



Visual Goal-Step Inference using wikiHow

Yue Yang, Artemis Panagopoulou, Qing Lyu, Li Zhang,
Mark Yatskar, Chris Callison-Burch

Department of Computer and Information Science, University of Pennsylvania



Introduction

- Teach AI system to understand complex events
- Goal – Step representation
- Reasoning relationship between high-level goals and the steps
 - E.g. Goal: “change a tire” → Steps: “place the jack under the car”, “raise the jack”,
- Past work examined the task of goal-step inference for **text**
- However, different modalities existed in real world



During my trip to CA this summer, I got a flat tire on highway one...

Introduction

- Learning goal-step relations in a multimodal fashion.
- Vision-enabled dialog systems
 - Recognize the goal of human activities
 - provide helpful recommendations
- Mitigate reporting bias
 - Induce lower-level steps
- We propose the Visual Goal-Step Inference (VGSI) task
 - Given given a textual goal
 - Infer which image represents a plausible step towards that goal
- More challenging

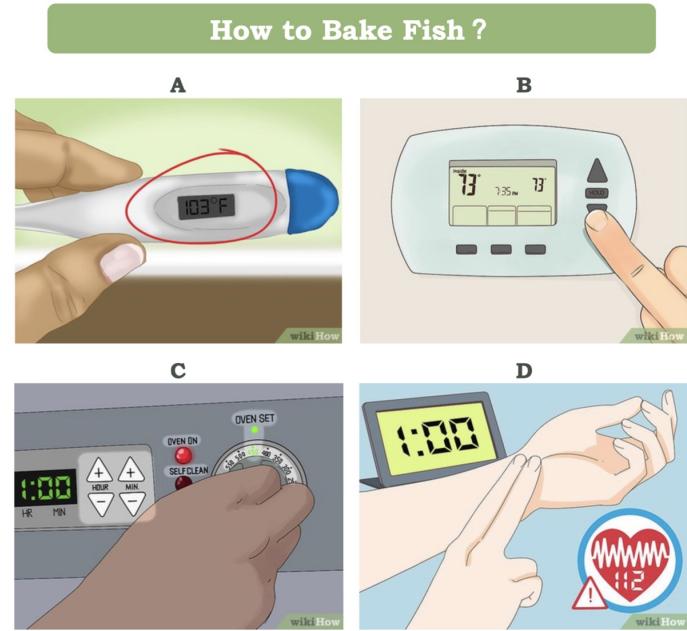


Figure 1: An example Visual Goal-Step Inference Task: given a text goal (*bake fish*), select the image (C) that represents a step towards that goal.

Dataset

- Harvested from wikiHow
- Goal – Method – Step structure
- The corpus consists
 - 53,189 wikiHow articles across various categories
 - 155,265 methods, 772,294 steps/images

Category	Goals	Methods	Steps	Images
Health	7.8k	19.1k	97.5k	111.8k
Home and Garden	5.9k	16.0k	82.9k	85.4k
Education & Communications	4.7k	12.4k	61.2k	66.1k
Food & Entertaining	4.6k	11.6k	62.0k	69.0k
Finance & Business	4.4k	11.8k	59.3k	66.8k
Pets & Animals	3.5k	9.5k	45.3k	48.0k
Personal Care & Style	3.4k	9.0k	46.1k	48.9k
Hobbies & Crafts	2.8k	7.5k	40.9k	42.7k
Computers & Electronics	2.6k	6.1k	31.5k	36.2k
Arts & Entertainment	2.5k	6.8k	35.4k	37.2k
Total	53.2k	155.3k	772.3k	772.3k

Table 1: Number of goals, methods, steps and images in the top 10 wikiHow categories.

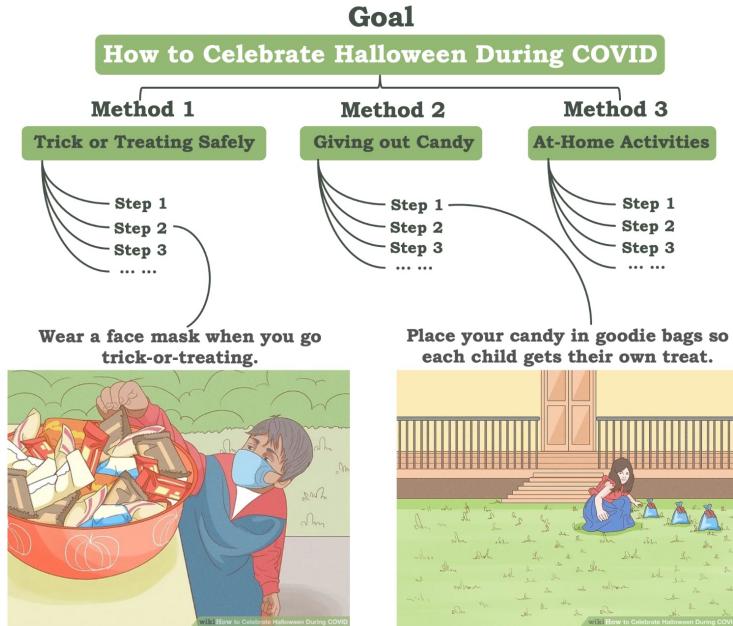


Figure 2: Hierarchical multimodality of wikiHow.

Methods

- Problem Formulation:
 - Input: a high-level goal G , an Image I
 - The model outputs the matching score:
- Baseline Models:
 - **DeViSE**: maps the source vector onto the span of the target vector.
 - **Similarity Network**: 2 branches with pointwise multiplication, train as binary classifier.
 - **Triplet Network**: 3 branches, minimize the positive pair distance, maximize negative pair distance.
 - **LXMERT**: transformer encoders, attention layers to ground text to objects.

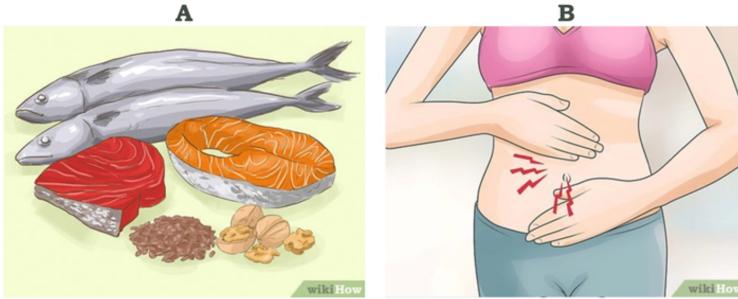
Experimental Setup

- Sampling Strategies
 - Random Sampling:
 - randomly pick three different articles
 - select one image by chance from each article
 - Similarity Sampling:
 - greedily select the most similar images based on the feature vectors
 - use FAISS to retrieve the top-3 most similar images
 - Category Strategy:
 - randomly select three different articles within the same wikiHow category as the goal
 - select one image by chance from each article

Experimental Setup

(C) Category Sampling

How to Prevent Kidney Disease ?



How to Walk Your Dog at Night ?



(C.1) Correct Answer is A



(C.2) Correct Answer is D

Results

Model	Sampling Strategy (Test Size)		
	Random (153,961)	Similarity (153,770)	Category (153,961)
Random	.2500	.2500	.2500
DeViSE	.6719	.3364	.4558
Similarity Net	.6895	.6226	.4983
LXMERT	.7175	.4259	.2886
Triplet Net (GloVe)	.7251	.7450	.5307
Triplet Net (BERT)	.7280	.7494	.5360
Human	.8450	.8214	.7550

Table 2: Accuracy of SOTA models on the wikiHow VGSI test set with different sampling strategies (sample size is shown in parentheses).

Transfer Learning

- Most wikiHow images are drawings.
- Can we apply the knowledge learned from wikiHow to real photos?
- Target dataset (Instructional Videos + Keyframe extraction)
 - Howto100m
 - randomly select 1,000 goals, one video per goal
 - k-means clustering to select key frames,
 - Split goals into 8:2 for training and testing
 - COIN
 - 180 goals, 5 videos per goal for testing, the remaining 9,709 videos for training
 - The videos in COIN are segmented into clips, we randomly pick one frame per video
- Caption-based dataset for comparison
 - Flickr30K
 - MSCOCO

Transfer Learning

		Sampling Strategy		
PT-Data	FT?	Random	Similarity	Category
-	✓	.6649	.5085	.5216
Flickr30K	✗	.4903	.5103	.3919
	✓	.7006	.5823	.5495
MSCOCO	✗	.5349	.5401	.4071
	✓	.7481	.6180	.5536
Howto100m	✗	.5694	.5811	.3989
	✓	.6948	.6104	.5436
wikiHow	✗	.6245	.6309	.4586
	✓	.7639	.6854	.5659
Human	-	.9695	.8500	.8682

Table 3: Transfer performance (4-way multiple choice accuracy) on COIN. PT stands for pre-training, FT for fine-tuning. FT results are obtained by fine-tuning the model on 5 examples of the COIN training set (i.e., 5-shot). **Red** numbers indicate the best zero-shot performance. **Blue** numbers are the best fine-tuned results.

		Sampling Strategy		
PT-Data	FT?	Random	Similarity	Category
-	✓	.6005	.6096	.4434
Flickr30K	✗	.4837	.5398	.3856
	✓	.6207	.6408	.4740
MSCOCO	✗	.5099	.5715	.3958
	✓	.6340	.6640	.4794
COIN	✗	.5067	.5161	.3978
	✓	.6170	.6343	.4638
wikiHow	✗	.6556	.6754	.4750
	✓	.6855	.7249	.5143
Human	-	.8300	.7858	.7550

Table 4: Transfer performance (4-way multiple choice accuracy) on Howto100m. FT results are obtained by fine-tuning the model on the full training set.

Transfer Learning

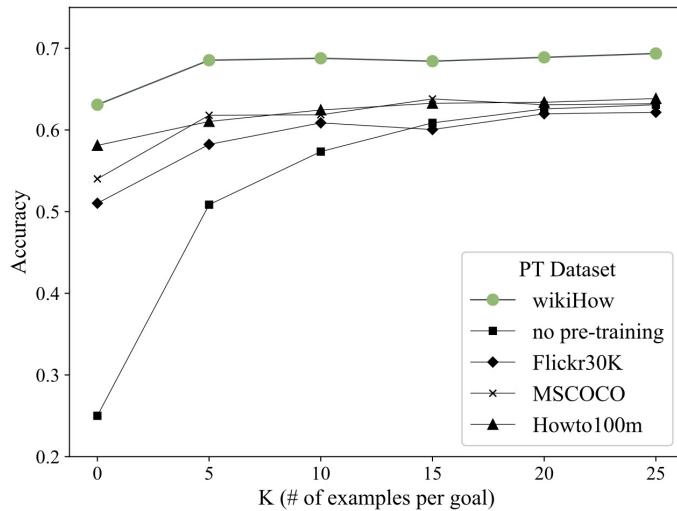


Figure 4: Few-shot performance on COIN (similarity sampling) with different pre-training datasets vs. the number of examples per goal.

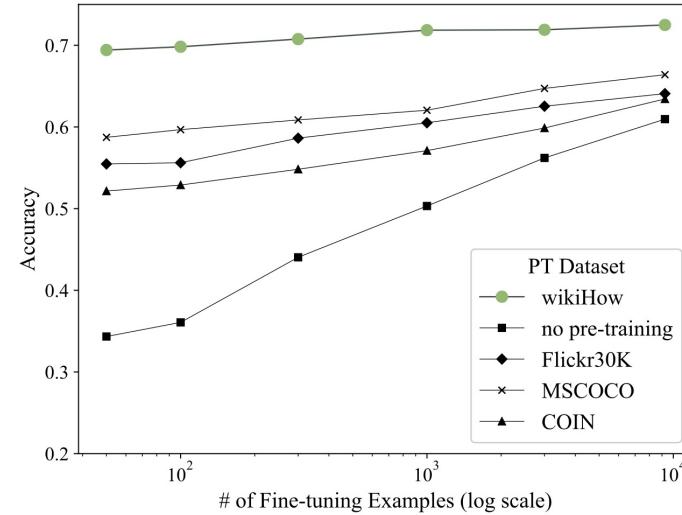


Figure 5: Transfer performance on Howto100m (similarity sampling) with different pre-training datasets vs. the number of training examples.

Conclusion

- We propose the novel Visual Goal-Step Inference task (VGSI), a multimodal challenge for reasoning over procedural events.
- We construct a dataset from wikiHow and show that SOTA models struggle on it.
- The knowledge harvested from our dataset could transfer to other domains.
- The multimodal representation learned from VGSI has strong potential to be useful for NLP applications such as multimodal dialog systems, multimodal schema induction systems.

Thank you!

Dataset and code are available at
<https://github.com/YueYANG1996/wikiHow-VGSI>