# Symbolic VQA on Visual Advertisements with SymViSe Networks

Rohan Doshi William Hinthorn
Princeton University
{rkdoshi, hinthorn}@princeton.edu

## Abstract

*Understanding the symbolic rhetoric of advertisements is difficult. An agent must recognize creatively portrayed objects, their relationships and cultural connotations, and perform common-sense reasoning. To that end, we will be tackling a new task defined by the CVPR Automatic Understanding of Visual Advertisements workshop. For each visual advertisement image, we must predict one of the three correct labels amongst a set of 15 action-reason statements. An advertisement of a woman squeezing a dog as a perfume bottle and causing the dog pain would match the following action-reason statement: "I should not buy products that are tested on animals because it is very cruel." To attempt this task, we adapt a popular VQA architecture to implement a 2-stream ViSe (Visual Semantic) Network, which takes an image and candidate action-reason statement and predicts a probability that the two match; this helps demonstrate the difficulty of the task and serves as a baseline of existing techniques. Then, we add a third symbolic stream to ViSe to construct a novel model, which we coin as the SymViSe (Symbolic Visual Semantic) Network. We further demonstrate how applying attention mechanisms can boost the performance of these two networks. Our best SymViSe model with attention achieves a 37.62% validation accuracy, significantly higher than the naive approach of guessing randomly (20%) and anything found in the literature. We also further isolate the magnitude of influence of the symbolic stream and attention mechanisms on task performance. We conduct a further error analysis to understand common failure modes.*

## 1. Introduction

Recent developments in computer vision and natural language processing have set the stage for more challenging AI-complete visual reasoning tasks. Current VQA datasets primarily contain natural images (e.g. images taken by photographers in everyday life) that lack the higher level of rhetorical complexity that is found in other contexts such as media and advertisements. Thus, while visual question answering (VQA) and other similar tasks have made progress in visual reasoning, more complex datasets are required to teach models to learn and leverage higher levels of reasoning necessary for understanding images in real world applications.

Reading visual advertisements requires the inference not only of the physical contents of an image but also of their cultural connotations, expectations, and other properties that contribute to the complex abstract ”message” being conveyed. For example, an advertisement of a woman squeezing a dog as a perfume bottle and causing the dog pain would match the following action-reason statement: ”I should not buy products that are tested on animals because it is very cruel.” This type of rhetorical reasoning is challenging because (1) objects are often portrayed in unorthodox ways (e.g. dog as a perfume bottle), (2) objects are juxtaposed in unconventional ways (e.g. human face and dog), and (3) common knowledge is required (e.g. animal testing is done on dogs). Additional emotional concepts, such as the love pet owners have for their pets, are extremely difficult to understand.

Based on the discussion above, it is apparent that advertisements expose the types of visual understanding challenges that could help spur the next wave of computer vision developments. To this end, CVPR has recently released the Visual Advertisement Dataset. In this work, we wish to decode the messages that ads seek to convey by identifying the implied action the ad wishes the viewer to take alongside the reasoning for performing this action that the image depicts. For each ad, a model must select one of three such correct action-response[1] pairs from amongst a set of 15 choices.

This complex task requires us to explore new ways of representing visual information so that models can connect abstract concepts with visual content.

## 2. Related Work

### 2.1. Visual Question Answer (VQA)

Developments in natural language processing (NLP) and computer vision (CV) tasks have enabled new tasks at

[1] [8] refers to these as “question/answer” pairs.

the intersection of the two fields. Image captioning has built upon computer vision developments in object recognition by integrating convolution neural network architectures such as VGG [13] and ResNet[6] for learning image embeddings [7]. For instance, [4] infers captions from visually similar images that are clustering nearby in image embedding space. Going further, [9] leverages NLP work on bidirectional Recurrent Neural Networks (RNNs) to align image embeddings in sentence embedding space for better generating image descriptions.

More recently, researchers have turned their attention to visual question answering (VQA), a more challenging task that requires higher level reasoning enabled by the integration of vision and language models. The VQA task aims to provide a natural language answer (***A***) to an input visual image (***I***) and natural language question (***Q***). VQA tasks are more difficult than the aforementioned tasks for many reasons stemming both from the vision and language domains:

1. **Vision:** Apart from identifying objects, higher level visual reasoning is required to resolve references to objects, the physical and functional relationships between them, and their meaning within a larger cultural context.

2. **Language:** Apart from capturing n-gram priors, advanced semantic reasoning is required to go from perception (e.g. identifying attributes such as color, number, and location) to deduction (e.g. how, what, and why). And, given the many types of questions, the model must scale to thousands of question types, which inevitably blurs semantic boundaries to challenge the model with ever greater levels of semantic ambiguity.

To facilitate VQA, new datasets have evolved. The first major dataset was Darquar, which made the first attempt at establishing a "Visual Turing challenge." Building upon this, the VQA1 dataset provided two orders of magnitude more data and a wider set of questions across two broad classes (open-ended and multiple choice) to drive the development of a new generation of VQA models, yet even with this improved dataset, models were able to ignore visual information and leverage language priors to artificially predict answers well. This bias-driven accuracy motivated the creation of VQA2, which balances the probability of an answer and its logical compliment given an image-question pair (a.k.a. $P(A \mid Q\&I) = P(A' \mid Q\&I)$).

However, VQA still faces considerable challenges, especially in reasoning over more complex images and questions reflecting metaphorical and abstract concepts. Many of these shortcomings become more evident in datasets such as the visual advertisement dataset, which we'll focus on now.

## 3. Visual Advertisements Dataset

The Visual Advertisements Dataset is a collection of 64,832 image ads and 3,477 video ads organized and annotated by Hussain et al. in [8]. Due to the inherent difficulty and cost of model development for the video understanding task, we choose to focus only on the still image ads for this project. These images already are difficult to parse, requiring a significant amount of prior cultural knowledge and physical reasoning abilities to understand the multiple levels of hidden messages they contain. Thus, we believe it will be worthwhile for us to focus on images. The dataset contains:

1. 204,340 topical annotations (i.e. the product or idea being sold or promoted by the advertisement)

2. 102,340 annotations indicating the sentiment an ad provokes (e.g. "amused" by a joke or "grateful" for the image subject's service)

3. 202,090 unprocessed, free-form action-reason responses to the question "what should I do according to this ad, and why should I do it?" (e.g. "I should buy Panasonic cameras because they have facial recognition software.")

4. 64,131 symbolic references grounded in bounding boxes to identify objects which alludes to certain abstract symbolic concepts (e.g. "danger" might be represented by a rattlesnake, or "ice" might symbolize freshness in the case of a gum ad)

5. 20,000 strategy annotations (e.g. whether an ad "contrasts" the product or idea being promoted with a popular competitor)

6. 11,130 slogans (e.g. always tea time) [2]

### 3.1. Difficulty

Advertisements pose many difficult visual reasoning challenges. Take for example 1. The simplicity of the image seeks to persuade the viewer to fly British Airways for its sophisticated elegance - one of the QA annotations states that "I should fly this airline because it is simple." The focus of the image is a Swiss Army Knife, a symbol for the nation Switzerland, the country in which Zurich lies. It is meant to be clever, since the knife is folded out slightly so as to imitate the silhouette of an airplane, the object symbolizing the services the airline is selling. The knife itself is known for its versatility and usefulness, both qualities the

[2] Slogans were developed by the ever-reliable and boundlessly creative Mechanical Turk workers.

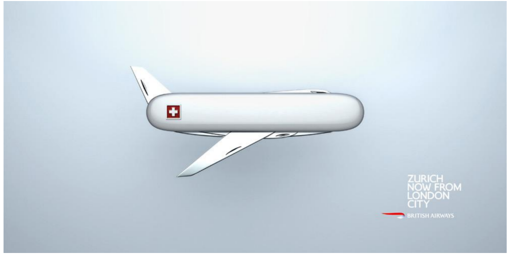


Figure 1: This British Airways ad promotes a new flight route “Zurich now from London City.” An analyzing agent with a strong center bias (commonly seen in image classification tasks) would, like one of the careless MTurk annotators, incorrectly believe this ad to be selling a pocket knife.

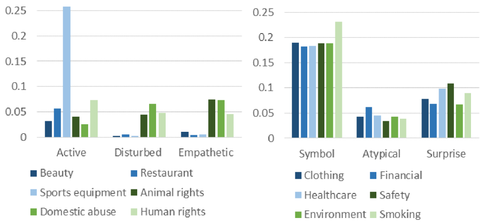


Figure 2: Strong correlations exist between the the topics of an ad and the sentiments the advertisers seek to evoke (left) as well as the strategies used to persuade the viewer (right). The categories visualized are a smalls subset of those used for the challenge selected for visualization purposes.

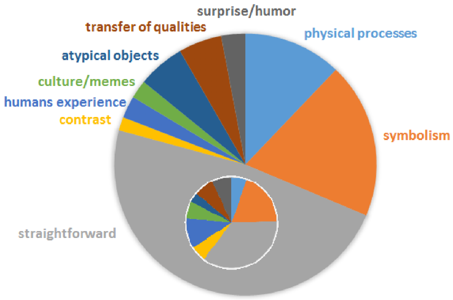


Figure 3: Frequency of common ad strategies in the Visual Advertisement dataset. Note that the main graph depicts the annotations from the authors of [8]; the inset shows annotations collected on Amazon Mechanical Turk.

airline wish the viewer to subconsciously attribute to themselves. Current computer vision models struggle to recognize this type of subtext and will continue to be imperfect even after using this dataset due to the incredible noisiness of the annotations. One of the annotators responded “BUY THIS SWISS KNIFE BECAUSE ITS COOL,” failing to realize that the knife isn’t actually the object being sold. Even careless humans would perform poorly on this task. If the dataset indeed is well-balanced, any model which performs well for this challenge would have to learn a broad set of cultural priors with which to reason about complex messages.

### 3.1.1 Dataset Analysis

Advertisers employ a wide-ranging set of strategies in order to push their products and ideas. For instance, in Figure 1, British Airways employs the symbolism of a swiss army knife to make customers associate their airline with efficient and elegance. The organizers of the Visual Advertisement challenge would categorize this ad strategy as “symbolism”. Figure 3 shows the breakdown of categories into which the authors have binned the ads.

As seen in 3, about half of the ads employ a straightforward messaging strategy; they are the most easily decoded ads and so could be interpreted using traditional object detection and scene parsing models. However, the next most common category, responsible for about a quarter of the images, requires comprehending the abstract symbolic meaning of the objects within the image. To drive performance among these images, we focus our research efforts towards inferring symbolic meaning from image objects. There are a total of 46,627 annotated symbols in the ad dataset, though many of these responses provide multiple symbolic labels separated by commas or a forward-slash (“/”). If we split the annotations on the “/”, the total number of annotations rises to 66,376; however, by grouping similar symbols together, we may decrease the total number of unique labels from 24,171 to 14,153. Furthermore, we recognize that the distribution of symbols has an extremely long tail. As figure 4 shows, the 10 most common symbols all have frequencies ranging from 411 to 863 occurrences, [3] while 9382 symbols appear only once. We prune uncommon symbol annotations for .all of our training samples to focus on tactics frequently employed by advertisers.

## 3.2. Models

The authors of the challenge proposed a simple VQA model as a baseline for a related task (221-way symbolic classification). First, this model concatenates features extracted by VGG-16 with those encoded from the word embeddings of a question processed by a 2-layer LSTM. Then, it runs this visual-semantic vector through a 1000-way clas-

[3] The top 10 categories ( “fun”, “beauty,” “nature,” “danger,” “sex,” “adventure,” “health,” “natural,” “love,”, and “power”) appear 863, 806, 749, 666, 606, 524, 493, 450, and 411 times, respectively.

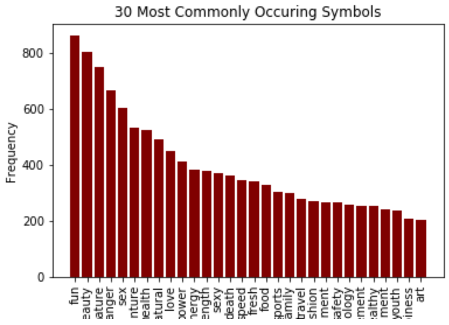


Figure 4: The most common symbols appear at sufficient frequencies to use for inference.

sification via a single fully-connected layer followed by a softmax layer. Each label corresponds to a single word, and the inference is considered "correct" if the word matches the word with the highest tf/idf score in any of the three responses. They obtain an accuracy of 11.96% on this related but different question-answer task but don't report results on the challenge task we tackle in this work. This shows the degree of complexity involved with understanding visual rhetoric.

# 4. Methodology

## 4.1. Network Proposal Overview

We propose a baseline 2-stream ViSe (Visual Semantic) Network based on existing visual-semantic VQA approaches and a novel 3-stream SymViSe (Symbolic Visual Semantic) Network for attempting the visual advertisement task.

- **ViSe Network** This network is composed of distinct visual and semantic streams that take an image and a candidate action-reason sentence, respectively, as input to extract visual and semantic features. This builds on the visual-semantic VQA framework that won the 2017 VQA challenge; the traditional "question" phrase used as input in a VQA formulation here is taken to be the concatenated "action-reason" statement [15]. The visual stream extracts visual features using bottom-up attention, as proposed in [2] and further explained in Section 4.1.4. The semantic stream looks up and processes word embeddings for each word in the input sentence, as explained in Section4.1.2. At a high level, the network jointly embeds visual and semantic features into a joint space used to infer two output probabilities for the 0 and 1 classes. The output indicates the level of correspondence between the textual label and the image. This approach optionally can use the text embeddings to attend to the image features in a manner similar to [12].
- **SymViSe Network** As shown in Figure 5, this network integrates a third symbolic stream with the ViSe Network, which takes the image as input and outputs symbolic features using a second deep visual network constructed like the one used in the visual stream. This visual encoding network is fine-tuned to infer symbolic meaning from objects in the image using the symbolic annotations in the Visual Advertisement dataset. These symbolic features are combined with the visual and semantic features extracted from the other streams and projected into a joint-embedding space. The visual and symbolic streams both are multiplied with the semantic stream, concatenated, and sent to a linear fully-connected classification layer to infer a confidence score for agreement between the image and input sentence. This approach also can have the text embeddings attend to the visual and symbolic embeddings in a manner similar to [12].

### 4.1.1 Visual Stream

One way to represent a visual advertisement is through its objects. Since no objects are annotated in the ads dataset, we pre-train a Faster R-CNN model to detect the 80 objects in the COCO [10] dataset. We then directly apply this model as an image encoder using the methods presented in Section 4.1.4 to produce $K = 100$ feature vectors of 1024 dimensions. These vectors are L2 normalized to be scaled into a standard range to facilitate cross-modal transfer learning.

### 4.1.2 Semantic Stream

As shown in figure 5, we choose to model each action-reason sentence as a 20x300 feature vector composed of concatenated GloVe embeddings for each word, where sentences of lengths less than 20 words are zero-padded to maintain a constant size. For sentences longer than 20 words, only the first 20 words are used. These textual features are encoded using a two-layer LSTM network and projected into a joint-embedding space via a fully connected non-linear layer.

### 4.1.3 Symbolic Stream

Visual advertisements contain abstract semantic messages that cannot be captured directly by the objects they contain. We wish to teach our model to understand the symbolic meanings of objects and the importance of their position within the image. We train a Faster R-CNN model identical to that used in Section 4.1.1 to extract image regions that correspond to the 1000 most frequent symbols

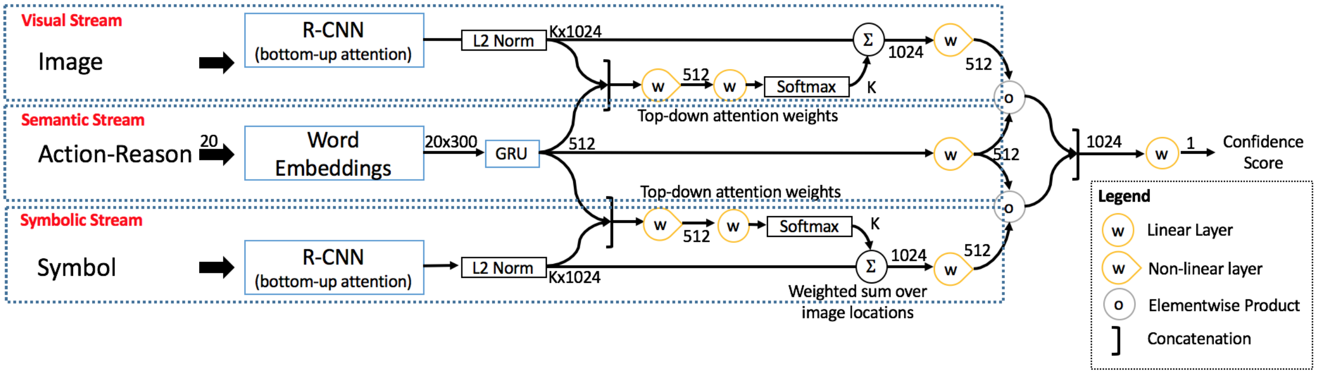


Figure 5: The proposed SymViSE network to attempt the symbolic multi-choice VQA task on the visual advertisement task. SymViSe extracts a visual, semantic, and symbolic feature vector from three different streams before combining and projecting them into a joint-embedding space for ultimate binary classification.

in the Visual Advertisement dataset. To select the candidate symbol classes, we first preprocess the free-response symbol annotations. We attempt to correct misspelled symbol names using an auto-correct tool. [4]. Words are lemmatized to disambiguate different conjugations of a single symbol. In order to count the total number of instances of each at symbol, we map each word in the training set to its corresponding stem before obtaining the tf/idf scores of the words. We select the 1000 words with the highest rate of occurrence in the training set as the candidate set. We wish to associate each bounding box with a single valid symbol name, so for each box, we select the symbol with the highest tf/idf score as its associated symbol. We drop all bounding boxes that cannot be mapped to a valid symbol class.

We train this model separately using the ground-truth symbolic bounding box annotations present in the ads dataset. This task is more difficult than normal object detection since symbol labels such as “fun” or “dangerous” are ascribed to objects with very different visual appearances.

In the ViSe and SymViSe models, the R-CNN outputs $K = 100$ symbolic feature vectors of 1024 dimensions, which are L2 normalized to be scaled into a standard range to facilitate cross-modal transfer learning.

#### 4.1.4 Bottom-Up Attention for Visual/Symbolic Feature Extraction

One common technique for modeling visual features is to extract the activity maps from a model trained on a basic task like image classification. This method retains representations of the background class and can become dominated by features are less representative of the salient objects in the image. When a human examines a scene, s/he may focus on objects consciously using high level attention, and s/he also may be drawn to regions based on low-level characteristics. Bottom-up attention attempts to model this latter effect by extracting features corresponding with the objects in an image. We adapt the bottom-up attention model proposed by [2] to encode image features in the visual and symbolic streams. We adapt a Faster-RCNN [11] model with a 101 layer ResNet feature extractor to the task. We rank bounding box predictions based on the model’s confidence in the presence of an object or symbolic feature. For each of the top $K$ predictions, we mean-pool the feature vectors in the upsampled activity map within the region enclosed by the corresponding box. Thus, each prediction is associated with a single feature vector representing the semantic content of the salient region. We use extraction modules fine-tuned to detect different types of semantic content.

#### 4.1.5 Top-Down Attention Mechanism

The model described in section 4 does more than vote based on a set of visual, semantic, and symbolic features we have selected. While it does exclude image regions of low importance using a bottom-up attention mechanism, this representation is generic. It would be better if the model could attend to certain extracted features conditioned on the semantic content of the action-reason statement. To accomplish this, we incorporate a top-down, text-guided attention mechanism via gating to select the visual and symbolic features that best correspond with the correct label. We adapt the attention mechanism described in [12] and [14] to attend to image features based on linguistic cues. Concretely, we concatenate the language embedding $q_i$ to each spatial image embedding $v_i$ and use a two-layer CNN $f$ with 1x1 kernels and a gated tanh activation function to output a binary attention map normalized with the softmax function. This is used to compute a weighted sum over the feature

[4] https://github.com/phatpiglet/autocorrect/

map. Formally:

$$a_i = f_a([v_i\{q_i, r_i\}])$$

$$\alpha = \text{softmax}(\alpha)$$

$$\hat{v} = \Sigma_i \alpha_i v_i$$

The attention function here $f_a$ is modeled using a nonlinear gated tanh module, the main non-linearity used throughout the ViSe and SymViSe networks,[5] composed with a $1 \times 1$ convolution. The gated tanh module's output $y$ is calculated for a given input $x$ as defined below:

$$\tilde{y} = \tanh\left(Conv_1(x)\right)$$

$$g = \sigma(Conv_2(x))$$

$$y = \tilde{y} \odot g$$

Here, both $Conv_1$ and $Conv_2$ are $1 \times 1$ convolutions. This module is used in highway networks to improve gradient flow and allow the network to select different regions to retain for further computation. Many existing attention-based VQA models pool image features to reason about the entire image. We choose to let the linguistic features separately attend to different features of the two image-based representations of the input: the visual and symbolic features.

If running either model without the top-down attention mechanism, directly average the $K$ different 1024-dimension vectors extracted by the R-CNN for the visual and symbolic stream without incorporating the top-down attention weights at all.

#### 4.1.6 Multi-modal Fusion: Combining the Streams

After applying another non-linear layer of weights to the visual, semantic, and symbolic streams, we get a 512-dim vectors from each of the three streams. We have the semantic stream act to gate the other two streams by element-wise multiplying (a.k.a. taking the Hadamard product) the semantic features with each the visual and symbolic feature vectors and then concatenating the two product vectors for a final 1024-dim vector that is linearly mapped via a fully-connected layer to a binary class output.

### 4.2. Training Methodology

We update the weights of the model using an Adam optimizer with a learning rate of 1e-3, using a batch size of 512 image-sentence pairs. We train for 20 epochs.

### 4.3. Evaluation Metrics

When determining model performance, accuracy is a clear starting point. For each candidate image, we consider 15 action-reason statements. If one of three correct statements are chosen, a hit is recorded in favor of accuracy. Otherwise, it is a miss. Thus, the naive baseline of randomly guessing a statement will return an accuracy of 20%. Yet, accuracy fails to capture information about the false positives and false negatives, which is why we compute the model's precision and recall. Precision communicates the number of action-response pairs predicted to describe an image that are correct. Recall displays the portion of the true action-response pairs that are correctly identified by the model. One may combine these two scores in a single summary metric, the F1 score. This is the harmonic mean of the precision and recall. We denote the number of true positives, false positives, true negatives, and false negatives as TPs, FPs, TNs, and FNs respectively. Using these values, we can calculate the accuracy, precision, recall, and F1 score using the following equations:

$$accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{1}$$

$$precision = \frac{TP}{TP + FP} \tag{2}$$

$$recall = \frac{TP}{TP + FN} \tag{3}$$

$$F1 = 2\frac{precision * recall}{precision + recall} = 2\frac{TP}{2TP + FP + FN} \tag{4}$$

### 4.4. Code Implementation

The code is implemented using PyTorch, a python deep learning library optimized for GPU utilization. The code for this project is publicly available on GitHub and will remain open-sourced [5]. It adapts an existing VQA implementation [1] based on the approach of the first place winner of the 2017 VQA challenge [15].

## 5. Results

### 5.1. Accuracy

We report the evaluation metrics of the ViSe and SymViSe models on the visual advertisement task in table 1 with and without top-down attention. The results show that both of our novel contributions—(1) the symbolic stream in SymViSe and (2) using the semantic embeddings to guide top-down attention on visual and symbolic embeddings—are effective at improving the models' ability to interpret visual rhetoric.

First, we see that SymViSe outperforms ViSe both without top-down attention (34.58% vs 30.84% accuracy) and

[5] The Visual and Symbolic streams use rectified linear units in their base ResNet structures.

| Model | Metrics (%) | | | |
|---|---|---|---|---|
| | Acc | Precision | Recall | F1 |
| ViSe | 30.84 | 32.35 | 27.34 | 29.64 |
| ViSe w/ Attention | 32.24 | 34.39 | **32.19** | 33.25 |
| SymViSe | 34.58 | 36.11 | 30.08 | 32.82 |
| SymViSe w/ Attention | **37.62** | **36.94** | 31.95 | **34.26** |

Table 1: Accuracy, precision, recall, and F1 scores for the ViSe Network (2-streams: visual and semantic) and SymViSe Network (3-streams: visual, semantic, and symbolic) with and without top-down attention mechanisms

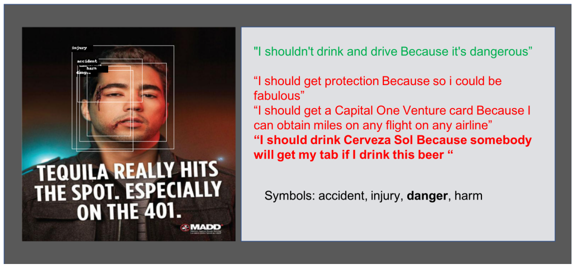
with top-down attention (37.62% vs 32.24% accuracy). Thus, the introduction of a symbol stream results in a performance boost of roughly 3 to 5%. This implies that the symbolic stream allows the model to correlate relevant symbolic image features with other visual and semantic features for greater higher-level reasoning ability on the symbolic visual rhetoric.

Second, when comparing models that incorporate top-down attention with those that directly predict correspondences, we find that the attention improves scores across the board (32.24% vs 30.84% accuracy for ViSe and 37.62% vs 34.58% accuracy for SymViSe). It appears top-down attention gives roughly a 3% accuracy boost.

## 5.2. Error Analysis

We examine images where the ViSe and SymViSe networks differ in their predictions to understand common modes of error and the unique behaviors of each model. Take for example Figure 6 . ViSe incorrectly predicts: ”I should drink Cerveza Sol because somebody will get my tab if I drink this beer.” It was unable to interpret the text in the image and furthermore does not have the cultural understanding that the 401 is a road with a high rate of traffic accidents. While the allusion to Cerveza, a form of alcohol, is correctly hinting at the true meaning of the advertisement since this poster is meant to steer young people away from drinking and driving, the network fails to understand the symbolic meaning *behind* the bruised face and appearance of intoxication. Figure 6 summarizes the symbol annotations that the symbolic stream attempts to capture, such as “accident,” or “danger.”’This perhaps helps explain why SymViSe was able to predict correctly: ”I shouldn’t drink and drive because it is dangerous.” It appears that SymViSe, unlike ViSe, was able extract and reason over the symbolic meaning of the bruised and bloodied face.

To explore the predictive relevance of each of SymViSe’s branches, we trained a network using top-down attention and added a simplified classification head. We replace the final combination of each stream’s feature vectors via



"I shouldn't drink and drive Because it's dangerous”

“I should get protection Because so i could be fabulous”

“I should get a Capital One Venture card Because I can obtain miles on any flight on any airline”

**“I should drink Cerveza Sol Because somebody will get my tab if I drink this beer “**

Symbols: accident, injury, **danger**, harm

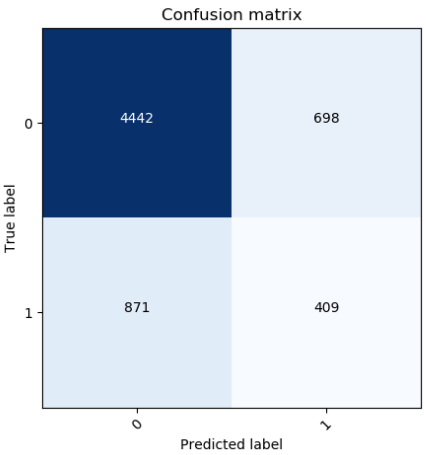
Figure 6: An ad incorrectly identified by ViSe, along with incorrect action-reason statements (red) and one of the correct statements. Symbol predictions are shown in bounding boxes, with the categories listed at the bottom of the light grey box.

element-wise multiplication with a direct concatenation of the features extracted from the visual, semantic, and symbolic streams. We then classify using a linear layer. This predicts a confidence score on the likelihood that a single action-reason sentence corresponds with an image. Each prediction is independent of any other sentence in the dataset. The $L2$-norms of the weights connecting these three feature vectors to the output tensor are 5.0378, 1.4754, and 1.5540 respectively. This indicates that the majority of the predictive power of the model is derived from the features yielded by the visual stream; thus, the language and symbol features are given a lower priority.

When we run the SymViSe model without the top-down attention mechanism on the same image, we find that the $L2$-norms of the visual, semantic, and symbolic classifier weights are 0.8108, 1.5355, and 1.1223, respectively. Note how visual and symbolic features decrease in priority, but the semantic embeddings increase in priority. The semantic embeddings are no longer attending to the other two streams, so the only way for the semantic content to contribute to the final inference is through this final classifier. This information isolation forces the network to increase the overall weight assigned to the semantic stream.

### 5.2.1 Confusion Matrix

We wish to understand the network’s points of failure by analyzing the positive and negative instances of each action-reason independently agreeing with the image. Table **??** depicts a confusion matrix for the test predictions for the SymViSe Network with top-down attention. Note how the model was far more likely to predict a positive instance as a negative sample than vice versa. In other words, most of the positive samples (image and sentence agree) are misclassified as not agreeing, so the recall of positive examples is low. This can be attributed to how there are 4x as many negative samples as positive samples. After all, training im-



Testing Confusion Matrix for the SymViSe Model with Attention

ages come with 12 incorrect action-reason statements for every 3 correct action-reason statements. So the model is unintentionally applying a biased prior resulting in a high frequency of false negatives.

### 5.2.2 Over-fitting

It appears the model may be over-fitting the training data, resulting in lower generalized performance on the validation dataset. For instance, as we see in Figure 7, after roughly 15 epochs, the training accuracy converges to 1. This causes the network's total loss to converge to 0, as shown in Figure 8. It appears that our network might be over-parametrized, enabling it to over-fit the small 10K image training dataset. In other words, our model has enough degrees of freedom and weights to ”memorize” the training dataset. This reduces the ability of our model to generalize to new data.

There are a number of steps we can take to address this. First, our network could train on a larger dataset (e.g. on the order of magnitude of the VQA dataset, which has roughly 60 times as many training images as the Ads dataset). Second, the network could reduce the number of trainable parameters by removing unnecessary layers and weights or by using bottleneck functions similar to ResNet. We also could use a regularization technique such as Dropout.

## 6. Conclusion

In this paper, we have shown that the use of top-down attention and the explicit encoding of symbolic content in images greatly assists computer vision systems in their ability to reason about visual rhetoric, especially as found in visual advertisements. Our proposed SymViSe network using attention, which leverages three processing streams (visual, semantic, and symbolic), performs surprisingly well on the multiple-choice VQA task for the visual advertisement dataset, achieving a 37.62% accuracy rate on the validation set, out-performing baseline models without top-down attention or without the symbolic stream.

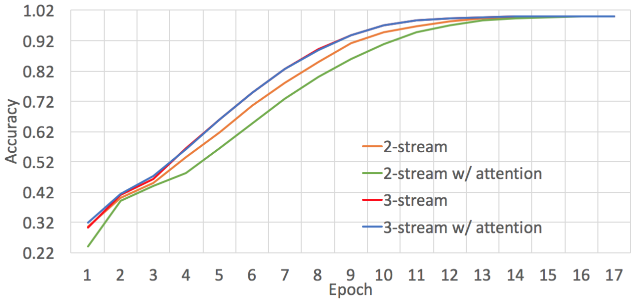


Figure 7: Training accuracy curve for 2/3-stream model with/without attention mechanisms

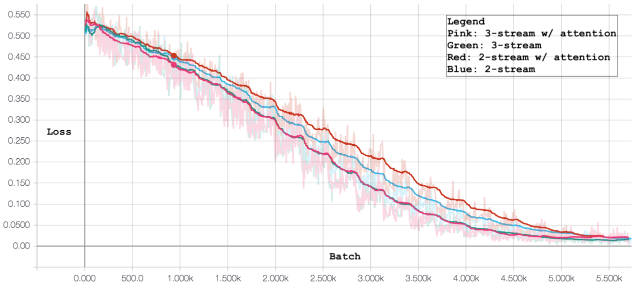


Figure 8: Training Loss Curve for 2/3-stream model with/without attention mechanisms

However, many straightforward modifications to the inference pipeline are likely to bring about large improvements.In order to strengthen the signal captured in the semantic stream, we plan on incorporating the Stanford sentence parser ([3]) to bin classes in a manner similar to [12]. This may give us more expressive representations of the text input than what we are getting by padding and cropping variable length sentences. This will build on the accuracy gains we already see from our top-down attention module, since more semantically meaningful word embeddings will be used to attend to specific image features.

We additionally would like to extend the cross-domain fusion techniques proposed above to incorporate direct supervision of ad topics and strategies. It seems intuitive that conditioning the model's inference on an understanding of the product or company behind an ad would help the model learn richer features and allow it to better discern between appropriate action-response statements.

## References

[1] vqa-winner-cvprw-2017, 2017.

[2] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, and L. Zhang. Bottom-up and top-down attention for image captioning and visual question answering. *arXiv preprint*.

[3] M.-C. De Marneffe, B. MacCartney, C. D. Manning, et al. Generating typed dependency parses from phrase structure parses. In *Proceedings of LREC*, volume 6, pages 449–454. Genoa Italy, 2006.

[4] J. Devlin, S. Gupta, R. Girshick, M. Mitchell, and C. L. Zitnick. Exploring Nearest Neighbor Approaches for Image Captioning. pages 1–6, 2015.

[5] R. Doshi and W. Hinthorn. Adsvqa, May 2018.

[6] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. *CoRR*, abs/1512.03385, 2015.

[7] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. corr abs/1512.03385 (2015), 2015.

[8] Z. Hussain, M. Zhang, X. Zhang, K. Ye, C. Thomas, Z. Agha, N. Ong, and A. Kovashka. Automatic understanding of image and video advertisements. *CoRR*, abs/1707.03067, 2017.

[9] J. Johnson, A. Karpathy, and L. Fei-Fei. DenseCap: Fully Convolutional Localization Networks for Dense Captioning. 2015.

[10] T. Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft COCO: Common objects in context. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 8693 LNCS(PART 5):740–755, 2014.

[11] S. Ren, K. He, R. Girshick, and J. Sun. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. 2015.

[12] K. J. Shih, S. Singh, and D. Hoiem. Where To Look: Focus Regions for Visual Question Answering. 2015.

[13] K. Simonyan and A. Zisserman. Very Deep Convolutional Networks for Large-Scale Image Recognition. pages 1–14, 2015.

[14] D. Teney, P. Anderson, X. He, and A. v. d. Hengel. Tips and tricks for visual question answering: Learnings from the 2017 challenge. *arXiv preprint arXiv:1708.02711*, 2017.

[15] D. Teney, P. Anderson, X. He, and A. van den Hengel. Tips and tricks for visual question answering: Learnings from the 2017 challenge. *CoRR*, abs/1708.02711, 2017.