Breaking Shortcuts by Masking for Robust Visual Reasoning (Supplementary File)

Keren Ye, Mingda Zhang and Adriana Kovashka
Department of Computer Science, University of Pittsburgh, Pittsburgh PA, USA
{yekeren, mzhang, kovashka}@cs.pitt.edu

In this document, we include more information and statistics regarding the knowledge and the models mentioned in our paper. We also provide additional qualitative experimental results.

We first show the exact SPARQL query we used to query the DBpedia in Sec. 1. We also provide examples of the retrieved DBpedia comments. They qualitatively demonstrate that the retrieval expands the understanding beyond the image and text within the ad image in that the original queries themselves are abstract notations or symbols.

Next, we provide statistics of the expanded knowledge of the PittAds dataset in Sec. 2. It shows that in our problem, irrelevant pieces of knowledge dominate the relevant ones. Hence, it is a challenge for the model to selectively use them.

We show the user interface for collecting ground truth knowledge pieces in Sec. 3. Such annotations provide hints for what are the appropriate types of knowledge to use, to understand a given ad. Based on the annotations, Sec. 4.3 in the main paper measures if our models reason correctly, i.e. similar to how a human would.

Finally, we provide model training details in Sec. 5 and additional qualitative experimental results in Sec. 4.

1. SPARQL query and retrieved knowledge

We show our SPARQL query in Fig. 1. It returns DBpedia comments that meet any of the following conditions: (1) they use the keywords as their labels; (2) they have related abbreviations to the keywords; (3) they have been linked with `wikiPageDisambiguates`, or `wikiPageRedirect` pages and the source page is labeled with the keyword. After retrieving the contents, we filtered out non-English content and limited the `rdf:type` in our 19 predefined types (`dbo:Company`, `dbo:Organisation`, etc.), which are chosen by one of our authors and are thought to be ads-related.

```
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX dbpedia2: <http://dbpedia.org/property/>
PREFIX dbo: <http://dbpedia.org/ontology/>
SELECT DISTINCT ?entry ?comment
WHERE {{
        ?entry rdf:type ?type.
        ?entry rdfs:label "[QUERY]"@en.
        ?entry rdfs:comment ?comment.
    } UNION {
        ?entry rdf:type ?type.
        ?entry dbpedia2:abbreviation ?abbreviation.
        ?entry rdfs:comment ?comment.
        FILTER regex(?abbreviation, "^[QUERY][^a-zA-Z].*?")
    } UNION {
        ?entry rdf:type ?type.
        ?entry_dup rdfs:label "[QUERY]"@en.
        ?entry rdfs:comment ?comment.
        ?entry ^dbo:wikiPageDisambiguates|^dbo:wikiPageRedirects ?entry_dup. }
    FILTER langMatches(lang(?comment),'en').
    FILTER (?type IN (dbo:Company, dbo:Organisation, dbo:Person, dbo:Software, dbo:VideoGame
        , dbo:Food, dbo:Weapon, dbo:Place, dbo:Country, dbo:Location, dbo:Beverage, dbo:
        Agent, dbo:Bank, dbo:EducationalInstitution, dbo:University, dbo:Actor, dbo:
        Publisher, dbo:Department, dbo:School))
} LIMIT 3
```

Figure 1: **SPARQL query for retrieving the DBpedia comments.** We denote the actual textual keyword using [QUERY].

Keywords	Url	`rdfs:comment`
Nike	`http://dbpedia.org/page/Nike_(mythology)`	In ancient Greek religion, Nike was a goddess who personified victory...
	`http://dbpedia.org/page/Nike,_Inc.`	This article is about the sportswear and apparel company. For other uses of the name "Nike", see Nike (disambiguation).) Nike, Inc. is an American multinational corporation that is engaged in the design, development, manufacturing and worldwide marketing and sales of footwear, apparel, equipment, accessories and services...
	`http://dbpedia.org/page/307_Nike`	307 Nike is a sizeable asteroid of the main belt. It was discovered by Auguste Charlois on March 5, 1891 while working at the Nice Observatory...
DMV	`http://dbpedia.org/page/DMV_(song)`	"DMV" is a song by the rock band Primus. Interscope Records asked Primus to release this song together with its video...
	`http://dbpedia.org/page/Department_of_Motor_Vehicles`	In the United States, a department of motor vehicles (DMV) is a state-level government agency that administers vehicle registration and driver licensing...
WWF	`http://dbpedia.org/page/Windows_Workflow_Foundation`	Windows Workflow Foundation (WF) is a Microsoft technology that provides an API, an in-process workflow engine, and a rehostable designer to implement long-running processes as workflows within .NET applications...
	`http://dbpedia.org/page/Words_with_Friends`	Words with Friends is a multi-player word game developed by Newtoy, Inc. Players take turns building words crossword puzzle style in a manner similar to the classic board game Scrabble...
	`http://dbpedia.org/page/World_Wide_Fund_for_Nature`	The World Wide Fund for Nature (WWF) is an international non-governmental organization founded in 1961, working in the field of the wilderness preservation, and the reduction of humanity's footprint on the environment...

Table 1: **Knowledge examples retrieved by the SPARQL query**. We see from these examples that some contents retrieved are not relevant to advertisements involving the keywords (here, products and organizations). For example, most ads referring to WWF are about the World Wide Fund only, but *all* entries on the right will be retrieved for all ads that mention WWF. Thus, we require the models to find the most suitable information to help the prediction.

2. Statistics of the additional data to the PittAds dataset

We mentioned in the paper that there are 6.9 slogans and 27.5 DBpedia comments on average, for each image. Here, we provide more details regarding the distribution of the expanded annotations. In general, images with less than 20 slogans constitute 94.2% of the dataset (0-9 slogans: 79.4%; 10-19 slogans: 14.8%), while images with less than 20 pieces of related knowledge only constitute 59.3% (<10 comments : 38.4%; 10-19 comments: 20.9%). This means that models need to make a choice among a large candidate pool based on their judgment (40.7% images are associated with ≥20 knowledge pieces).

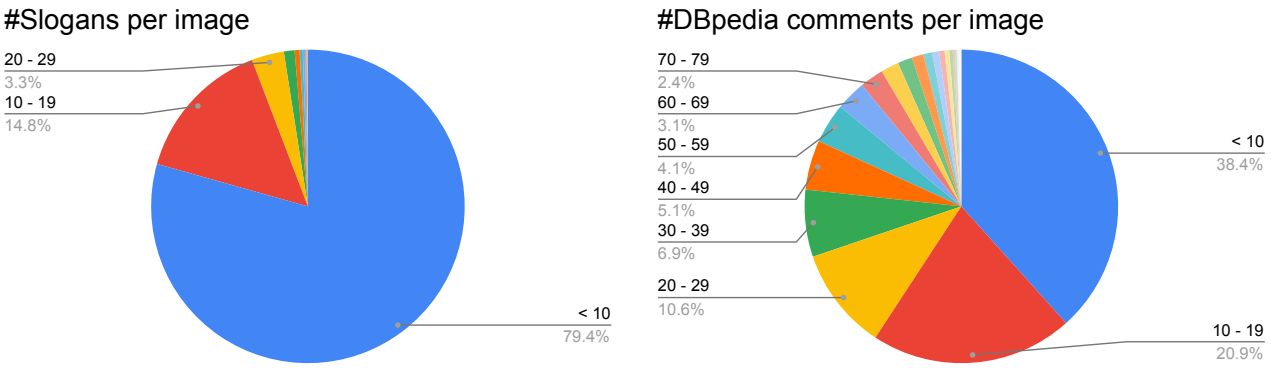


Figure 2: **Statistics of the additional data to the PittAds dataset.**

3. Collecting ground truth knowledge pieces

Fig. 3 shows our user interface for collecting ground truth knowledge pieces. These annotations are used only for evaluation purposes. We show our guidelines on the left and show the annotation task on the right. The interface gathered personal opinions regarding how to reason based on knowledge.

More specifically, we asked human participants to provide a "gold standard": for a given advertisement, we show all retrieved knowledge pieces to human annotators and ask them to select whether each piece is helpful or not, for the ad understanding task.

We provide 410 images to 10 different human annotators (authors are not involved), and 270 of them were annotated with the helpful knowledge; for the remaining images, all retrieved knowledge was marked irrelevant by our annotators.

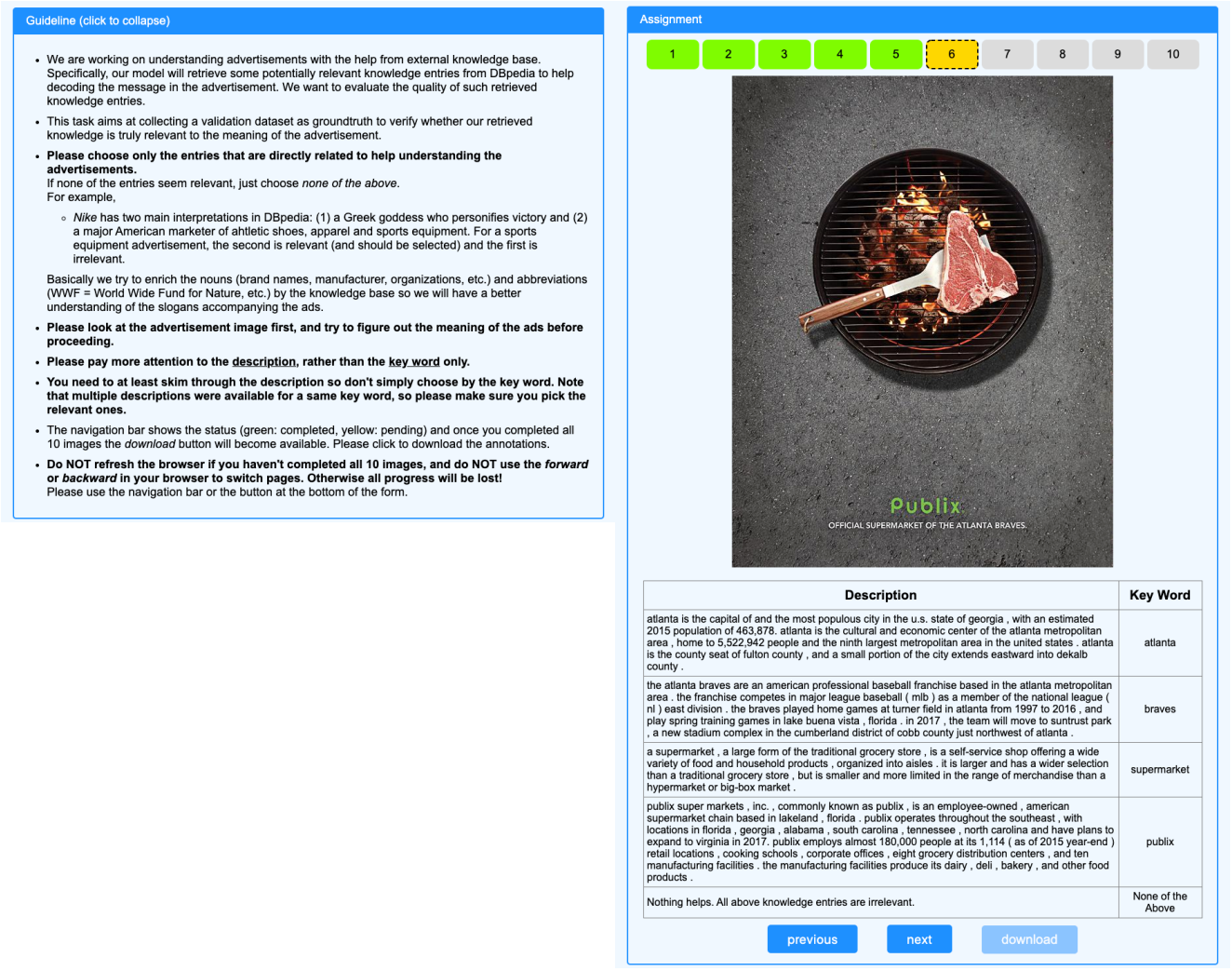


Figure 3: **User interface for collecting ground truth knowledge pieces as described in Sec. 4 in the main paper.**

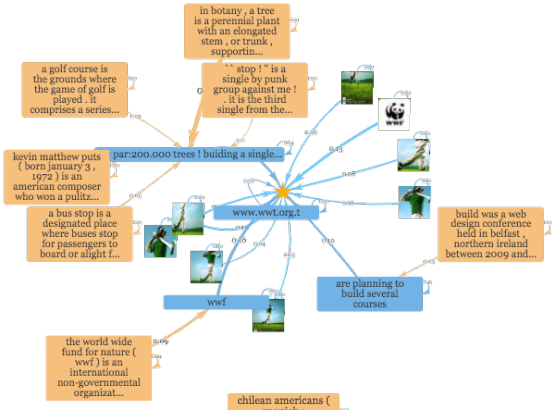
4. More qualitative examples

Fig. 4 shows the learned graphs of some PSAs. In the first example, the model referred to the knowledge of "tree" and "WWF" since the two pieces of knowledge have large edge weights. In the second example, the model examined "suicide" and "smoking" in detail. In the last example, the model paid more attention to the description of "abuse". All these examples show that our model expands reasoning beyond the ad in a reasonable manner.

I should support the WWF because they are trying to protect the wildlife from becoming a golf course

I should support the WWF because they help protect nature

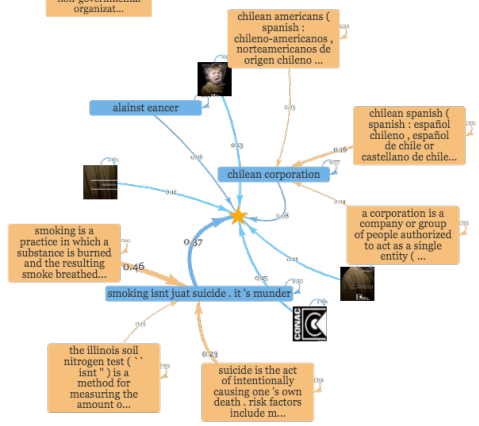
I should be more environmentally conscious because of all the trees getting cut down

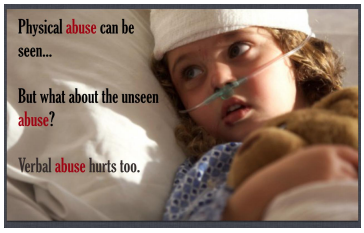


I should not smoke because it will kill more than just me

I should not smoke around children because it can give them cancer

I should not put a plastic sack on my kid's head because he'll be mad





I should prevent verbal abuse because its as bad as physical abuse

I should help stop abuse because it saves lives

I should not verbally abuse my children because it hurts as much as physical abuse

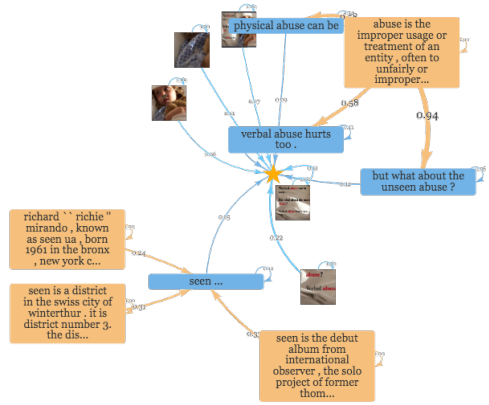


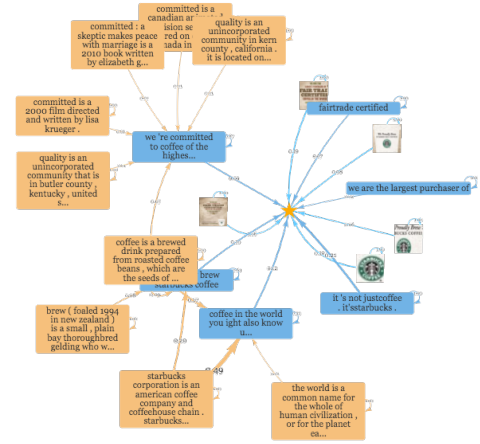
Figure 4: **PSAs: more qualitative examples.** We show the ad image on the left, the graph learned by our model on the rights. We show slogans in blue and DBpedia comments in orange, and the global node is represented by a star. The width of arrow is correlated with learned weights α, β discussed in the main paper. We used a threshold of 0.0001 to remove unimportant edges.

Fig. 5 shows the learned graphs of some product advertisements. We observed that the model benefits from knowing the background of the brand name. In these examples, the explanations of "Starbucks", "Louis Vuitton", and "Sony" have larger edge weights in the graph. This is similar to human reasoning in that we also based our ad reasoning on our former understanding of the brands.

I should drink Starbucks because they are fair trade certified

I should support this company because it supports coffee workers and fair trade

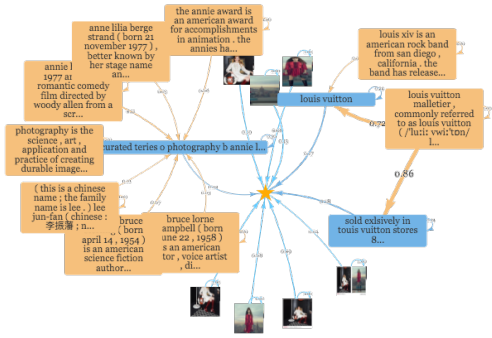
I should buy this product because it will quench my thirst

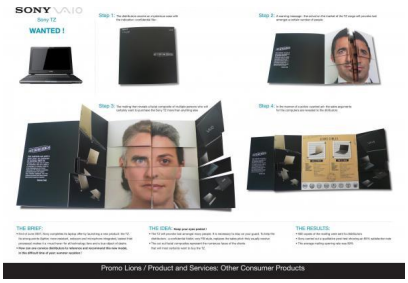


I should buy Louis Vuitton because they are trend setters

I should buy Louis Vuitton handbags because they are original

I should buy this product because it is a high quality brand and very tending





I should buy a Sony laptop because they let me create

I should buy a Sony laptop because it looks creative

I should buy this laptop because I can customize it to meet my needs

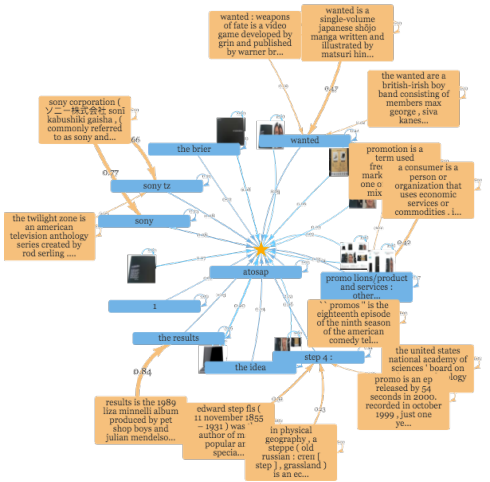


Figure 5: **Product ads: more qualitative examples.** We show the ad image on the left, the graph learned by our model on the rights. We show slogans in blue and DBpedia comments in orange, and the global node is represented by a star. The width of arrow is correlated with learned weights α, β discussed in the main paper. We used a threshold of 0.0001 to remove unimportant edges.

5. Training details

Before training, we use a pre-trained object detection model [5] to generate at most 10 ads object proposals per image, represented using InceptionV4 [3]. For the OCR detected slogans, we sort them by their areas and keep only the biggest 20 regions. Our vocabulary for slogan, knowledge and statements consists of words that appeared more than 5 times in human-annotated statements or more than 20 times in either the OCR slogans or DBpedia comments. We use a larger threshold for the latter two because these two corpora involve more diverse contents. To extract the text features, we use a 300-D word embedding initialized from GloVe [2] and the BiLSTM encoders with 200 hidden units with a dropout keep probability of 0.8. The dimensions of the image-text joint feature space is set to 200. We set the hidden units of all relation MLPs to 200 and add a dropout layer with keep probability of 0.5 after their `tanh` activation. We choose Tensorflow [1] framework and use the RMSprop optimizer with learning rate of 0.001. We use batch size of 128, and set η in the triplet loss to 0.2 based on [4].

References

[1] Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, et al. Tensorflow: A system for large-scale machine learning. In *12th USENIX Symposium on Operating Systems Design and Implementation (OSDI)*, 2016.

[2] Jeffrey Pennington, Richard Socher, and Christopher Manning. Glove: Global vectors for word representation. In *The 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2014.

[3] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander A. Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning. In *The Thirty-First AAAI Conference on Artificial Intelligence (AAAI)*, 2017.

[4] Keren Ye and Adriana Kovashka. Advise: Symbolism and external knowledge for decoding advertisements. In *The European Conference on Computer Vision (ECCV)*, 2018.

[5] Keren Ye, Mingda Zhang, Adriana Kovashka, Wei Li, Danfeng Qin, and Jesse Berent. Cap2det: Learning to amplify weak caption supervision for object detection. In *The IEEE International Conference on Computer Vision (ICCV)*, 2019.