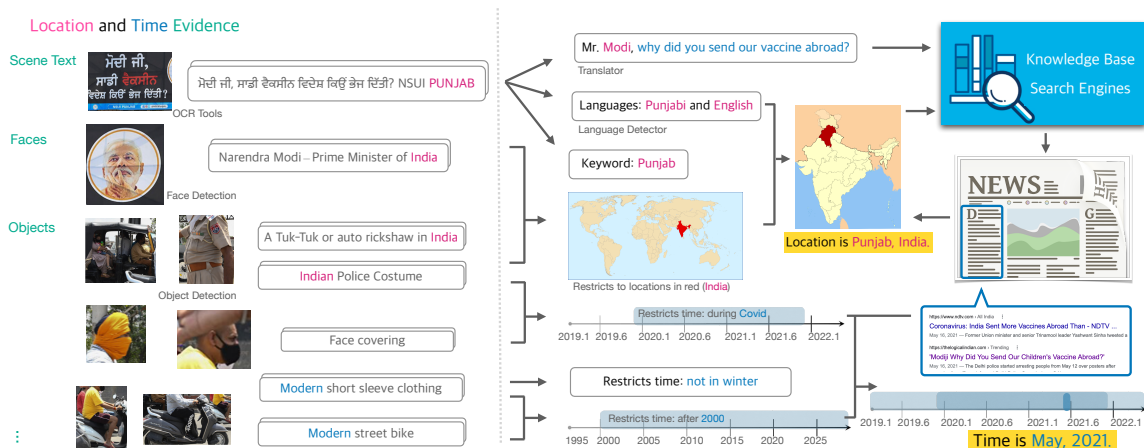


Figure 3: An example of potential joint reasoning on Figure 2 to ground its time and location. Note that people with different backgrounds may need to use different levels of reasoning, resulting in a completely accurate or just partial grounding (e.g., the decade and country), and we only show one such reasoning route. We start with grounding multiple scene text, faces, and objects segments from the image, and use the information to conduct a constrained search in a large news-base, until it locates specific textual information related to the image.

task aiming at identifying the *time* and *location* in which the given image was taken. Specifically, we develop a novel dataset TARA, (Time and plAce for Reasoning beyond the imAge), a challenging dataset that tasks models with grounding images to real-world spatial and temporal information. In our collection, we make sure that for models to accurately find images’ creation time and location, they would need to successfully ground the visual clues in texts such as news, stories and encyclopedias. As a result, this task motivates models to consider the association between visual information, language, and background knowledge, more closely and in a more open-ended setting. Figure 2 shows an example from TARA, and Figure 3 shows a possible way for a model to ground the image to its spatio-temporal information. The system starts with grounding multiple segments from the image, and uses the information to conduct a constrained

search in a large news-base, until it locates specific textual information related to the image. This demonstrates the complexity and significance of this task.

TARA is collected via a rigorous process that involves rule-based distant supervision extraction from news-images data which results in 16k image examples. While the training data has high label correctness (around 95%), we further run a crowdsourced validation on 3k examples to form the evaluation dataset. During the validation, annotators are asked to verify that there exists a potential path for humans to derive the correct answer, which encourages proper reasoning in future works. To better support the study of domain transfer and supervision for the reasoning process, we collect an additional 61k examples from the Wikipedia domain. We apply the state-of-the-art joint model CLIP (Radford et al., 2021) and show that it only

achieves accuracy of 11.11% and 0.46% for time and location, respectively, on our dataset.

Additionally, we present a new CLIP-based baseline model that reasons on object and facial segments and achieves 16.46% and 1.07% accuracy for time and location, respectively. We show that there exists a large gap (around 70% in accuracy) between state-of-the-art models and human performance, suggesting that the TARA data will provide a benchmark to motivate reasoning based approaches and support significant future work.

2 Related Work and Datasets

Vision and Language Learning Language understanding in the context of images has been widely studied in various datasets covering a wide range of tasks including visual question answering, image retrieval, image and video captioning, etc. Earlier datasets mostly focus on simple local object properties identification (Antol et al., 2015; Chen et al., 2016). Later on, datasets start to focus on compositional visual reasoning. For example, Suhr et al. (2017) and Johnson et al. (2017) use synthetic images or synthetic language to study spatial relations. Recently, datasets using real images and real languages such as (Hudson and Manning, 2019; Liu et al., 2021) were proposed for reasoning about natural language descriptions of photos. However, all of the datasets focus on local grounding on segments inside the image, but not globally ground beyond the image with open-ended reasoning.

While there are various tasks and datasets, the underlying associations between language and visual concepts are often common across different tasks (Lu et al., 2020). Therefore, we use CLIP (Radford et al., 2021) to study the TARA dataset in this paper. CLIP is a recently released state-of-the-art image representation model which has shown impressive performance on various tasks through pre-training on 400 million image and captions pairs collected from the internet.

Spatio-temporal IE from Texts There has been extensive work on identifying temporal expressions and their associations with events in texts. Uz-Zaman et al. (2013); Ning et al. (2018) focus on temporal information extraction within the local contexts, and Zhou et al. (2020, 2021) further extends the scope to consider contextual information from external texts. The NLP community has also investigated spacial information extraction, with geocoding (Gritta et al., 2018; Kulkarni

et al., 2020), which maps mentions to geological coordinates, being closest to our scope.

3 Dataset Collection

Each example in TARA includes a news image, along with its time, location, caption, and corresponding news background such as headline, abstract, and news type. These are included for training or analysis purposes, but the task is to guess the correct time and location as accurately as possible given only the image. In developing the dataset, our goal is to collect a large corpus of semantically rich images that human with world knowledge can correctly identify the time and location, using evidence from the image, background knowledge, and appealing to external knowledge (which we call "reasoning" here). We design the process of collecting and identifying the images so that it facilitates this type of reasoning, and then use crowd sourcing to label a random 20% of high-quality images for development and testing. Figure 4 illustrates our data collection procedure.

3.1 Image collection

We first collect all the news between January 2010 and May 2021 using the NYT API². We did not collect news that are earlier than 2010 because earlier news articles contain much fewer images. Each news article comes with a list of attributions³ such as headline, abstract, news type, and possibly a main image. We first filter the news articles that has a valid image, and then scrape image caption for each image. Since the NYT covers news in several multimedia formats, the images follow a range of formatting practices, such as representative news images, image collages, images sampled from slideshows and descriptive natural thumbnails for videos. We setup a NYT specific pipeline to scrape image captions. We define a separate scraping procedure to get image specific text information for the different media types mentioned above and remove instances where multiple and/or ambiguous captions are returned.

Image Pruning and Labeling Next, we describe how we automatically collect time and location of an image from corresponding news articles and captions. First, we filter out the images with

²<https://developer.nytimes.com/docs/archive-product/1/overview>

³For each news, the API provides attributes as listed here: <https://developer.nytimes.com/docs/archive-product/1/types/Article>

(a) **Collect Images:** We collect all the news with images from NYT between January 2010 and May 2021, and crawl captions.

Search: 2010-1, 2010-2, ..., 2021-5

The New York Times Developers

Captions

Headlines, Date, Section ...

P1: The National Guard on Springfield Avenue in Newark on July 14, 1967.
P2: The seven members of the Communist Party's ... with President Xi Jinping, in Beijing in 2016.
P3: Afghanistan's vice president Gen. Abdul Rashid Dostum, had been forced into exile this year.
P4: Bahia Amawi's contract as a speech pathologist in Texas was not renewed this year.
P5: Hossein Nayeri was led back to jail on Sunday after he was recaptured.

2017-07-11: Five Days of Unrest That Shaped, and Haunted, Newark: ...
2017-07-16: As China Prepares for New Top Leaders... ; Foreign; ...
2017-08-18: Afghanistan, a Destructive Game of Thrones; Foreign; ...
2018-12-19: She Wouldn't Promise Not to Boycott Israel; National; ...
2016-02-02: Fight Between 2 California Escapees; National; ...

(b) **Image Pruning and Labeling:** We use NER models to prune images using captions, and assign possible time and location labels. In this example, the last image is removed because there is not location detected in the caption.

1967-7-14	2016	2017	2018	2016-1-31
Newark, Essex County, New Jersey, United States, North	Beijing, Dongcheng District, Beijing, 100010, China, Asia	Afghanistan, Asia	Texas, United States, North America	None

(c) **Validation:** Crowdworkers are given only the images. They need to judge whether human, e.g. local people, can guess the time and location of the image, without searching online. If positive, we adjust label to the majority hierarchy; otherwise, the image will be removed.

Yes — Year	Yes — Date	Yes — Year	No
Yes — Exact Location	Yes — Exact Location	Yes — Exact Location	No

(d) **Validation:** Crowdworkers are given the image, news headline, possible labels, and possible main event we extracted from news. They decide whether the possible labels are correct. In this example, the majority workers think the time label is wrong, so we use Null as label.

1967	2016	2016 Null
Newark, Essex County, New Jersey, United States, North America	Beijing, Dongcheng District, Beijing, 100010, China, Asia	Afghanistan, Asia

Figure 4: Data collection process. Steps (a)–(b) are described in Section 3.1; and steps (c)–(d) in Section 3.2.

unwanted news types such as reviews, series, and obituaries, and unwanted news topics such as food, fashion, and movies, because images from these articles may not be informative enough. Then, we filter out the images whose caption does not contain location and time. For those that contain temporal and spacial cues, we assign each image a possible time label and location label. Specifically, we use the Spacy NER model⁴ to find if the caption has both exactly one “DATE” entity for time and one “GPE” or “LOC” typed entity for location. Note that each news comes with a publication date and possible locations in attributes. We would either directly use our NER-extracted time entity as the possible time label if it’s a valid time, or adjust the publication date using the time entity. For example, if the time entity is “1936” and publication date is “2021-05-01”, then we will use “1936” as the possible time label because it should be an old image occurring in a recent news; in the latter case, if the

⁴<https://spacy.io/models/en>

time entity is “last month” and publication date is “2015-07-18”, then we will use “2015-06” as the possible time label. We also compare our NER-extracted location entity with the news attribute locations. If the only difference is granularity, e.g. one is New York, United States and the other is United States, then we will use the fine-grained one “New York, United States” as possible location label. Otherwise, we will filter out this image.

Finally, we add missing hierarchies for each possible label. For time labels, we add the decade and the century. For location labels, we use Geopy⁵ to identify the location and add missing hierarchies such as country and continent.

3.2 Validation

We randomly select an equal number of images from each month, such that a total of about 20% images are assigned to development and test. On

⁵<https://geopy.readthedocs.io/en/stable/>

these images, we use two crowdsourcing tasks to (1) prune unanswerable images, and (2) verify correctness of the labels.

In the first task, we display a single image, and ask a worker to answer, without searching online, if any person can guess the time and location of the image. We offer different hierarchies in the choices – date, year, decade, and century for time and exact location, city, country, and continent for location – so that workers can choose one of these. If the majority of workers agree that human cannot reason time or location based on the image itself, we will mark the corresponding label as null. Otherwise, if the majority of them agree on a certain hierarchy, we adjust the possible label to that specific hierarchy. Check step(c) in Figure 4 for criteria and positive and negative examples.

The second task further verifies the correctness of current time and location labels. Specifically, we provide the same image, but including its caption, news headline, abstract, and extracted time and location labels. We ask the workers to verify if the background event is the same as in image, and if the labels are correct after reading the additional information. We use the Semantic Role Labeling (SRL) model⁶ from AllenNLP to detect the main verb in the image caption by selecting the verb with most arguments, and mark it as the possible main event to provide to the workers. Detailed examples can be found in step(d) in Figure 4.

3.3 Test Set of Interest

We further select a small set of 30 interesting images as shown in Figure 5, that are related to most famous news happening after January 2021, the CLIP model date.⁷ This adversarial test set is specifically chosen to cover unseen images by baseline models to better test their generalization instead of memorization.

Additionally, regarding to human baseline, annotators need to have enough knowledge to extract and interpret the key evidence segments, in order to reason about the answer. For instance, a person with an American cultural background and speaks English but not Hindi may find Figure 1 is easier to infer the precise time and location than Figure 2, compared to a person with Indian cultural background and speaks Hindi but not English, and vice

⁶<https://demo.allennlp.org/semantic-role-labeling>

⁷<https://github.com/openai/CLIP/blob/main/model-card.md>

Dataset	Train	Dev	Test	All
TARA before validation	12,306	1,644	1,644	15,652
TARA	12,306	1,552	1,571	15,429
WIT				61,325

Table 1: Dataset statistics for TARA and additional WIT supervision.

versa. This test set of interest is chosen to cover most well-known news for the purpose that human baseline annotators are more likely to have enough knowledge about the key evidence so that the comparison with neural models can be more fair.

3.4 Additional Weak Supervision

We apply the same image pruning and labeling procedures on the WIT dataset (Srinivasan et al., 2021), which contains 11.5M Wikipedia images and the surrounding paragraphs and captions. Since this dataset is much unorganized, we only select images in English Wikipedia articles, and apply two additional NER models (Lample et al., 2016; Peters et al., 2017) from AllenNLP⁸ to select locations. We further use zero-shot CLIP model to prune unwanted image types. Specifically, we provide each image with text sentences in the format of “a photo of [type]”, with *type* being *photograph*, *map*, *paint*, and *paper*, and retrieve the sentence with highest similarity score. We only keep images of type *photograph*, and use these as additional weak supervision. The benefit of adding this additional weak supervision is that it has a wider range of time and location labels than the NYT images, especially because that all the NYT images are taken from news between 2010 and 2021.

4 Dataset Analysis

4.1 Dataset Statistics

Dataset statistics can be found in Table 1. TARA contains about 16K images from New York Times. After crowd-sourcing validation on development and testing, about 94% of the images that either has a valid location label or time label are kept, indicating that our training set can serve as a good weak supervision. In addition, TARA provides a 61K weak supervision dataset built upon WIT.

⁸<https://demo.allennlp.org/named-entity-recognition/fine-grained-ner>



Figure 5: Some example images in our test set of interest as described in Section 3.3. These very recent images require open-ended reasoning with world knowledge and are specifically chosen such that our human baseline annotators probably have enough knowledge about the key evidence. For example, in the first image, people need to know what “BLM” is so that they can start to search statues in United States. Also in the second image, people need to know it is the President Biden for further reasoning.

4.2 Time and Location Distribution

Figure 6 shows the time and location distribution in TARA. We can see that most images are taken in North America, Asia, and Europe, between 2010 and 2021. This can be the effect of using NYT as image source.

5 Baselines

We assess the quality of our dataset through human annotation, and evaluate on existing visual reasoning approaches.

5.1 Human Performance

As introduced in Section 3.3, an expert annotator works on our test set of interest to gain a better understanding of the human performance on TARA. The expert is not allowed to directly search the image online, but can search for anything else such as the keywords she/he infers from the image. The expert is presented with all the labels in the test set just as neural models.

5.2 Evaluation Systems

We use the state-of-the-art systems in machine reading comprehension for this task: **CLIP** (Radford et al., 2021). CLIP is the state-of-the-art image representation model and has shown impressive progress on visually grounded language understanding tasks. Specifically, we use the “ViT-B/32” model⁹ for zero-shot classification and analysis. During prediction, the model is given a single image and needs to classify the correct label. We use a similar prompt template “A photo taken in {label}.” following the original paper, to encode all the labels. We compare the similarity between the image and each label prompt, and the highest one is the predicted label.

We also add several variants of CLIP. The first is **CLIP+**, which is the zero-shot CLIP model finetuned on NYT training data. Note that CLIP uses contrastive loss to train on image and text pairs. We concatenate the time and location labels into a

⁹<https://github.com/openai/CLIP>

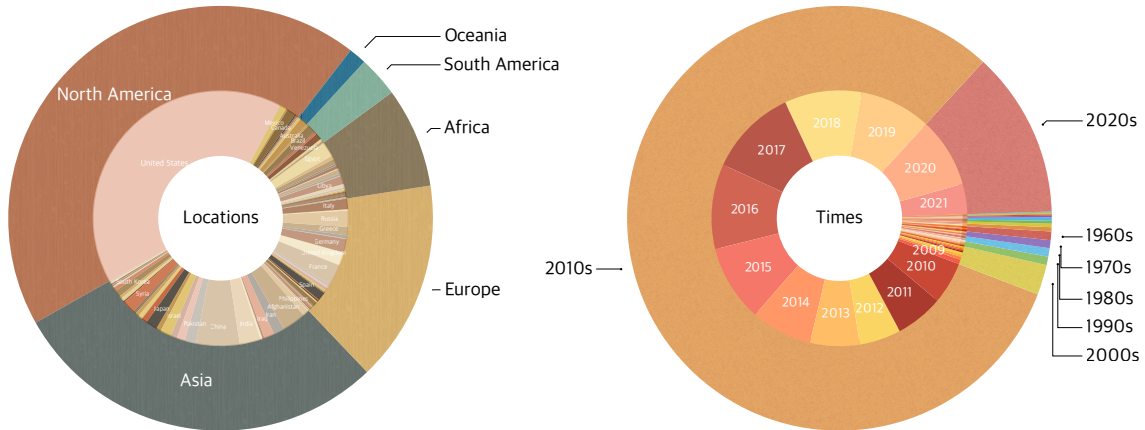


Figure 6: Label distribution in TARA. All of the training, development, and testing data are considered.

natural language sentence to serve as the text part for an image.

CLIP+Seg is another variant where we first extract object and face segments, and then finetune the CLIP model on the whole images along with the segments, both with time and location labels concatenated together as the final goal. As for object detection, we use the YOLOv5¹⁰ method, specifically with model “yolov5s”. The intuition is that for objects such as iPhone, the model benefits from training it to times later than 2010. We add a limit to the segments so that we only consider important objects that have size larger than 50. We further restrict the number of people segments to be no more than 3, since many of the images have crowds and adding more people do not bring in much additional information. As for face segments, we use the InsightFace (Guo et al., 2022) facial detection model¹¹. The intuition is that for famous people such as President Biden, we will benefit from training the segments to location “United States”. During implementation, we also add a limit to the segments so that we only consider face that have size larger than 50, which are more likely to be most important faces.

CLIP+WIT is the variant of CLIP where we finetune on the training images along with the 61K weak supervision Images extracted from WIT. We concatenate the possible time and location labels as the paired text.

¹⁰<https://github.com/ultralytics/yolov5>

¹¹<https://github.com/deepinsight/insightface>

6 Experiments and Results

6.1 Evaluation metrics

Two metrics are adopted in this work: Accuracy and **Example-F1** (also known as micro-Dice coefficient) following previous studies (Shen et al., 2021). Accuracy is calculated without considering hierarchies – the predicted label needs to exactly match the gold label. In contrast, **Example-F1** calculates the average F1 scores considering each hierarchy as follows:

$$\text{Example-F1} = \frac{1}{N} \sum_{i=1}^N \frac{2 |\mathbb{L}_i^{\text{true}} \cap \mathbb{L}_i^{\text{pred}}|}{|\mathbb{L}_i^{\text{true}}| + |\mathbb{L}_i^{\text{pred}}|} \quad (1)$$

where $\mathbb{L}_i^{\text{true}}$ ($\mathbb{L}_i^{\text{pred}}$) is the true (model predicted) hierarchical label set of image i . For example, if the true labels for an image are “1967-7-14” and “Newark, New Jersey, United States, North America” respectively, then its true hierarchical label sets are [“Newark, New Jersey, United States, North America”, “United States, North America”, “North America”] and [“1967-7-14”, “1967-7”, “1967”, “1960s”, “20th century”].

6.2 Experimental results

In Table 2, we report the experimental results using the CLIP based baselines on the TARA. We can see that all of the model performance still have a large gap with human performance. Also, the object and facial segments boosts the model to be the highest on location prediction, proving that segment level reasoning is needed in this task. In contrast, adding the WIT weak supervision does not show consistent improvement or reduction on the performance. It can be due to that WIT images are not similar

Model	Accuracy	Example-F1
CLIP	11.11	44.96
CLIP+	15.72	49.74
CLIP+WIT	11.11	45.20
CLIP+Seg	16.46	50.52
Human	86.21	92.41

Model	Accuracy	Example-F1
CLIP	0.46	39.90
CLIP+	1.00	43.09
CLIP+WIT	1.07	41.73
CLIP+Seg	0.92	42.82
Human	75.86	91.63

Table 2: Summary of the performance(%) for different baselines on the image location prediction (above) and time prediction (bottom). Definition of Example-F1 is in Equation 1. Note that human performance here is evaluated on the test set of interest instead of on the whole test set, please see Section 6.3 for more details.

to news images, and that WIT images are mostly taken in older times than 2010, thus not providing enough supervision for our test set. There is also an obvious gap between the location prediction and time prediction, showing that temporal reasoning in vision language learning is much under explored and needs further research. Note that the Example-F1 value is consistently higher than accuracy because if the model predicts the highest two hierarchies correctly (e.g. century and decade), then it gets an Example-F1 around 40%.

6.3 Analysis

We perform qualitative and quantitative analysis of the baseline results to better understand the strengths and weaknesses of CLIP based models, and hypothesize avenues for future work. Specifically, we look into: model performance on test set of interest; effects on performance by using news abstract.

Test Set of Interest Since we conduct human evaluation only on the test set of interest, we examine how models perform on this set and show the results in Table 3. Note that we use the same setting for the models and human experts – both are given the entire test set labels. From the results, we observe a large gap between between the model performance and human performance, indicating that existing sota model still lacks a certain level of reasoning capability required to solve a such hard task as defined in the TARA dataset. Comparing the results in Table 3 to those in Table 2, we can see that there is little performance difference for each

Model	Accuracy	Example-F1
CLIP	13.33	56.44
CLIP+	13.33	58.67
CLIP+WIT	10.00	55.11
CLIP+Seg	23.33	63.11
Human	86.21	92.41

Model	Accuracy	Example-F1
CLIP	0.00	24.65
CLIP+	0.00	26.49
CLIP+WIT	0.00	29.83
CLIP+Seg	3.33	24.43
Human	75.86	91.63

Table 3: Performance(%) of different baselines evaluated on the test set of interest for image location prediction (above) and time prediction (bottom).

Model	Accuracy	Example-F1
CLIP	28.18	61.63
CLIP+	26.49	62.68
CLIP+WIT	11.11	50.00
CLIP+Seg	26.96	62.41

Table 4: Performance(%) for different baselines predicted towards news abstracts.

model, indicating that our human performance on the test set of interest can serve as a good reference to human performance on the whole test set, under the assumption that the annotators have enough knowledge about the key evidence segments.

News Abstracts We also experiment with news abstracts being the classification goal instead of time and location labels given an image, under the assumption that models are given corresponding news abstract for each label. The intuition is that the news abstract might provide more descriptions that can map to several local segments, and thus providing additional information. Comparing the results shown in Table 4 to Table 2, we can see that providing news abstracts improves the performance a lot, despite that there is still a large gap with human performance.

7 Conclusion

In this work, we introduce TARA, a new dataset and task for spatio-temporal grounding of images that requires open-ended joint reasoning with world knowledge. TARA provides a dataset of 16K high-quality images from NYT and Wikipedia-based supervision for additional 61K images. Compared to previous visual-language understanding datasets, TARA requires more complicated reasoning ability and existing state-of-the-art models such as CLIP

are far from human levels, suggesting that our task remains a significant challenge with large room for improvement. We hope that TARA will inspire future work on reasoning beyond image’s local segments in vision-language understanding.

8 Ethical Considerations

We collected data for TARA by downloading raw data from the official NYT API at <https://developer.nytimes.com>. According to the Terms of Use at <https://developer.nytimes.com/terms> and NYTimes.com Terms of Service located at <https://help.nytimes.com/hc/en-us/articles/115014893428-Terms-of-service>, NYT granted us a license to access the NYT APIs and scrape their data. We ensure that our dataset has been collected in a manner which is consistent with the terms of use of NYTimes.

We only release our dataset TARA for academic purpose. In order to retrieve the same raw data we scraped from the NYT API, multiple requests for months between January 1, 2010 and May 31, 2020 need to be made following the instructions at <https://developer.nytimes.com/docs/archive-product/1/overview>.

As introduced in Section 3.2, we annotated the data using crowd-workers through Amazon Mechanical Turk. They are voluntary participants who were aware of any risks of harm associated with their participation. We require the workers to be located in either Australia, Canada, Great Britain or the United States such that they are English speakers. We also require the workers to have HIT Approval Rate (%) for all Requesters’ HITs greater than or equal to 98%. All crowd-workers were compensated by a fair wage determined by estimating the average completing time of each annotation task. Each worker earn \$2.4 per 10 queries and each query should take less than a minute to annotate. Example screenshots of the NYT data and our annotation interface can be found in Appendix A.

Acknowledgments

This research is based upon work supported in part by the office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via Contract No. 2019-19051600006 under the BETTER Program, by Contracts FA8750-19-2-1004 and FA8750-19-2-0201 with the US Defense Advanced Research Projects

Agency (DARPA), and by a grant from the Allen Institute for Artificial Intelligence (allenai.org). The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of ODNI, IARPA, the Department of Defense, or the U.S. Government.

References

- Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. VQA: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433.
- Wenhu Chen, Aurelien Lucchi, and Thomas Hofmann. 2016. A semi-supervised framework for image captioning. *arXiv preprint arXiv:1611.05321*.
- Yuqing Cui, Apoorv Khandelwal, Yoav Artzi, Noah Snaveley, and Hadar Averbuch-Elor. 2021. Who’s Waldo? linking people across text and images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1374–1384.
- Milan Gritta, Mohammad Taher Pilehvar, Nut Lim-sopatham, and Nigel Collier. 2018. What’s missing in geographical parsing? *Language Resources and Evaluation*, 52:603 – 623.
- Jia Guo, Jiankang Deng, Alexandros Lattas, and Stefanos Zafeiriou. 2022. [Sample and computation redistribution for efficient face detection](#). In *International Conference on Learning Representations*.
- Drew A Hudson and Christopher D Manning. 2019. GQA: A new dataset for real-world visual reasoning and compositional question answering. *Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. 2017. CLEVR: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2901–2910.
- Sayali Kulkarni, Shailee Jain, Mohammad Javad Hosseini, Jason Baldridge, Eugene Ie, and Li Zhang. 2020. Spatial language representation with multi-level geocoding. *ArXiv*, abs/2008.09236.
- Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. [Neural architectures for named entity recognition](#). In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 260–270, San Diego, California. Association for Computational Linguistics.

- Fangyu Liu, Emanuele Bugliarello, Edoardo Maria Ponti, Siva Reddy, Nigel Collier, and Desmond Elliott. 2021. [Visually grounded reasoning across languages and cultures](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10467–10485, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Jiasen Lu, Vedanuj Goswami, Marcus Rohrbach, Devi Parikh, and Stefan Lee. 2020. 12-in-1: Multi-task vision and language representation learning. In *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Qiang Ning, Ben Zhou, Zhili Feng, Haoruo Peng, and Dan Roth. 2018. [CogCompTime: A tool for understanding time in natural language](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 72–77, Brussels, Belgium. Association for Computational Linguistics.
- Matthew E. Peters, Waleed Ammar, Chandra Bhagavatula, and Russell Power. 2017. [Semi-supervised sequence tagging with bidirectional language models](#). In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1756–1765, Vancouver, Canada. Association for Computational Linguistics.
- Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR.
- Jiaming Shen, Wenda Qiu, Yu Meng, Jingbo Shang, Xiang Ren, and Jiawei Han. 2021. [TaxoClass: Hierarchical multi-label text classification using only class names](#). In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4239–4249, Online. Association for Computational Linguistics.
- Krishna Srinivasan, Karthik Raman, Jiecao Chen, Michael Bendersky, and Marc Najork. 2021. [WIT: Wikipedia-Based Image Text Dataset for Multimodal Multilingual Machine Learning](#), page 2443–2449. Association for Computing Machinery, New York, NY, USA.
- Alane Suhr, Mike Lewis, James Yeh, and Yoav Artzi. 2017. [A corpus of natural language for visual reasoning](#). In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 217–223, Vancouver, Canada. Association for Computational Linguistics.
- 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.
- Naushad UzZaman, Hector Llorens, Leon Derczynski, James F. Allen, Marc Verhagen, and James Pustejovsky. 2013. SemEval-2013 Task 1: TempEval-3: Evaluating Time Expressions, Events, and Temporal Relations. In **SEMVAL*.
- Ben Zhou, Qiang Ning, Daniel Khashabi, and Dan Roth. 2020. [Temporal common sense acquisition with minimal supervision](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7579–7589, Online. Association for Computational Linguistics.
- Ben Zhou, Kyle Richardson, Qiang Ning, Tushar Khot, Ashish Sabharwal, and Dan Roth. 2021. [Temporal reasoning on implicit events from distant supervision](#). In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1361–1371, Online. Association for Computational Linguistics.

A Example Screenshots from the NYT Website and MTurk Annotation Interface

In this section, we first show an example news screenshot taken from the NYT website located at <https://www.nytimes.com/>, where we use the provided API to download the data, as in Figure 7. We then show example screenshot of our data annotation process as described in Section 3.2. For the data annotation tasks, we present the Turkers with step-by-step instructions of the tasks that we require them to do, along with carefully selected examples. More details can be found in Figure 8.

French Strikers Shut Down the Louvre, Setting a New Target in a Pension Fight

Protesters angry about the government's pension reforms prevented visitors from entering the museum, which is one of the French capital's most visited sites.



Union workers protested the government's pension plan outside the Louvre in Paris on Friday. Christophe Petit Tesson/EPA, via Shutterstock



By Aurelien Breeden

Jan. 17, 2020

PARIS — Visitors were turned away from the Louvre on Friday after strikers protesting the French government's planned pension overhaul blocked entrances to the museum, leaving tourists from around the world befuddled and frustrated.

Figure 7: This is an example news from the New York Times(NYT) website.

Instructions 1:

Below is an image extracted from a New York Times article.



Do not search online, by only looking at the image, Answer the Questions:

Can you tell which specific event the image is about? Events can be like a speech, celebration, etc. Non-events can be like city view, natural views, etc.

- Yes, I know the exact event.
- No, but it should be an event.
- No, I don't think the image is about an event. It's just random views.

Do you think human (you or anyone you know or the local people) can guess the image creation TIME through the information in this image, without searching online?

- Yes, I/someone can guess the date.
- Yes, I/someone can roughly guess the year.
- Yes, I/someone can roughly guess the decade.
- Yes, I/someone can only guess a rough century.
- No, because there is hardly any information about the time.

Do you think human (you or anyone you know or the local people) can guess the image creation LOCATION through the information in this image, without searching online?

- Yes, I/someone can guess the exact location (e.g. Building, Road name, etc.).
- Yes, I/someone can roughly guess the city.
- Yes, I/someone can roughly guess the country.
- Yes, I/someone can only guess a rough continent.
- No, because there is hardly any information about the location.

Instructions 2:

Look at this image extracted from a NYT article again. Read its caption, article publication date, article headline and article abstract. Answer the questions.



Image Caption: Union workers protested the government's pension plan outside the Louvre in Paris on Friday..

Article publication date: 2020-01-17

Article headline: French Strikers Shut Down the Louvre, Setting a New Target in a Pension Fight

Article abstract: Protesters angry about the governments pension reforms prevented visitors from entering the museum, which is one of the French capitals most visited sites.

We extracted **Caption-Time** and **Caption-Location** from the image caption and publication date. They should be the time and location of the **main event** in the caption.

Caption-Time: 2020-01-17

Caption-Location: Louvre Museum, Rue Croix des Petits Champs, Quartier Les Halles, 1st Arrondissement, Quartier du Palais Royal, 1st Arrondissement, Paris, Ile-de-France, Metropolitan France, 75001, France, Europe

Answer the question: Does the image show this main event?

For example, if the image is a cuisine and the caption is about a restaurant, then no; if the image is a building, and the caption is about the building being sold, then no; if the image is about car crash and caption is also about the car crash, then yes.

Does the image reflect the caption's main event? (By our extraction, the main event (not 100% correct) might possibly be **protested**.)

- Yes.
- No.
- Partially -- The image can only reflect partial event discussed in the caption.

Do you think that the image creation time is same as Caption-Time: 2020-01-17 (extracted from the image caption and publication date)?

- Yes, I think the image creation time is EXACTLY the same as Caption-Time, because the main event for Caption-Time is the same event described in the image.
- No, the image creation time is not the same as Caption-Time, because the main event for Caption-Time is not the same event described in the image.
- I guess no, even though the main event for Caption-Time is the same event described in the image, I don't think the image creation time must be Caption-Time.

Do you think that the image creation location is same as Caption-Location: Louvre Museum, Rue Croix des Petits Champs, Quartier Les Halles, 1st Arrondissement, Quartier du Palais Royal, 1st Arrondissement, Paris, Ile-de-France, Metropolitan France, 75001, France, Europe (extracted from the image caption)?

- Yes, I think the image creation location is EXACTLY the same as Caption-Location, because the main event for Caption-Location is the same event described in the image.
- No, the image creation location is not the same as Caption-Location, because the main event for Caption-Location is not the same event described in the image.
- I guess no, even though the main event for Caption-Location is the same event described in the image, I don't think the image creation location must be Caption-Location.

Figure 8: A screenshot of the MTurk annotation instructions for data validation as introduced in Section 3.2.