Equal But Not The Same: Understanding the Implicit Relationship Between Persuasive Images and Text

Mingda Zhang
mzhang@cs.pitt.edu
Rebecca Hwa
hwa@cs.pitt.edu
Adriana Kovashka
kovashka@cs.pitt.edu

Department of Computer Science
University of Pittsburgh
Pittsburgh, PA, USA

Abstract

Images and text in advertisements interact in complex, non-literal ways. The two channels are usually complementary, with each channel telling a different part of the story. Current approaches, such as image captioning methods, only examine literal, redundant relationships, where image and text show exactly the same content. To understand more complex relationships, we first collect a dataset of advertisement interpretations for whether the image and slogan in the same visual advertisement form a *parallel* (conveying the same message without literally saying the same thing) or *non-parallel*relationship, with the help of workers recruited on Amazon Mechanical Turk. We develop a variety of features that capture the creativity of images and the specificity or ambiguity of text, as well as methods that analyze the semantics within and across channels. We show that our method outperforms standard image-text alignment approaches on predicting the parallel/non-parallel relationship between image and text.

1 Introduction

Modern media are predominantly multimodal. Whether it is a news article, a magazine story, or an advertisement billboard, it contains both words and images (and sometimes audios and videos as well). Effective communications require the different channels to complement each other in interesting ways. For example, a mostly factual news report about a politician might nonetheless be accompanied by a flattering photo, thus subtly conveying some positive sentiments towards the politician. A car ad might juxtapose evocative adjectives such as "powerful" next to images of horses and waterfalls, inviting the viewers to make the metaphorical connections.

While recent work has made advances in making *literal* connections between image and text (e.g., image captioning, where the text describes what is seen in the image [1, 6, 9, 16, 18, 35, 38, 40]), recognizing *implicit* relationships between image and text (e.g., metaphorical, symbolic, explanatory, ironic, etc.) remains a research challenge.

In this work, we take a first step toward the automatic analysis of non-literal relations between visual and textual persuasion. To make the problem concrete, we focus on the relationship between the visual component of an image advertisement and the slogan embedded

Figure 1: Images and text in persuasive advertisements reinforce each other in complex ways. Even for image-text pairs that are *parallel*, the image and text are complementary rather than redundant (e.g. the first image shows a saw but the text only mentions risks). This is in contrast to image captioning approaches that assume redundant image-text pairs.

in the ad. In particular, we propose a method for determining whether the image and the slogan are in a *parallel* relationship, i.e. whether they convey the same message independently.

Cases A and B in Fig. 1 are two examples of the *parallel* relationship: the image and text in A each warns of commonplace dangers; the image and text in B each warns of unhappiness caused by abuse. Note, however, that in both cases, the image and text do not obviously repeat the same concept: the text in A does not mention any finger or a saw, and the image in B only subtly suggests the aftermath of violence. These examples are challenging because their messages are parallel *but not equivalent*. Cases C and D are examples of the *non-parallel* relationship. In C, the text alone does not establish the topic of smoking, and the image alone does not establish a connection between the shortening of the cigarettes' lengths and the shortening of one's life. The image and text in D even appear to be contradictory, because a fuzzy duckling is "not soft enough."

As the above examples show, elements from the image and text do not always align, suggesting that standard image-text alignment methods alone are insufficient for detecting parallel relationships. We hypothesize that additional cues, such as surprise, ambiguity, specificity and memorability, are needed. To validate our claim, we first collected a dataset of human perceptions of parallelity: we asked annotators on Amazon Mechanical Turk (MTurk) to determine whether the image and slogan convey the same message for a variety of ads. Then, we evaluated different strategies for their ability to identify ad image/slogan in parallel/non-parallel relationships on the collected dataset. We find that the proposed ensemble of cues outperforms standard image-text alignment methods.

The ability to determine whether an image and text are parallel without literally describing the same thing, has many applications. For example, advertising is a common marketing communication strategy that utilizes a visual form to deliver persuasive messages to its target audiences and sell a product, service or idea. A system that can understand the complex relationship between image and text is better equipped to understand the underlying message of the ad. Conversely, this system could also be used to help design more sophisticated ads. Other applications include: flagging poor journalism by detecting unsubstantiated bias (e.g., recognizing that an unflattering image of a celebrity has no semantic association with the corresponding news story other than to bias the audience), generating creative captions (e.g., creating memes that deliver diverse messages vividly by accompanying different text with the same image) and promoting media literacy for educational purposes (e.g., teaching young consumers to recognize persuasive strategies employed by ads).

2 Related Work

Prior work in matching images and text is primarily applied for image captioning or learning visual-semantic spaces. We show in experiments that this is not sufficient for understanding the relationship between persuasive images and persuasive text. A more sophisticated strategy is to determine how logically coherent an image and text pair is. This is similar to modeling *discourse relations* in natural language processing, which aims to identify the logical connection between two utterances in a discourse. Image-text coherence does not yet have the theoretical foundation or resource support that textual coherence has. We also discuss tasks related to judging image memorability and text abstractness, which we show have some relation to image-text parallelity (e.g. if the text is too abstract, it may not be able to convey the full message without help from the image). Finally, we describe prior work on understanding image advertisements and visual persuasion.

Visual-semantic embeddings and image captioning. Work in learning joint image-text spaces [3, 5, 10, 12, 20, 36] aims to project an image and a matching piece of text in such a way that the distance between related text and image is minimized. For example, [20] use triplet loss where an image and its corresponding human-provided caption should be closer in the learned embedding space than image/caption that do not match. [10] propose a bi-directional network to maximize correlation between matching images and text. None of these consider images with implicit or explicit persuasive intent, as we do. Image and video captioning [1, 6, 9, 11, 16, 18, 21, 35, 38, 39, 40] is the task of generating a textual description of an image, or similarly, finding the textual description that best matches the image. These have been successful when depicting observable objects and their spatial relationships, but in our project, the interaction between visual and linguistic information can be hidden from the surface, which is more difficult to detect and requires novel techniques.

Discourse relations. Linguistic units in text are connected logically, forming *discourse relationships* with each other. For example, a later sentence might *elaborate* on an earlier sentence, it might serve as a *contrast*, or it might state the *result* of the earlier one. In this light, our problem of determining the relationship between image and text bears some resemblance to the problem of automatically determining discourse relations between sentences, for which there is extensive prior work [2, 4, 28, 29, 30, 31]. However, unlike the text-only case, we can neither rely on explicit linguistic cues nor on annotated resources such as the Penn Discourse Treebank.

Concreteness and specificity of text. Concrete words are grounded in things that we can sense (e.g., birds, flying, perfume) while abstract words refer to ideas that we do not physically perceive (e.g., cause, "think different," tradition). A related concept is *specificity*, the level of details of the text being communicated (e.g. "robin" is more specific than "bird"). Resources and methods exist to predict concreteness and specificity [7, 22, 37]. These textual stylistic clues may help to assess whether the text and image match in their specificity and concreteness.

Image creativity and memorability. Images in the media are often more creative than typical photographs. Several works examine the aesthetics of photos [26, 34] and the creativity of videos [32]. Some study the memorability of images [15, 19]. In our work, we consider memorability as a cue for the relationship between image and text; memorable images might be more creative (such as image C in Fig. 1) and might indicate a "twist" and *non-parallel* relationship between the image and text.

Visual rhetoric and advertisements. There has *not* been extensive work in visual rhetoric, which is especially common in advertisements. In a first work on visual rhetoric, [17] devised a method to predict which of two photos of a politician portrays that person in a more positive light, detecting implications that the photo conveys beyond the surface. [14, 41] take another step towards understanding visual rhetoric, focusing on image and video advertisements. They formulate ad-decoding as answering a question about the reason why the audience should take the action that the ad suggests, e.g. "I should not be a bully *because words hurt as much as fists do*." However, their prediction method ignores any text in the advertisement images, which is crucial for decoding the rhetoric of ads.

3 Dataset

We use images from the dataset of [14], and annotate them on MTurk for our task. First, we randomly sample 1000 ads images that were predicted to be modern and not too wordy. We request a transcription of the important text (e.g. slogans) in the image, and labels regarding the relationship of the transcribed text and the image. We will make the dataset available for download upon publication.

Image selection. We want to determine how a concise piece of text interacts with the image. In order to focus our task on interesting relationships, we opted to filter images in two ways. First, we noticed that vintage ads often did not seem very creative in their use of language, compared to contemporary ads. Often the text was fairly straightforward and literal, while modern ads make clever use of idioms and wordplay, and aim to be concise and "punchy". We observed that wordy ads were oftentimes also not creative. Thus, we implemented automatic methods to filter vintage or wordy ads. We trained a classifier to distinguish between 2080 vintage ads crawled from the Vintage Ad Browser[1] and 4530 modern ads from Ads of the World[2]. We used the ResNet-50 architecture [13] pre-trained on ImageNet [33] and further trained on these vintage/modern ads, achieving 0.911 precision and 0.950 recall on a held-out validation set. We applied it to images from the dataset of [14], and removed 11582 images that were predicted to be vintage. We also used optical character recognition (OCR) from the Google Cloud Vision API[3] and counted the length of the automatically extracted text fragments in each image. We removed 33503 images containing less than 10 or more than 80 words. In the end, we were left with 19747 images, from which we sampled 1000 for human annotation.

Transcription. Automatic text recognition did not work sufficiently well, so we asked humans to manually transcribe the persuasive text in the ads images. We instructed workers to ignore the following: logos, legalese, date and time, long paragraph of description, urls, hashtags and non-English characters. We emphasized they should transcribe that text which carries the core message of the image.

Relationship annotation. For each image, the key question we want to answer is: Are the text and image *parallel*? We define *parallel* as image and text individually aiming to convey the same message as the other; see the first two examples in Fig. 2. We instruct annotators that: "If you have no difficulty understanding the ad with only the highlighted text, and you have no difficulty understanding it with only the image, the relationship is *parallel*." In contrast, for *non-parallel* pairs, "one of text or image might be fairly unclear

[1] http://www.vintageadbrowser.com
[2] https://www.adsoftheworld.com
[3] https://cloud.google.com/vision/

Text: "Purified drinking water" **Relation:** Parallel **Fine-grained:** Equivalent

Text: "63% of strawberries are contaminated." **Relation:** Parallel **Fine-grained:** Non-equivalent

Text: "A bite's not right without it" **Relation:** Non-parallel **Fine-grained:** One ambiguous

Text: "It's yellow, it's ugly, it doesn't match anything, but it can save lives" **Relation:** Non-parallel **Fine-grained:** One ambiguous

Text: "You've just hired a technology wizard" **Relation:** Non-parallel **Fine-grained:** Opposite

Text: "An apple a day didn't work?" **Relation:** Non-parallel **Fine-grained:** Decorative

Figure 2: Image-text pairs and their relationship (parallel/non-parallel) with rationales. Note: In the second image, Ra is the chemical element Radium.

or ambiguous without the other, text and image alone might imply opposite ideas than when used in combination, or one might be used to attract the viewer's attention while providing no useful information."

Additional questions. We also ask our annotators two additional questions. First, we ask them to describe the message that they perceive in the ad: "Could you figure out the message of this ad? Please give an explicit statement in one or two sentences." We found that asking annotators to type the message of the ad helped them to think more thoroughly about the ad image, which resulted in higher-quality results. We are not interested in this annotation beyond quality control; further, [14]'s dataset already provides this info.

Fine-grained relationships. Second, we asked annotators to provide some rationale for their answer in the main parallel/non-parallel question. Annotators could choose from the following five rationale options (or type their own):

1. **Equivalent parallel:** The text and image are completely equivalent: They make exactly the same point, equal in strength (e.g. the first image in Fig. 2).
2. **Non-equivalent parallel:** They try to say the same thing, but one is more detailed than the other: They express same ideas but at different level (e.g. the second image in Fig. 2).
3. **One ambiguous (non-parallel):** Text or image is fairly unclear or ambiguous without the other.
4. **Opposite (non-parallel):** Text and image alone imply opposite ideas than when used together.
5. **Decorative (non-parallel):** The main idea of the ad is conveyed by just the text (or image) alone; the other is primarily decorative.

The first two rationales are meant to accompany a *parallel* response, and the next three a *non-parallel* one. The boldface fine-grained relationship label is internal, and not provided to annotators. We designed this question for improved annotation quality, as we found that in our pilot collection, annotators chose the *parallel* option too often due to carelessness. By making them think about the ad harder, we obtained higher-quality responses.

Quality control. We only allow MTurk workers who have completed over 1000 HITs and have 95%+ approval rate to participate in our study. For each image/text pair, we collected five relationship labels, and performed a majority vote to obtain the final ground-truth. To obtain high-quality labels, we launched the annotation task in batches (each containing 20 image/text pairs), and embedded two quality control questions in each batch. One is a question for which we know the relation between text and image in advance, and the answer is unambiguous. The other is a repeat; we expect the annotators to answer the same way both times if they were paying attention. We removed data from annotators that failed both of these tests, or wrote clearly careless answers to our question about the ad's message.

Ambiguity. One characteristic we observed from collecting such subjective annotations is the lower agreement across workers compared with other objective annotation tasks. For

example, for the parallelity annotation task, nearly 40% of all collected annotations have a 3/2 or 2/3 split between positive and negative choices. After manual inspection we noticed that the difference might come from the fact that image and text might be parallel *on different levels*. Although this interesting trend deserves further exploration, in this study we removed these ambiguous cases and only focused on the instances with high inter-rater agreement.

4 Approach

We develop an ensemble of predictors for the relationship between image and text. Each captures a different aspect of parallelity, and when all are combined, they predict the relationship much more reliably than prior methods for image captioning. Note that even though some classifiers use information only from a single channel (text or image), they still demonstrate competitive predictive power for modeling the relationship. One possible explanation is they capture how ads are created by designers. For example, if wordplay is used in the slogan it is usually a strong indicator of ambiguity and non-parallelity.

4.1 Base features

We develop the following features to recognize parallel relationships, then train classifiers using them. These capture a variety of aspects, including alignment and distances between channels, as well as stylistic and rhetorical nuances of individual channels.

1: Visual Semantic Embedding (VSE). Deep neural networks trained on massive datasets have demonstrated their potential to capture semantic meaning in many tasks. To modify existing visual-semantic embeddings for our task, we apply the pre-trained VSE model from [20] on our dataset, converting ad images and the accompanying slogans into fixed-length embeddings which are concatenated as a feature set. VSE is expected to be good at detecting literal parallel cases (i.e. *equivalent parallel*).

2: Concepts alignment. While VSE offers a means to compare the image and text in their entirety, it may also be useful to determine whether a certain concept is explicitly mentioned in both the image and the text. As a heuristic, concept alignment may give us a high-precision classifier for identifying parallel cases (and as a corollary, a high-recall classifier for identifying non-parallel cases). We use the general-purpose recognition API provided by Clarifai[4] to extract dominating concepts from ad images. Typically each image will be tagged with 20 relevant concepts together with probabilities, and the concepts cover a wide range of objects, activities and scenes. For text, we preprocess the transcribed slogans using the NLTK package [23] to perform automatic part-of-speech tagging. After removing stop-words, we empirically select `NOUN`, `ADJ`, `ADV` and `VERB` to be key concepts in text. We then compute pairwise semantic distances between the collected image concepts and text concepts. We use a pre-trained word2vec [24] model as the bridge to convert concepts into a semantically meaningful vector representation, then measure their cosine distances.

3: Agreement from single channel prediction. If the viewer can deduce the correct message from either channel alone, then using either text or image would be sufficient for advertising, and the relationship might be *parallel*, but if the meaning is unclear with one channel disabled, then both channels are indispensable, and the relationship might be *non-parallel*. To capture this intuition, the most fitting task is to evaluate whether the intended message can be perceived with information from a single channel, but this task is very challenging as shown in [14]. Here we choose the prediction of the ad's topic as a proxy task.

[4] https://clarifai.com/developer/guide/

For images, we infer the ad topic using a fine-tuned ResNet [13] model, and take the predicted probability distribution as the information provided from the image channel. For text, since we have only a limited number of transcribed slogans, we use the Google Cloud Vision API for OCR and extract text for the full ads dataset from [14]. After removing non-English words and stop-words, we group the words in the extracted text according to their ad topic annotation and perform TF-IDF analysis. The resultant adjusted counts represent the relative frequency for individual words to appear in various topics. For inference, we let individual words vote for the topics using weights proportional to the TF-IDF scores. We apply softmax to convert the votes into a probabilistic distribution over topics, and the two predicted topic distributions are then concatenated as the feature.

4: Surprise factor. In addition to the contrast across image and text channels, empirical evaluations indicate that the relationship between concepts within a single channel is informative as well. For example, using words that contradict each other in the slogan is sufficient to make an impression ("less is more") by itself, and an image with objects that do not normally co-occur (like cats in a car ad) catches the viewer's attention. We calculate the pairwise distances between concepts extracted within the image and text channels, respectively, and use the normalized histogram of the distribution of cosine distances as a feature vector.

5: Image memorability. Studies have shown that memorable and forgettable images have different intrinsic visual features. For ads, memorability is an especially interesting property as one of the goals for ads is to be remembered. Therefore a reasonable guess is that visual memorability can affect how designers arrange elements within ads, including both image and text. Here we adopt the model in [19] and perform inference on our collected ads, and the 1D resultant memorability score (higher means more memorable) is used as a feature.

6: Low-level vision features. Low-level vision features like HoG [8], which capture gradient orientation statistics, have been widely used for image recognition and object detection. To obtain a fixed-length representation for ads with varied sizes, we resize all the images to be 128×128 and use 9 orientations, 16×16 pixels per cell and 2×2 cells per block.

7: Slogan specificity and concreteness. Sentence specificity and word concreteness have been recognized as important characterizations of texts [27] for language tasks such as document summarization and sentiment analysis. Here we adopt the model from [22] to predict a specificity score for the slogans we collected from manual transcription. We transformed the transcriptions into four different forms (original, all uppercase, all lowercase and only first character of sentence is capitalized), and the specificity scores for all these four forms are concatenated. Similarly, we extract the concreteness score for individual words within slogans using the MRC dataset [7], and calculate the max, median, mean and stdev.

8: Lexical ambiguity and polysemy. Polysemy is a prevailing phenomenon in languages, and this lexical ambiguity has been widely used in advertising because the varying interpretation makes the viewer more involved in understanding the ads, and subsequently makes the ad more impressive. We use WordNet [25] to query each word within the slogan, and collect statistics on the number of plausible meanings to estimate the ambiguity in the slogan.

9: Topics. We use a one-hot vector for the ground-truth topic annotation of ads [14], following an intuitive clue that ads belonging to the same theme could use similar strategies for image and text composition.

	Counts	Parallel Score	Equivalent Score
Non-parallel	169	0.291 (0.141)	0.143 (0.148)
Parallel Equivalent	84	0.924 (0.097)	0.767 (0.114)
Parallel Non-equivalent	84	0.857 (0.090)	0.210 (0.127)

Table 1: Statistics on the dataset constructed for our experiments. The parallel score mean (stdev) come from the main annotation, and the equivalent score from the fine-grained rationale. We analyze performance on the non-equivalent examples as they are more challenging.

4.2 Combining clues from different classifiers

We attempted several ways to combine our cues from the previous section. One can concatenate all previously described features, perform a majority vote over the individual classifier responses, or learn a weighted voting scheme via bagging and boosting. While several of these gave promising results, the best approach was the following. Rather than constructing one classifier towards the overall goal, we first focus on how each individual classifier performs, and then use auxiliary features to help choose reliable classifiers under different scenarios. Specifically, since our base classifiers capture different perspectives of parallelity, they demonstrate varied performance on different samples. Intuitively, if an auxiliary classifier can capture contextual information to judge which classifier to trust for a specific test instance, performance should improve compared to weighing all classifiers equally.

Consider one sample for illustration so we can omit the sample subscript. We construct m sets of features (corresponding to the subsections in Sec. 4.1); let X_i be the i-th set. We first train m SVM classifiers $f_i(X_i)$, one per feature set. Then we take the output $\hat{y}_i = f_i(X_i)$ and its probability estimate $\hat{p}_i$. We further construct a new set of labels, $o_i = \mathbb{1}_{\{y=\hat{y}_i\}}$, and train $g_i(\hat{p}_1,...,\hat{p}_m)$ for each o_i, to obtain prediction $\hat{o}_i = g_i(\hat{p}_1,...,\hat{p}_m)$, measuring whether each base prediction is reliable for certain contexts. Finally, we take the k largest $\hat{o}_i$, and use the corresponding $\hat{y}_i$ to get a majority vote (we choose $k = 5$ in our experiments).

5 Experimental Validation

In this section, we verify that our proposed features can help distinguish between *parallel* and *non-parallel* ads better than standard methods (e.g. VSE) can, especially for the harder cases (such as distinguishing between *non-equivalent parallel* ads and *non-parallel* ads).

5.1 Experimental setup

For each image-text pair in our collected dataset, we calculate the *parallel score* according to the responses from all annotators by the percentage of *parallel* responses (e.g. 0.6 if 3 out of 5 annotators choose *parallel*). To validate our proposed strategy, we remove some ambiguous cases and construct a balanced dataset. We choose those image/text pairs for which more than half of annotators agree to be *non-parallel* as the negative samples, and sample roughly the same number of positive samples from ads that have *parallel* score higher than 0.8 (over four-fifths agreement on *parallel*). Since *parallel* ads are varied, we use the fine-grained rationales (Sec. 3) to construct a balanced set within *parallel*. We make sure that *equivalent* and *non-equivalent* ads (corresponding to the first and second rationales) are balanced within the *parallel* category. In Table 1 we report statistics about the constructed dataset, which clearly show the distinct rating scores between *parallel* and *non-parallel* ads, as well as *equivalent* and *non-equivalent* ads. Note that our main task is to predict *parallel* while the equivalent/non-equivalent task is only used for further analysis.

Feature Set	Dim	Avg accuracy	Per-class avg accuracy		
			Non-parallel	Parallel	
				Equivalent	Non-equiv
1: VSE	2048	0.602	0.574	*0.667*	0.595
2: Alignment	85	0.581	0.598	*0.595*	0.536
3: Topic Agreement	80	0.516	0.728	0.298	*0.310*
4a: Image Surprise	65	0.549	0.550	*0.595*	0.500
4b: Text Surprise	70	0.596	0.598	0.548	*0.643*
5: Memorability	1	0.567	0.408	*0.774*	0.679
6: HoG	1764	0.582	0.728	0.405	*0.464*
7a: Specificity	4	0.534	0.491	0.524	*0.631*
7b: Concreteness	4	0.525	0.361	0.679	*0.702*
8: Polysemy	4	0.504	0.308	0.643	*0.762*
9: Topic	40	0.539	0.402	*0.702*	0.655
Combination	11	**0.655**	0.633	0.702	0.655

Table 2: Experimental results using individual classifiers as well as our combined classifier, on predicting *parallel* vs *non-parallel*. Accuracy is whether the prediction matches the ground truth, and we measure it over the entire dataset and over three fine-grained subcategories. The method performing best overall is in **bold**. For the last two columns, we *italicize* whether an individual feature does better on the harder case of correctly classifying non-equivalent parallel ads, or the easier case of equivalent parallel. Our combined method outperforms the VSE baseline, and our features generally do better on the harder case.

5.2 Results

We show our quantitative results in Table 2, including performances of individual classifiers from Sec. 4.1 and the combination of features (Sec. 4.2). For fair comparison we use SVMs with linear kernels for all feature sets, and report performance measured from five-fold cross validation. We observe the five individual methods that are most competitive in terms of overall average (third column) are VSE, Text Surprise, HoG, Alignment and Memorability. However, the combined classifier performs best, and outperforms VSE by 9%.

We are particularly interested in the type of samples that each method classifies correctly. The distinction between *parallel equivalent* (such as the first image in Fig. 2) and *non-parallel* is much more obvious, than the distinction between *parallel non-equivalent* (such as the second image in Fig. 2) and *non-parallel*. Thus, we hope our methods which were specifically designed to handle less obvious cases can do well on the harder cases. This is indeed true for most of our methods, especially Text Surprise, HoG, Specificity, Concreteness and Polysemy. Topic Agreement does not help much for overall accuracy but it performs best for capturing *non-parallel* image-text pairs, tied with HoG. Memorability is the best across all methods for distinguishing between *parallel equivalent* and *non-parallel* ads.

Finally, we give some qualitative examples in Fig. 3, comparing our model to the second-best method, VSE. One example in which our model outperforms VSE is the electric car ad. Here, VSE mistakenly predicts *parallel*, but the difference between “Mom”, “Dad” and “electricity” is captured by Text Surprise. Additionally, HoG and Topic Agreement also predict "non-parallel." While some individual classifiers choose *parallel*, our combination model has correctly relied on Text Surprise, HoG, and Topic Agreement to make the right

Figure 3: Qualitative examples from our combined model and VSE. (Top) False positive: Non-parallel cases classified as parallel by VSE or our model. (Bottom) False negative: Parallel cases classified as non-parallel. (Left): Our combined model correctly captures the image/text relationship but VSE fails. (Right) Both our combined model and VSE fail.

prediction. Another success is the "Picture perfect!" example. The detected image concepts are all semantically close (child, family, people, etc.) and such trends are captured by Image Surprise for *parallel*. Our combination model picks it as a trustworthy classifier, along with Alignment and Topic, and makes the correct prediction.

Our model makes mistakes, especially for semantically subtle ads. In the "Reveal deep clean and brighter skin" example, the text highlights the *effect* after using the makeup, while image illustrates the makeup itself. Human raters have no problem recognizing the differences, while VSE, Topic Agreement (both image and text are indicative of makeup), Image Surprise (common and natural image, little surprise factor) and Concreteness (very concrete expression) all predict *parallel*. Although Alignment predicts *non-parallel* (image and text do not align well), our combination model incorrectly goes with the majority prediction.

6 Conclusion

We proposed the novel problem of analyzing non-literal relationships between persuasive images and text. We developed features that aim to approximate some of the persuasive strategies used by ad designers. We tested these features on the task of predicting whether an image and text are *parallel* and convey the same message as each other without saying the exact same thing, for which we crowdsourced high-quality annotations. We showed that our combined classifier, which uses our proposed features, significantly outperforms a baseline method developed for literal image-text matching, as in image captioning.

Acknowledgements. We would like to thank Daniel Zheng for helping in analyzing the sentence concreteness, and Keren Ye and Xiaozhong Zhang for valuable discussions throughout the project. We would also like to thank the anonymous reviewers for their helpful comments. This material is based upon work supported by the National Science Foundation under Grant Number 1718262 and an NVIDIA hardware grant. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

References

[1] Lisa Anne Hendricks, Subhashini Venugopalan, Marcus Rohrbach, Raymond Mooney, Kate Saenko, and Trevor Darrell. Deep compositional captioning: Describing novel object categories without paired training data. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[2] Or Biran and Kathleen McKeown. Aggregated word pair features for implicit discourse relation disambiguation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 69–73, Sofia, Bulgaria, August 2013.

[3] Yue Cao, Mingsheng Long, Jianmin Wang, and Shichen Liu. Deep visual-semantic quantization for efficient image retrieval. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[4] Jifan Chen, Qi Zhang, Pengfei Liu, Xipeng Qiu, and Xuanjing Huang. Implicit discourse relation detection via a deep architecture with gated relevance network. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1726–1735, Berlin, Germany, August 2016.

[5] Kan Chen, Trung Bui, Chen Fang, Zhaowen Wang, and Ram Nevatia. Amc: Attention guided multi-modal correlation learning for image search. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[6] Tseng-Hung Chen, Yuan-Hong Liao, Ching-Yao Chuang, Wan-Ting Hsu, Jianlong Fu, and Min Sun. Show, adapt and tell: Adversarial training of cross-domain image captioner. In *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[7] Max Coltheart. The mrc psycholinguistic database. *The Quarterly Journal of Experimental Psychology*, 33(4):497–505, 1981.

[8] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2005.

[9] Jeffrey Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, and Trevor Darrell. Long-term recurrent convolutional networks for visual recognition and description. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

[10] Aviv Eisenschtat and Lior Wolf. Linking image and text with 2-way nets. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[11] Chuang Gan, Zhe Gan, Xiaodong He, Jianfeng Gao, and Li Deng. Stylenet: Generating attractive visual captions with styles. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[12] Lluis Gomez, Yash Patel, Marcal Rusinol, Dimosthenis Karatzas, and C. V. Jawahar. Self-supervised learning of visual features through embedding images into text topic spaces. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[14] Zaeem Hussain, Mingda Zhang, Xiaozhong Zhang, Keren Ye, Christopher Thomas, Zuha Agha, Nathan Ong, and Adriana Kovashka. Automatic understanding of image and video advertisements. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[15] Phillip Isola, Jianxiong Xiao, Devi Parikh, Antonio Torralba, and Aude Oliva. What makes a photograph memorable? *IEEE transactions on pattern analysis and machine intelligence*, 36(7):1469–1482, 2014.

[16] Justin Johnson, Andrej Karpathy, and Li Fei-Fei. Densecap: Fully convolutional localization networks for dense captioning. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[17] Jungseock Joo, Weixin Li, Francis F Steen, and Song-Chun Zhu. Visual persuasion: Inferring communicative intents of images. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.

[18] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

[19] Aditya Khosla, Akhil S Raju, Antonio Torralba, and Aude Oliva. Understanding and predicting image memorability at a large scale. In *The IEEE International Conference on Computer Vision (ICCV)*, 2015.

[20] Ryan Kiros, Ruslan Salakhutdinov, and Richard S Zemel. Unifying visual-semantic embeddings with multimodal neural language models. In *TACL*, 2015.

[21] Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. Dense-captioning events in videos. In *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[22] Junyi Jessy Li and Ani Nenkova. Fast and accurate prediction of sentence specificity. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence (AAAI)*, 2015.

[23] Edward Loper and Steven Bird. Nltk: The natural language toolkit. In *Proceedings of the ACL-02 Workshop on Effective tools and methodologies for teaching natural language processing and computational linguistics-Volume 1*, pages 63–70, 2002.

[24] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems (NIPS)*, pages 3111–3119, 2013.

[25] George A Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41, 1995.

[26] Naila Murray, Luca Marchesotti, and Florent Perronnin. Ava: A large-scale database for aesthetic visual analysis. In *CVPR*, pages 2408–2415. IEEE, 2012.

[27] A. Paivio. *Imagery and Verbal Processes*. Holt, Rinehart and Winston, 1971.

[28] Joonsuk Park and Claire Cardie. Improving implicit discourse relation recognition through feature set optimization. In *Proceedings of the 13th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, SIGDIAL '12, pages 108–112, 2012.

[29] Emily Pitler, Annie Louis, and Ani Nenkova. Automatic sense prediction for implicit discourse relations in text. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2 - Volume 2*, ACL '09, pages 683–691, 2009.

[30] Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind Joshi, and Bonnie Webber. The penn discourse treebank 2.0. In *Proceedings of LREC*, 2008.

[31] Lianhui Qin, Zhisong Zhang, and Hai Zhao. A stacking gated neural architecture for implicit discourse relation classification. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2263–2270, Austin, Texas, November 2016.

[32] Miriam Redi, Neil O'Hare, Rossano Schifanella, Michele Trevisiol, and Alejandro Jaimes. 6 seconds of sound and vision: Creativity in micro-videos. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4272–4279, 2014.

[33] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115 (3):211–252, 2015.

[34] Rossano Schifanella, Miriam Redi, and Luca Maria Aiello. An image is worth more than a thousand favorites: Surfacing the hidden beauty of flickr pictures. In *ICWSM*, pages 397–406, 2015.

[35] Rakshith Shetty, Marcus Rohrbach, Lisa Anne Hendricks, Mario Fritz, and Bernt Schiele. Speaking the same language: Matching machine to human captions by adversarial training. In *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[36] Jifei Song, Yi-Zhe Song, Tao Xiang, and Timothy Hospedales. Fine-grained image retrieval: the text/sketch input dilemma. 2017.

[37] Peter Turney, Yair Neuman, Dan Assaf, and Yohai Cohen. Literal and metaphorical sense identification through concrete and abstract context. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 680–690, Edinburgh, Scotland, UK., July 2011.

[38] Ramakrishna Vedantam, Samy Bengio, Kevin Murphy, Devi Parikh, and Gal Chechik. Context-aware captions from context-agnostic supervision. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[39] Subhashini Venugopalan, Lisa Anne Hendricks, Marcus Rohrbach, Raymond Mooney, Trevor Darrell, and Kate Saenko. Captioning images with diverse objects. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[40] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3156–3164, 2015.

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