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

Anonymous ACL submission

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 on the signs, the buildings, the crowds, and more. This tells us the time when and the location where the image is taken, which will help us in subsequent tasks, such as evidence retrieval for criminal activities, automatic storyline construction, and upper-stream processing such as image clustering.

In this work, we formulate this problem and introduce TARA[1]: a dataset with 16k images with their associated news, time and location automatically extracted from New York Times[2] (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 images are believed to be feasible to find their 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.

1 Introduction

Vision and language are two of most important information sources, and the fact that humans can reason jointly with both sources at the same time has motivated artificial intelligence research to consider visually-grounded language understanding. Most work in this area has focused on reasoning with local evidence (Suhr et al., 2018; Hudson and Manning, 2019; Lu et al., 2020; Liu et al., 2021), e.g. asking about factoid questions such as the colors or shapes of objects and numbers of people, yet few of works 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 a 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, entity names, building styles, and other details in the input image.

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

In this work, we identify and formulate this problem, spatio-temporal grounding of images, a task aiming at identifying the *time* and *location* the given image was taken. Specifically, we de-

[1]Code and Data will be fully released to foster further studies.

[2]https://developer.nytimes.com/docs/archive-product/1/overview

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

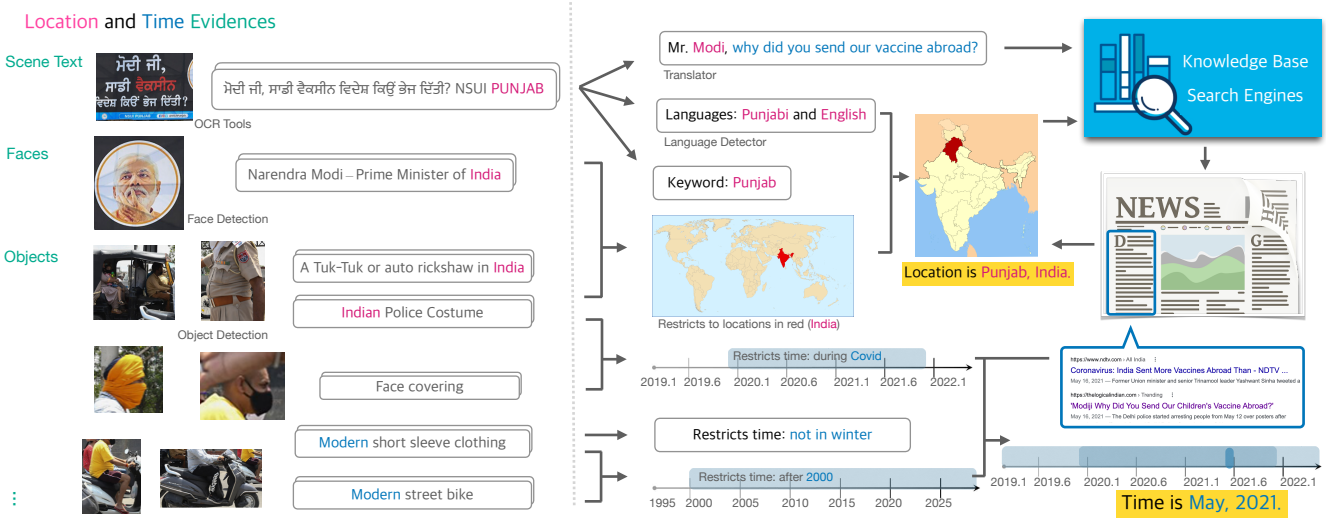

Figure 3: An example of potential joint reasoning on Figure 2 to ground its time and location. Note that people with different backgrounds may need to use different levels of reasoning, resulting in a completely accurate or just partial grounding (e.g., the decade and country), and we only show one possibility. 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.

velop a novel dataset TARA, (Time and plAce for Reasoning beyond the imAge), a challenging and important task that tasks models with grounding images to real-world spatial and temporal information. In our collection, we make sure that if models can accurately find images' creation time and location, they would need to successfully link the visual clues with contexts, which are often only found in texts such as news, stories and encyclopedias. As a result, this task motivates models to consider the association between visual information and language 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 validate 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 sizes, we collect an additional 61k examples from the Wikipedia domain. We ap-

ply 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. The earliest 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) are 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, and 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. Captions and corresponding news backgrounds such as headline, abstract, news type are also included for training or analysis purposes, but the task is to guess the correct time and location as detailed as possible given only the image. Our goal is to collect a large corpus of semantically rich images that human with world knowledge can correctly label time and location, with use of key evidence detection and visually-grounded world knowledge reasoning. We design a process to collect and identify images that enable such types 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 API [3]. We did not collect news earlier than 2010 because earlier news articles contain much fewer images. Each news article comes with a list of attributions[4] such as headline, abstract, news type, and possible main image. We first filter the news with 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 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 unwanted news types such as reviews, series, and

[3] https://developer.nytimes.com/docs/archive-product/1/overview

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

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

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 model[5] 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 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 our 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 Geopy[6] to identify the location and add missing hierarchies such as country and continent.

3.2 Validation

We randomly select equal number of images from each month, such that a total of about 20% images are assigned to devlopment and testing. On these images, we use two crowdsourcing tasks to

[5] https://spacy.io/models/en

[6] https://geopy.readthedocs.io/en/stable/

(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 local people 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) model[7] 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.[8] 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 vise 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.

[7] https://demo.allennlp.org/semantic-role-labeling

[8] 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.

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 AllenNLP[9] 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.

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

[9] 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.

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" model[10] 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 fine-tuned 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 natural language sentence to serve as the text part for an image.

CLIP+Seg is another variant where we first ex-

[10] https://github.com/openai/CLIP

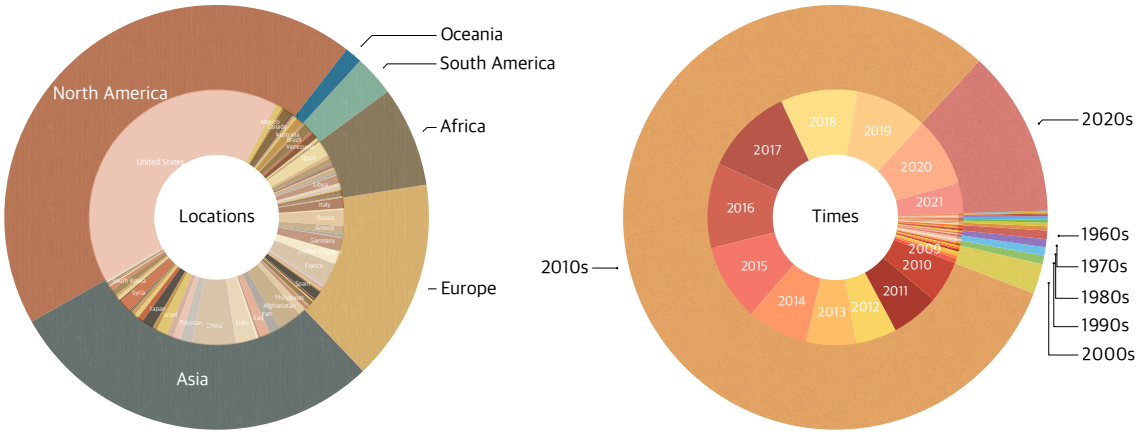

Figure 6: Label distribution in TARA. All of the training, development, and testing data are considered.

tract 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 YOLOv5[11] 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., 2021) facial detection model[12]. 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.

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:

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

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 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

[11] https://github.com/ultralytics/yolov5
[12] https://github.com/deepinsight/insightface

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.

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).

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 baseline results to better understand 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 same setting for the models and human experts – both are given the entire test set labels. From the results, we can see that model performance has a large gap with human performance, indicating that existing sota model still lacks a certain level of reasoning capability to solve a hard task such as TARA. Comparing the results in Table 3 to those in Table 2, we can see that there is little performance difference for each 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.

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.

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 16K high-quality dataset from NYT and a 61K additional supervision from Wikipedia. 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.

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*.

Milan Gritta, Mohammad Taher Pilehvar, Nut Limsopatham, 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. 2021. Sample and computation redistribution for efficient face detection. *arXiv preprint arXiv:2105.04714*.

Drew A Hudson and Christopher D Manning. 2019. Gqa: a new dataset for compositional question answering over real-world images. *arXiv preprint arXiv:1902.09506*, 3(8).

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. *arXiv preprint arXiv:1603.01360*.

Fangyu Liu, Emanuele Bugliarello, Edoardo Maria Ponti, Siva Reddy, Nigel Collier, and Desmond Elliott. 2021. Visually grounded reasoning across languages and cultures. *arXiv preprint arXiv:2109.13238*.

Jiasen Lu, Vedanuj Goswami, Marcus Rohrbach, Devi Parikh, and Stefan Lee. 2020. 12-in-1: Multi-task vision and language representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10437–10446.

Qiang Ning, Ben Zhou, Zhili Feng, Haoruo Peng, and Dan Roth. 2018. Cogcomptime: A tool for understanding time in natural language. In *EMNLP*.

Matthew E Peters, Waleed Ammar, Chandra Bhagavatula, and Russell Power. 2017. Semi-supervised sequence tagging with bidirectional language models. *arXiv preprint arXiv:1705.00108*.

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. *arXiv preprint arXiv:2103.00020*.

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.

Krishna Srinivasan, Karthik Raman, Jiecao Chen, Michael Bendersky, and Marc Najork. 2021. Wit: Wikipedia-based image text dataset for multimodal multilingual machine learning. *arXiv preprint arXiv:2103.01913*.

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.

Alane Suhr, Stephanie Zhou, Ally Zhang, Iris Zhang, Huajun Bai, and Yoav Artzi. 2018. A corpus for reasoning about natural language grounded in photographs. *arXiv preprint arXiv:1811.00491*.

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 **SEMEVAL*.

Ben Zhou, Qiang Ning, Daniel Khashabi, and Dan Roth. 2020. Temporal Common Sense Acquisition with Minimal Supervision.

Ben Zhou, Kyle Richardson, Qiang Ning, Tushar Khot, Ashish Sabharwal, and Dan Roth. 2021. Temporal reasoning on implicit events from distant supervision. *NAACL*.