# ADVISE: Symbolism and External Knowledge for Decoding Advertisements (Supplementary Material)

Keren Ye[0000−0002−7349−7762] and Adriana Kovashka[0000−0003−1901−9660]

University of Pittsburgh, Pittsburgh PA 15260, USA
{yekeren,kovashka}@cs.pitt.edu

In this document, we include more information and statistics about our ADVISE model and some implementation decisions. We also provide additional quantitative and qualitative experimental results.

We first describe in more depth the implementation and evaluation setup. We provide details about the implementation and training of our ADVISE model in Sec. 1. In Sec. 2, we describe different ways of adapting Hussain et al. [22]'s method to the ranking task. In Sec. 3, we explain the reason why we choose to rank 50 statements for our main task.

Next, we provide additional quantitative results which demonstrate the contribution of different algorithmic choices that we made. In Sec. 4, we provide details justifying our choice of the value of $k'$ for hard negative mining. In Sec. 5, we demonstrate different strategies and justify our choice of the attention mechanism used. In Sec. 6, we quantitatively demonstrate that it is not sufficient to use any type of label in the domain of interest for the method component described in Sec. 3.4 of the main text; in particular, we show that using the symbol labels as constraints gives more improvement than using topic labels. In Sec. 7, we break down the ranking task evaluation into topics.

Finally, to enable a more intuitive understanding of our model, we provide more qualitative results of the ranking task in Sec. 8, including both statement ranking and hard-statement ranking. In Sec. 9, we show qualitatively that the ADVISE model learns not only the image representation but also meaningful *region* representations.

## 1 Training the ADVISE model

As shown in Figure 2 in our paper, there are two branches in our model, the main branch (top left), and the knowledge embedding branch (top right). We do not train the branches jointly at the very beginning, instead, we at first train the main branch till the model converges, then we add the knowledge base information. It is beneficial to incorporate knowledge additively using this two-step process. In our experiments, training the knowledge branch requires less time compared to training the main branch. Thus, one could efficiently experiment with multiple extra types of knowledge given that the main branch is trained and the entry-points of symbols are properly set. Another advantage of the two-step training is that the knowledge branch serves a role similar to a residual branch, that is it would not hurt the performance of the existing main branch.

We experimented with using Adagrad, Adam, RMSProp and found Adagrad to give the best results. We use a learning rate of 2.0 and apply no decay strategy on the learning rate. Also, we did not use different gradient multipliers for image and text embedding networks. For both the image embedding network ($\boldsymbol{w}$ in Eq. 4) and the attention prediction network ($\boldsymbol{w_a}$ in Eq. 5), we use batch normalization layer, a weight decay of 1e-6, and dropout keep probability of 0.7 for the input Inception V4 feature. For the text embedding network of the statement (Sec. 3.2), DenseCap captions, and symbols (Sec. 3.4), we use a weight decay of 1e-8 and a dropout keep probability of 0.7 for the embedding weights. The DenseCap captions do not share weights with the statements, but they are both initialized from GLOVE word embedding [48]. For the "unknown" words of DenseCap captions, Ads statements and the symbol embedding vectors, they are initialized from the uniform distribution ranging from -0.08 to 0.08. According to our experiments, adding a dropout layer (with keep probability of 0.5) after the pointwise multiplication of image and text embedding of $||x - y||_2^2$ is really important. It assures that the model will not overfit on the training set. For the triplet training, we mine the most challenging 32 negative examples in the 128-sized training batch, we weigh 0.1 for the symbol loss and object loss as mentioned in Eq. 8. Based on the settings mentioned above, we train the main branch for 100,000 steps and use Recall@3 as the metric to choose the best model on the validation set.

We build the knowledge branch after getting the checkpoint of the main branch. During this second phase, we freeze all of the parameters we harvest in the first step, saying $\boldsymbol{w}$ for the image embedding $\boldsymbol{z}$, $\boldsymbol{w_a}$ for attention prediction, $\boldsymbol{t}$ for Ads statement embedding, $\boldsymbol{c}$ for DenseCap embedding, and $\boldsymbol{s}$ for symbol embedding. In the meantime, we also freeze the parameters of the pre-trained symbol classifiers ($\boldsymbol{u_{symb}}$ as mentioned in Sec. 3.5) since the classifiers are part of our prior knowledge. Therefore, the only parameters of our ADVISE model in the second training phase are the 53 scalar values, that is, $\alpha_j^{symb}$ (for each of $j \in \{1, ..., C = 53\}$), denoting the importance of the 53 classifiers. If the main branch captures all the information during the training, assigning 0 to all $\alpha_j^{symb}$ would not hurt the performance of the model. In case the main branch

does *not* capture all relevant information, the knowledge branch may provide complementary information to help to improve the final performance. Thus the knowledge branch trained is similar to a residual branch. Please note that the symbol embedding $\boldsymbol{s}$ is learned in the first phase using symbol constraint, and so these symbol embedding vectors serve as entry points for external knowledge. In order to bound the $\alpha_j^{symb}$, we apply the *sigmoid* activation on them and multiply them by 2. Thus the 53 confidence scores of the classifiers are ranging from 0 to 2, in which 0 means the associated classifier is not useful. To train the knowledge branch, we use the Adam optimizer, and a learning rate of 0.01. Based on this setting, we train both the branches jointly (with main branch freeze) for 5,000 steps and cross-validate to get the best model.

In order to show that our ADVISE really learns the importance of different classifiers, we show the confidence scores of the top-10 most useful and bottom-10 least useful symbol classifiers in Figure 1. The figure accords to our expectation because our ADVISE model tends to weigh more on classifiers such as "smoking", "animal cruelty" while in the Ads dataset there are only a limited number of training examples for these symbols. Thus incorporating knowledge for these not well-trained symbols is necessary.

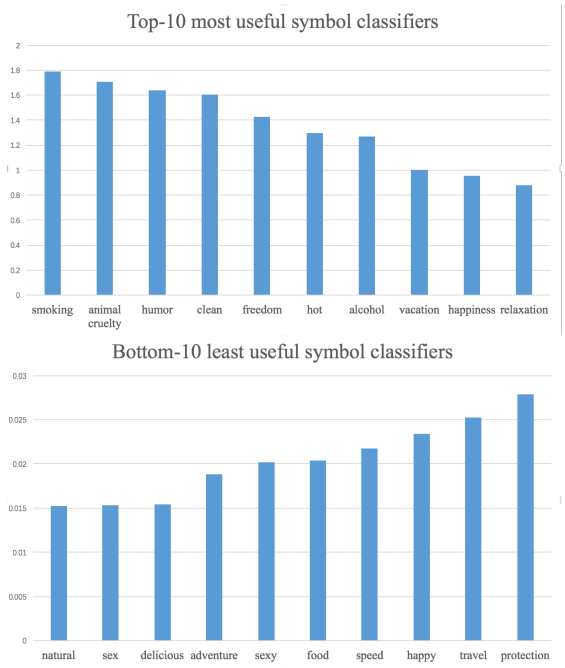


**Fig. 1.** Confidence scores of the symbol classifiers.

## 2 Adapting Hussain et al.'s method [22]

Hussain et al. [22] developed the only method we are aware of for understanding the messages of advertisements. In the main text, we show the results of the most promising way of adapting Hussain et al.'s question-answering method for our ranking task. The other ways of adapting Hussain's method that we tried include retrieving the statement that had the highest similarity between the single-word picked in [22] and the action-reason statements, using a standard GLOVE embedding or the embedding learned in our method. We also tried concatenating the Inception-V4 and symbol features rather than pointwise-adding them, and using the original VGG features used in [22].

## 3 Reason to choose 50 statements

Our evaluation is similar to [3] which provides 18 choices. The task is challenging, and enlarging the list would make all methods score so poorly (e.g. low Recall@3) that their performance would be hard to compare. Table 1 demonstrates this challenge and shows that the 3 correct statements are internally more similar than a correct and an incorrect statement. We computed $D^{min}_{within}$, $D^{min}_{between}$, $D^{avg}_{within}$, $D^{avg}_{between}$ as follows, where $\boldsymbol{x}$ is an image in Ads dataset $\mathbb{D}$, $P(\boldsymbol{x})$ is the set of related statements for $\boldsymbol{x}$, $N(\boldsymbol{x})$ are the randomly sampled statements from images other than $\boldsymbol{x}$, and $g(\cdot)$ computes average GLOVE embedding.

$$
\begin{aligned}
D^{min}_{within} &= \operatorname*{avg}_{x\in\mathbb{D}} \min_{a,b\in P(x),a\neq b} ||g(a)-g(b)||_2^2 \\
D^{avg}_{within} &= \operatorname*{avg}_{x\in\mathbb{D}} \operatorname*{avg}_{a,b\in P(x),a\neq b} ||g(a)-g(b)||_2^2 \\
D^{min}_{between} &= \operatorname*{avg}_{x\in\mathbb{D}} \min_{a\in P(x),b\in N(x)} ||g(a)-g(b)||_2^2 \\
D^{avg}_{between} &= \operatorname*{avg}_{x\in\mathbb{D}} \operatorname*{avg}_{a\in P(x),b\in N(x)} ||g(a)-g(b)||_2^2
\end{aligned}
$$

**Table 1.** The reason to choose 50 statements. We compute L2 distance between statements belonging to same (within) and different images (between). With more candidates, sampled negatives are hard to distinguish: the min "between" distance becomes similar to the "within" distance.

| # statements | $D^{min}_{within}$ | $D^{avg}_{within}$ | $D^{min}_{between}$ | $D^{avg}_{between}$ |
|---|---|---|---|---|
| 10 | | | 1.619 | 1.937 |
| 50 | 1.395 | 1.551 | 1.437 | 1.938 |
| 200 | | | 1.324 | 1.938 |

## 4 Hard negative mining

In Table 2, we show that negative mining strategy does matter in our task. Using 32 hard negatives is better than using all or just the most challenging example.

**Table 2.** Hard negative mining with different top-k hyperparam, using batch size 128.

| | Rank (Lower ↓ is better) | | Recall@3 (Higher ↑ is better) | |
|---|---|---|---|---|
| Method | PSA | Product | PSA | Product |
| 1 NEGATIVE | 6.033 (± 0.127) | 4.647 (± 0.026) | 0.853 (± 0.015) | 1.091 (± 0.003) |
| 32 NEGATIVES (VSE++) | **4.139** (± 0.094) | **3.110** (± 0.019) | **1.197** (± 0.017) | **1.510** (± 0.004) |
| ALL NEGATIVES (VSE) | 4.155 (± 0.091) | 3.202 (± 0.019) | 1.146 (± 0.017) | 1.447 (± 0.004) |

## 5 Region-based v.s. standard attention

We wish to verify that bottom-up region-based attention is more appropriate for our task than standard attention. We refer to the AMC model [6] to implement the baseline standard attention mechanism: we divide the image into $3 \times 3$ grids, apply Inception-v4 [58] to get features per cell, obtain regional features (including original image as proposal), then predict attention distribution. Thus we keep the basic feature, number of regions, and resolution the same as for our method. Table 3 shows two ablations of our method and standard attention, for three tasks. Note +ATTENTION from the main text is not our contribution; SYMBOL REGION is. Standard attention is inferior. For PSAs, which is our focus, our symbol-based attention is the strongest attention method overall.

**Table 3.** Region-based vs. standard image attention. SYMBOL REGION uses mean pooling, REGION-BASED ATT (+ATTENTION in Tab.2 of paper) uses attention pooling over region features, STANDARD ATT applies attention pooling on evenly split $3 \times 3$ grids.

| | Statement | | | | Slogan | | | | Clustering |
|---|---|---|---|---|---|---|---|---|---|
| | Rank ↓ | | Recall@3 ↑ | | Rank ↓ | | Recall@3 ↑ | | Homogen ↑ |
| Method | PSA | Product | PSA | Product | PSA | Product | PSA | Product | All |
| SYMBOL REGION | **3.174** | 2.539 | **1.442** | 1.697 | **3.774** | 3.344 | 1.121 | 1.182 | 0.331 |
| REGION-BASED ATT | 3.258 | 2.488 | 1.428 | **1.726** | 3.850 | **3.257** | **1.155** | **1.205** | **0.355** |
| STANDARD ATT | 3.382 | **2.482** | 1.415 | 1.720 | 3.954 | 3.320 | 1.073 | 1.185 | 0.339 |

## 6 Different types of proxies

In Table 4 and Table 5, we show that not any type of label would suffice as constraint. In particular, the Ads Dataset includes 6 times more topic labels that could be used as constraints compared to symbol labels (64,325[1] vs 10,493[2] images annotated; almost every image is annotated with topics yet only around 15% images are annotated with symbols). Despite this, symbol labels give much greater benefit. Thus, [44]'s proxy approach is not enough; the type of labels must be carefully chosen.

**Table 4.** Different types of labels as constraints. The baseline method ("No extra components") uses image attention (Sec. 3.3) but does not have the components from Sec. 3.4-3.5 of the main text.

| | Rank (Lower ↓ is better) | | Recall@3 (Higher ↑ is better) | |
|---|---|---|---|---|
| Method | PSA | Product | PSA | Product |
| No extra components | 3.258 (± 0.081) | 2.488 (± 0.015) | 1.428 (± 0.017) | 1.726 (± 0.004) |
| Symbol labels | 3.171 (± 0.081) | 2.465 (± 0.015) | 1.471 (± 0.017) | 1.726 (± 0.004) |
| Topic labels | 3.186 (± 0.079) | 2.477 (± 0.015) | 1.456 (± 0.017) | 1.728 (± 0.004) |

**Table 5.** % improvement for different types of labels as constraints.

| | PSA | | Product | |
|---|---|---|---|---|
| Method | Rank ↓ | Rec@3 ↑ | Rank ↓ | Rec@3 ↑ |
| Symbol labels | 3 | 3 | 1 | 0 |
| Topic labels | 2 | 2 | 0 | 0 |

[1] The Ads dataset [22] involves 64,832 images with topic annotations. However, the actual topic labels (64,325) we can use as the constraint is the intersection of the topic annotations and the format-filtered (Sec. 4.2) statement annotations.

[2] There are 13,938 images annotated with symbols in Ads dataset [22], and we use these images to train our region proposal network in Sec. 3.3. However, not all of the free-formed symbol annotations could be transformed into the 53 symbols that [22] used. 10,493 is the number of images that have symbol annotations for the 53 symbols.

## 7 In-depth quantitative results

**Quantitative results with extra measurement.** We provide Rank and Recal@3 as the measurement of our model in the paper. In Table 6, we also compute three other metrics: Recall@10, which denotes the number of correct statements ranked in the Top-10; RankAvg, the average ranking value of the averaged-ranked true matching statement; RankMedian, the average ranking value of the median-ranked true matching statement.

The reason that we only use Rank in the paper instead of RankAvg or RankMedian is that we found that in the Ads dataset [22], there are some really noisy annotations that ruin the metrics of RankAvg and RankMedian. This could be seen from Figure 6 where RankAvg is always worse than RankMedian.

**Table 6.** Our main result with extra measurements. The best method is shown in **bold**. For Recall@3 and Recall@10, higher values (↑) are better. For Rank, RankAvg, and RankMedian, lower values (↓) are better.

| | | VSE++ | Hussain-Ranking | ADVISE (ours) |
|---|---|---|---|---|
| Product | ↑ Recall@3 | 1.510 (± 0.004) | 1.515 (± 0.004) | **1.725** (± 0.004) |
| | ↑ Recall@10 | 2.379 (± 0.003) | 2.392 (± 0.003) | **2.527** (± 0.003) |
| | ↓ Rank | 3.110 (± 0.019) | 3.093 (± 0.019) | **2.469** (± 0.015) |
| | ↓ RankAvg | 7.311 (± 0.029) | 7.122 (± 0.028) | **6.143** (± 0.025) |
| | ↓ RankMedian | 6.392 (± 0.030) | 6.297 (± 0.029) | **5.252** (± 0.026) |
| PSA | ↑ Recall@3 | 1.197 (± 0.017) | 1.258 (± 0.017) | **1.509** (± 0.017) |
| | ↑ Recall@10 | 2.089 (± 0.017) | 2.151 (± 0.017) | **2.323** (± 0.015) |
| | ↓ Rank | 4.139 (± 0.094) | 3.854 (± 0.088) | **3.013** (± 0.075) |
| | ↓ RankAvg | 9.424 (± 0.135) | 8.718 (± 0.127) | **7.553** (± 0.119) |
| | ↓ RankMedian | 8.268 (± 0.143) | 7.741 (± 0.135) | **6.394** (± 0.121) |

**Per-topic evaluation.** We provide per-topic quantitative results to further compare our ADVISE model and the two strong baselines: VSE++ [11] and Hussain-Ranking [22]. Figures 2 and 3 show the per-topic evaluation results of the statement ranking and hard-statement ranking task respectively. In both figures, we show bar charts of the best performing 10 topics (left) and the worst performing 10 topics (right), in terms of the Rank measurement of our ADVISE model (shown in blue).

Intuitively, the hard-statement ranking task defined in our paper emphasizes more the ability to distinguish the within-topic nuances, yet our main ranking task focus on both the within-topic and between-topic differences. We see from Figures 2 and 3 that ads of different topics perform differently on the two tasks. For example "baby" (baby products) ranks the first in the main ranking task, yet it is also in the worst-10 in terms of Rank in the hard-statement ranking task. Our explanation is that "baby" ads are quite distinguishable from others, yet they do not have sub-categories within themselves. Another example is "clothing", which has good performance on both tasks. The reason is that "clothing" ads

are distinguishable from others, while they can still be further classified as, e.g., "jean", "watches", "outfit", and so on.

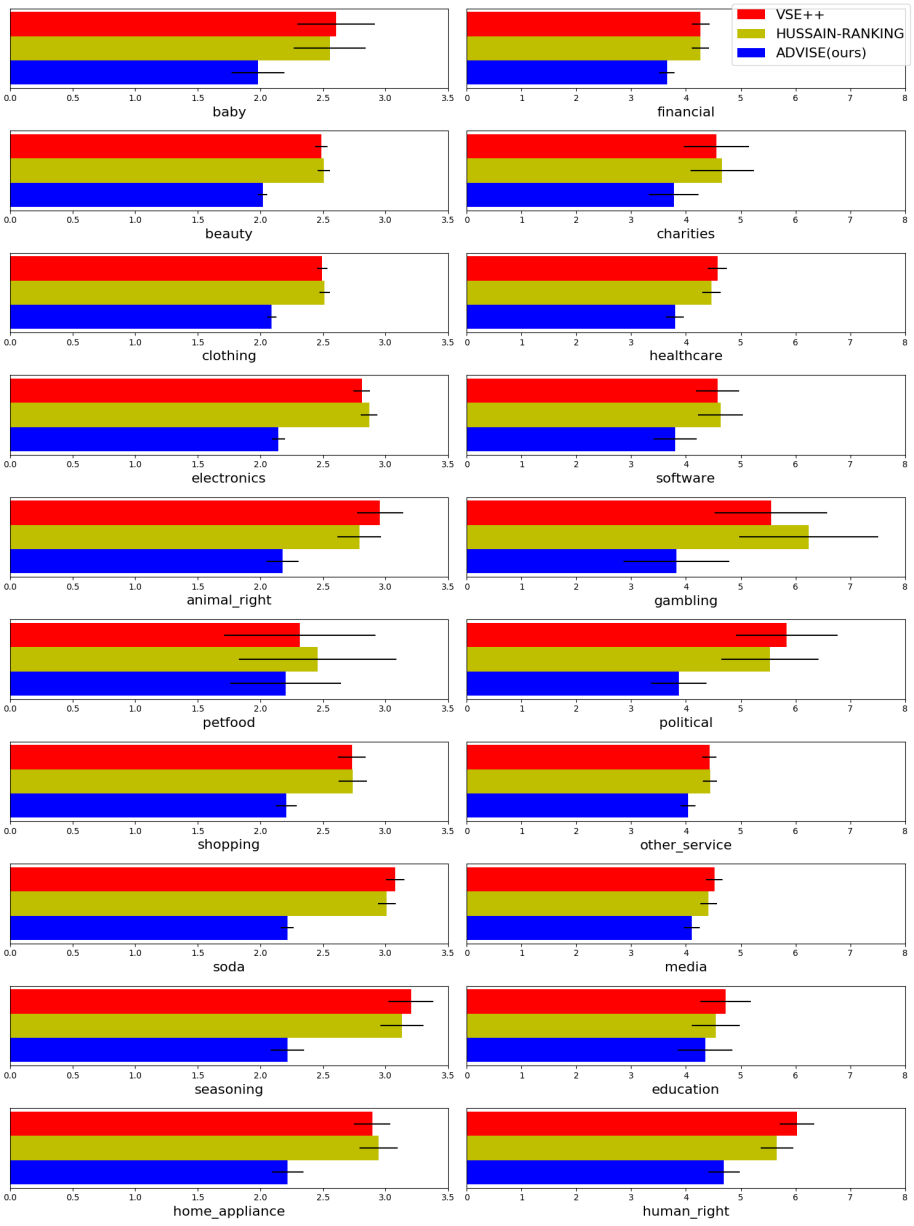


**Fig. 2.** Bar chart of the statement ranking task. The x-axis denotes the Rank measurement. Lower is better. The error bar shows the standard error, which is defined as $\delta/\sqrt{n}$ where $\delta$ is the standard deviation of the Rank and $n$ is the number of examples. Our ADVISE model is always better than the other two baselines.

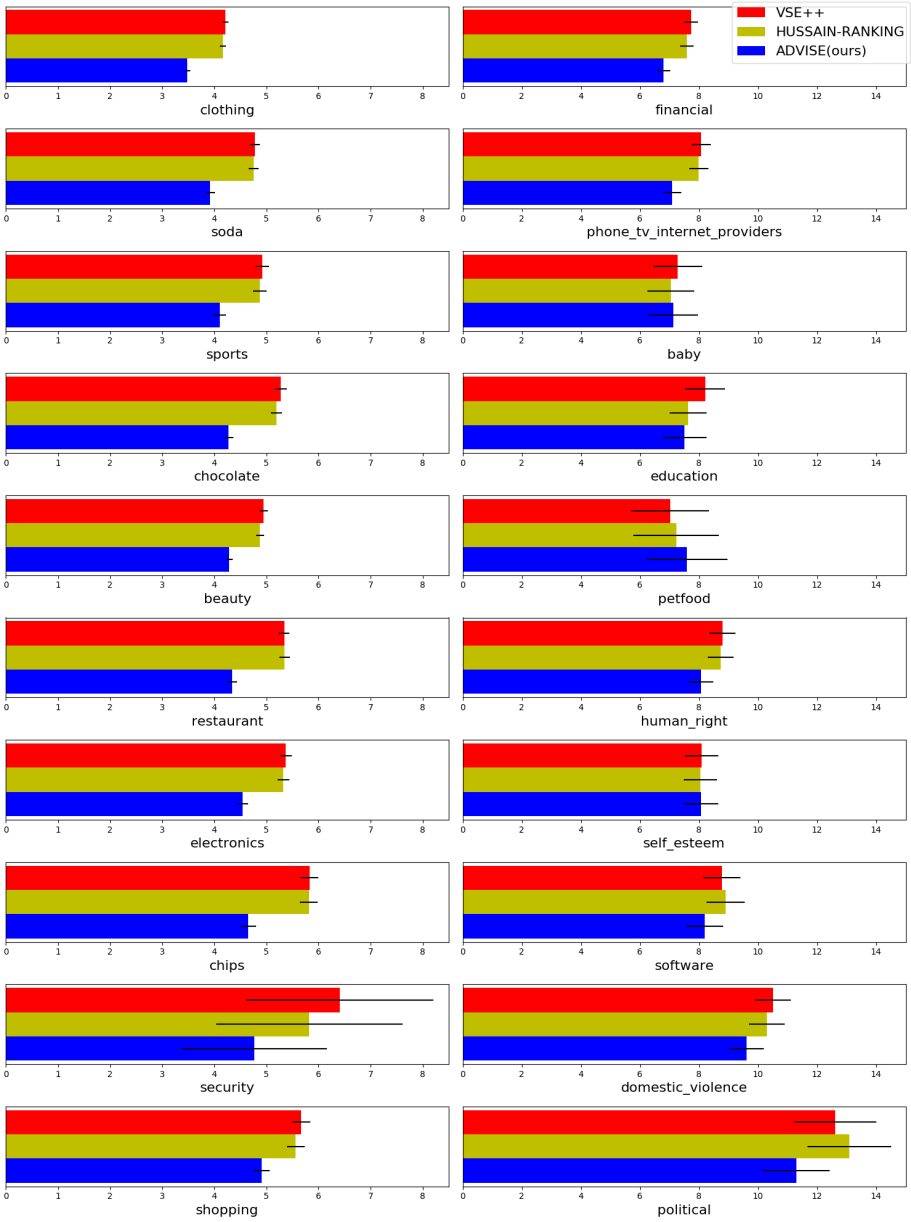


**Fig. 3.** Bar chart of the hard-statement ranking task. The x-axis denotes the Rank measurement. The error bar shows the standard error, which is defined as $\delta/\sqrt{n}$ where $\delta$ is the standard deviation of the Rank and $n$ is the number of examples. Our ADVISE model is better than the two baselines in most of the cases.

## 8 Ranking task - qualitative results

We provide more qualitative examples of both the statement ranking task and the hard-statement ranking task for both PSAs (Figure 4) and product ads (Figure 5). We can see from both figures, that hard-statement ranking task requires the model to have a deep understanding not only about the topic and the purpose of the ads but also the details such as the brand of product, the reasoning of causal relation, etc. We also see from the qualitative examples that failure in the hard-statement ranking task does not always mean we failed to understand the ads. For example, the top-1 hard-statement prediction of row 1 of Figure 4 has already captured all of the information in the ad, "I should not be on my phone while driving because it can cause an accident", yet this statement is from another similar image which causes the evaluation to not count this prediction as correct.

We see some interesting examples showing that the model understands both the images and statements reasonably well. For example, the result of row 3 in Figure 4, our ADVISE model ranks safety-related statements higher in the results, moreover, the top-ranked statements all involve the keyword "helmet" probably because the model associates the watermelon with head/helmet. Understanding ads is still challenging, and the result of row 4 in Figure 5 shows one obstacle. Our model should recognize beer, yet the 'pepsi' mentioned in the second-highest ranked statement is visually quite similar to the bottle in the image thus misleads the model.


**Fig. 4.** Statement/Hard-statement ranking task of PSA ads. We show the top-5 multiple-choice answers ranked by our ADVISE model for both statement and hard-statement ranking task. Statements in **bold** are the correct predictions.

| | | STATEMENT | HARD-STATEMENT |
|---|---|---|---|
| 1 |  | **[0.350] I should buy Revlon makeup because they are pretty and natural**<br>**[0.355] I should use Revlons lip balms and mascara because it will enhance the look of my lips and lashes**<br>**[0.392] I should buy Revlon makeup because it will enhance my features**<br>[0.444] I should use Heinz because it does not have unnatural things in it<br>[0.614] I should drink this bacardi because it makes the world seem different | [0.330] I should use Men Degree because it feels great<br>[0.346] I should buy this product because it's sexy.<br>[0.346] I should buy this lipstick because it has cool shades<br>[0.349] I should buy this product because its use keeps you alert<br>**[0.350] I should buy Revlon makeup because they are pretty and natural** |
| 2 |  | **[0.468] I should buy Polo clothes because it will make me handsome**<br>**[0.477] I should buy these clothes because handsome preppies wear them**<br>**[0.547] I should wear Polo by Ralph Lauren because it will make me attractive**<br>[0.595] I should eat Pringles because Pringles is down with the holidays.<br>[0.609] I should purchase these security cameras from this distributor because they offer a variety of products | [0.449] I should wear Chanel because it's sophisticated<br>**[0.468] I should buy Polo clothes because it will make me handsome**<br>**[0.477] I should buy these clothes because handsome preppies wear them**<br>[0.530] I should buy a Givenchy suit because it will look cool<br>[0.532] I should wear jean shorts because it's unexpected |
| 3 |  | **[0.578] I should shop on Google because there is a lot of information**<br>**[0.634] I should shop online with my phone because it is more convenient**<br>[0.711] I should sign up for this website to get black Friday deals because it helps you to save money when you're not paying regular price<br>[0.793] I should buy westinghouse appliances because they're easy to clean<br>**[0.853] I should use this because it has local inventory and related items** | [0.550] I should use a Window's Phone because I get free music for a year<br>[0.553] I should download this app because I want to hear about updates on Applebees<br>**[0.578] I should shop on Google because there is a lot of information**<br>[0.600] I should consider using this software because it can allow me to create and edit ads from my smartphone<br>[0.625] I should use admob because it is on my phone |
| 4 |  | **[0.348] I should drink Bud Light because I have good taste**<br>[0.516] I should drink more Pepsi because I cannot get enough<br>**[0.555] I should buy this beer because I can get a boyfriend**<br>[0.684] I should drink Budweiser because it will make sexy women appear<br>**[0.690] i should want to drink this because its good for women** | **[0.348] I should drink Bud Light because I have good taste**<br>[0.384] I should drink this beer because it tastes good<br>[0.430] I should drink Pilsner Beer because it will associate me with standards<br>[0.440] I should drink Schaefer beer because it is a fine tasting beer<br>[0.462] I should buy coors light because it is the worlds most refreshing can |
| 5 |  | **[0.591] I should buy a volkswagen because it can get you through the toughest terrain**<br>**[0.659] I should drive this car because its tough**<br>**[0.686] I should buy volkswagen because it can weather the storm**<br>[0.723] I should be drinking Coca-Cola because I am a star<br>[0.785] I should get an Ericsson because it takes pictures that look real | [0.536] I should use Shell because they are efficient<br>[0.536] I should purchase a boles-aero because travel in style<br>**[0.591] I should buy a volkswagen because it can get you through the toughest terrain**<br>[0.621] I should get an oldsmobile because it is a summer classic<br>**[0.659] I should drive this car because its tough** |
| 6 |  | **[0.382] I should buy a cycle ops bicycle because it will get me out of the house**<br>**[0.516] I should buy cycle oops because its healthy**<br>[0.779] I shouldn't wear fur because fur is unattractive on a person<br>[0.882] I should buy this golfing equipment because I want to use the best and not something that will fall apart<br>[0.882] I should give gifts because I can give a gift to someone religious | [0.312] I should visit the Museum of Science because there's a piece of history for everyone<br>[0.332] I should love to surf because the ad is telling me to<br>**[0.382] I should buy a cycle ops bicycle because it will get me out of the house**<br>[0.504] I should join this website So my daughter can have a better life too<br>**[0.516] I should buy cycle oops because its healthy** |

**Fig. 5.** Statement/Hard-statement ranking task of product ads. We show the top-5 multiple-choice answers ranked by our ADVISE model for both statement and hard-statement ranking task. Statements in **bold** are the correct predictions.

## 9 k-NN retrieval on image regions

After embedding each image region and weighing them using the attention mechanism, our ADVISE model learns a good image-level representation that could be used to distinguish among statements. Since the image representation is a weighted sum of region representations, regions serve as visual words in our model and the image is analogous to a sentence. In order to know if the model also has the ability to assign concept to these visual words (image regions), we made the following qualitative experiments.

**Product words as query.** We choose 11 discriminative words from the vocabulary of product ads and use k-NN to retrieve the 10 most related image regions from the test images. The retrieval results are shown in Figure 6. Though we manually choose the 11 words, the k-NN results are entirely generated by the ADVISE model. Please note that we have no such labels associated with regions at training time, that is, we only use the label-agnostic bounding box location annotations (we ignore the symbol categorical labels and semantic meaning of the box region), and the image-level statement annotations, to train the model. However, the model itself successfully learns to associate the concept with a specific region.

**PSAs words as query.** Aligning abstract words from PSA ads to the image regions is a more challenging task. We show in our paper quantitatively that our ADVISE model performs worse in PSAs than that in product ads. To understand the challenge, we show several qualitative examples in Figure 7. Similar to the previous visualization of product ads, we choose multiple discriminative words and retrieve image regions using k-NN. We retrieve the top-20 image regions for each query, merge some queries (such as “kill”, “kills”, and “killing”), and manually select (since the results are not as good as that of product ads) 10 typical examples to visualize.

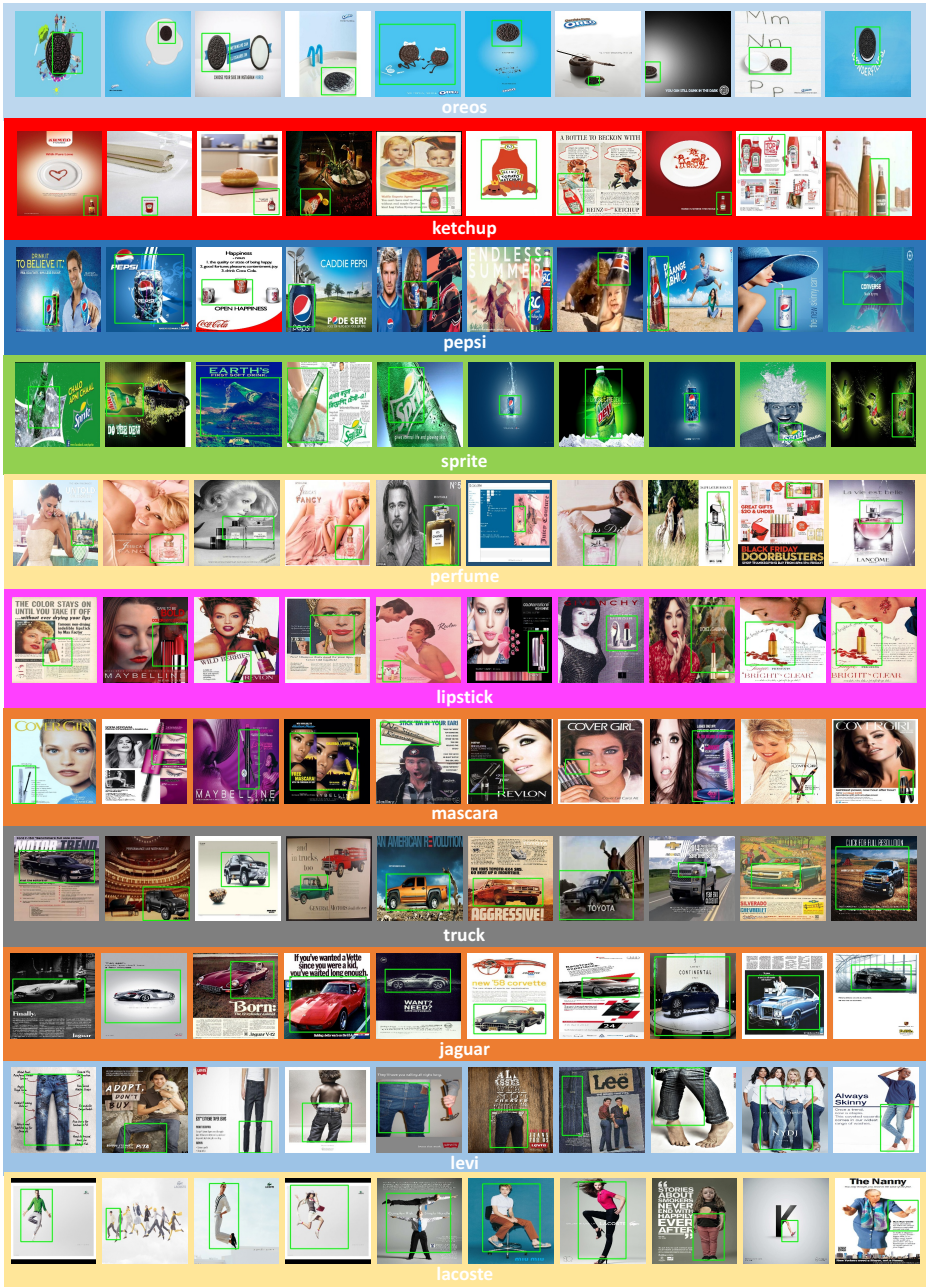


**Fig. 6.** Product words and the retrieved image regions. We select confusing word pairs such as "pepsi" and "sprite", "lipstick" and "mascara", "truck" and "jaguar". We see that our ADVISE model makes mistakes occasionally such as retrieving "pepsi" for query "sprite". However, the model knows the nuances in general.


**Fig. 7.** PSA words and the retrieved image regions. It is a more challenging task to associate PSA words with image regions since the words in PSAs tend to be more abstract than that in product ads. The qualitative examples such as "warming" and "litter" remind us that the embedding of image region and the attention mechanism may also depend on the other regions in the same image (e.g. in case that a beautiful woman exists, polar bear does not symbolize "warming" any longer), which is an interesting research direction for our future work.