Clue: Cross-modal Coherence Modeling for Caption Generation

Malihe Alikhani
Rutgers University
malihe.alikhani@rutgers.edu

Piyush Sharma
Google Research
piyushsharma@google.com

Shengjie Li
Rutgers University
sl1560@rutgers.edu

Radu Soricut
Google Research
rsoricut@google.com

Matthew Stone
Rutgers University
mdstone@rutgers.edu

Abstract

We use coherence relations inspired by computational models of discourse to study the information needs and goals of image captioning. Using an annotation protocol specifically devised for capturing image–caption coherence relations, we annotate 10,000 instances from publicly-available image–caption pairs. We introduce a new task for learning inferences in imagery and text, coherence relation prediction, and show that these coherence annotations can be exploited to learn relation classifiers as an intermediary step, and also train coherence-aware, controllable image captioning models. The results show a dramatic improvement in the consistency and quality of the generated captions with respect to information needs specified via coherence relations.

1 Introduction

The task of image captioning is seemingly straightforward to define: use natural language to generate a description that captures the salient content of an image. Initial datasets, such as MSCOCO (Lin et al., 2014) and Flickr (Young et al., 2014), approached this task directly, by asking crowd workers to describe images in text. Unfortunately, such dedicated annotation efforts cannot yield enough data for training robust generation models; the resulting generated captions are plagued by content hallucinations (Rohrbach et al., 2018; Sharma et al., 2018) that effectively preclude them for being used in real-world applications.

In introducing the Conceptual Captions dataset, Sharma et al. (2018) show that this dataset is large enough, at 3.3M examples, to significantly alleviate content hallucination. However, because the technique for creating such a large-scale resource relies on harvesting existing data from the web, it no longer guarantees consistent image–text relations. For example, along with descriptive captions (e.g.,"this is a person in a suit"), the dataset also includes texts that provide contextual background (e.g., "this is the new general manger of the team") and subjective evaluations (e.g., "this is stylish"). As a result, current captioning models trained on Conceptual Captions avoid content hallucination but also introduce different, more subtle and harder-to-detect issues related to possible context hallucinations (i.e., is this actually the new general manager?) or subjective-judgement hallucinations (i.e., whose judgment is this anyway?).

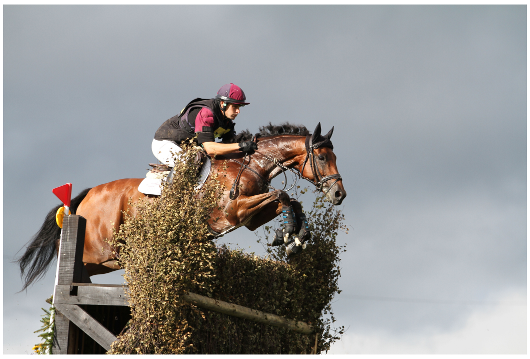


Figure 1: Output of a coherence-aware model for various coherence relations. Content that establishes the intended relation is underlined. (Photo credit: Blue Destiny / Alamy Stock Photo)
Visible: horse and rider jumping a fence.
Meta: horse and rider jumping a fence during a race.
Subjective: the most beautiful horse in the world.
Story: horse competes in the event.

In this paper, we propose to tackle this issue of large-scale image-caption consistency using a coherence-aware approach inspired by the framework of discourse coherence theory (Hobbs, 1978; Phillips, 1977). This framework characterizes the inferences that give discourse units a coherent joint interpretation using a constrained inventory of coherence relations. In multimodal presentations, discourse units can be images as well as text, so we appeal to new image–text coherence relations that

Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 6525–6535
July 5 - 10, 2020. ©2020 Association for Computational Linguistics

capture the structural, logical, and purposeful relationships between the contributions of the visual modality and the contributions of the textual modality. For instance, a *Visible* relation characterizes grounding texts that serve to make key aspects of the image content common ground (perhaps to a visually-impaired reader), analogous to *Restatement* relations between one text unit and another; *Visible* relations are key to traditional descriptive captions such as "this is a person in a suit." Meanwhile, a *Story* relation characterizes texts that develop the circumstances depicted in the image in pursuit of free-standing communicative goals, analogous to *Occasion* or *Narration* relations in text; *Story* relations can go far beyond image content ("I hiked this mountain as we found it on a list for good hikes for kids") and so pinpoint one kind of risk for context hallucinations. The key contribution of our work is to show that image–text coherence can be systematized, recognized, and used to control image captioning models.

To support our argument, we create a coherence-relation annotation protocol for image-caption pairs, which we use to annotate 10,000 image-caption pairs over images coming from the Conceptual Captions (Sharma et al., 2018) and Open Images (Kuznetsova et al., 2020) datasets. We release[1] this dataset, named Clue, to facilitate follow-up research. By annotating these coherence relations in the context of image captioning, we open up the possibility of analyzing patterns of information in image–text presentations at web scale.

In addition, we show that we can exploit these coherence-relation annotations by training models to automatically induce them, as well as by building models for coherence-aware image captioning. Because they are driven by input coherence relations, these captioning models can be used to generate captions that are better suited to meet specific information needs and goals.

2 Prior Work

There are diverse ways to characterize the communicative functions of text and images in multimodal documents (Marsh and Domas White, 2003), any of which can provide the basis for computational work. Some studies emphasize the distinctive cognitive effects of imagery in directing attention; engaging perceptual, spatial and embodied reasoning; or eliciting emotion (Kruk et al., 2019; Shuster et al., 2019). Some look at contrasts across style and genre (Guo et al., 2019). Others look holistically at the content of text and imagery as complementary or redundant (Otto et al., 2019; Vempala and Preotiuc-Pietro, 2019). Unlike our approach, none of these methodologies attempt to characterize information-level inferences between images and text, so none is suitable for building generation models that control the information that text provides.

While coherence theory has been applied to a range of multimodal communication, including comics (McCloud, 1993), gesture (Lascarides and Stone, 2009), film (Cumming et al., 2017), and demonstrations and other real-world events (Hunter et al., 2018; Stojnic et al., 2013), applying coherence theory specifically to text–image presentations is less well explored. The closest work to ours is Alikhani et al. (2019), who explore coherence relations between images and text in a multimodal recipe dataset. Their relations are specialized to instructional discourse and they do not build machine learning models combining imagery and text. We consider more general coherence relations and a broader range of machine learning methods.

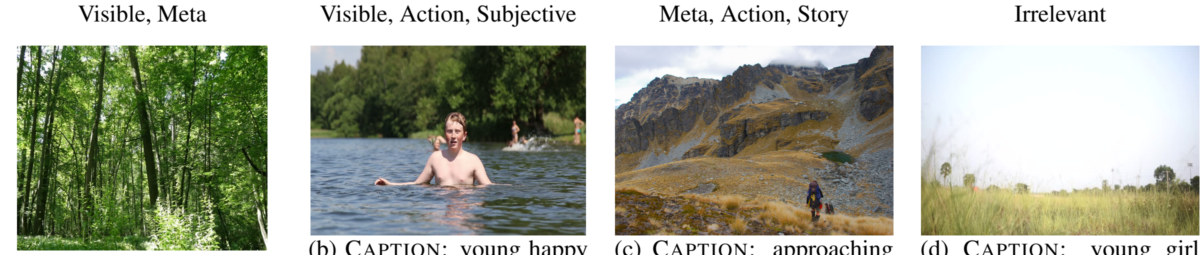
We use our relations and introduce a coherence-aware caption generation model that improves the rate of good *Visible* captions by around 30%. This is a considerable improvement over the recent models that have tried to achieve more control over neural language generation using an enhanced beam search (Anderson et al., 2017), a memory network with multiple context information (Chunseong Park et al., 2017), forced attentions (Sadler et al., 2019) and modeling and learning compositional semantics using fine-grained annotations of entities in MSCOCO (Cornia et al., 2019).

3 Coherence in Images and Captions

The first step toward our goals is to characterize image–text coherence and annotate a sizable corpus of image–text pairs with coherence relations.

We use an overlapping set of high-level relations, inspired both by theoretical work linking discourse coherence to discourse structure and discourse goals (Roberts, 2012; Webber et al., 1999), and by previous successful discourse annotation campaigns (Prasad et al., 2008). Crucially, following previous work on text (Rohde et al., 2018) and multimodal discourse (Alikhani et al., 2019),

[1] https://github.com/malihealikhani/Cross-modal_Coherence_Modeling



(a) CAPTION: forest on a sunny day

(b) CAPTION: young happy boy swimming in the lake.

(c) CAPTION: approaching our campsite, at 1550m of elevation on the slopes.

(d) CAPTION: young girl walking on the dry grass field under daylight.

Figure 2: We use a constrained set of coherence relations to summarize the structural, logical and purposeful relationships between the contributions of text and the contributions of images. Multiple coherence relations can be found simultaneously. (Image–caption pairs are chosen from the Conceptual Caption dataset; photo credits: Dmytro Zinkevych; Shutterstock user yauhenka; Danilo Hegg; Andre Seale)

we assume that several of these relations can hold concurrently. The relations are:

- *Visible*, where text presents information that is intended to recognizably characterize what is depicted in the image, analogous to *Restatement* relations in text (Prasad et al., 2008).
- *Subjective*, where the text describes the speaker's reaction to, or evaluation of, what is depicted in the image, analogous to *Evaluation* relations in text (Hobbs, 1985);
- *Action*, where the text describes an extended, dynamic process of which the moment captured in the image is a representative snapshot, analogous to *Elaboration* relations in text (Prasad et al., 2008);
- *Story*, where the text is understood as providing a free-standing description of the circumstances depicted in the image, analogous to the *Occasion* relation of Hobbs (1985) but including instructional, explanatory and other background relations; and
- *Meta*, where the text allows the reader to draw inferences not just about the scene depicted in the image but about the production and presentation of the image itself, analogous to *Meta-talk* relations in text (Schiffrin, 1980).

Figures 2(a), (b) and (c) show examples of image–caption pairs and the associated coherence relations. We can see that image–caption pairs often have multiple relations. For completeness, we also present in Figure 2(d) an example of an image–caption pair that does not fall into any of the above categories (and it is therefore labeled *Irrelevant*).

3.1 Data Collection

Clue includes a total of 10,000 annotated image–caption pairs. A first subset of 5,000 image–caption pairs was randomly selected from the training split of the Conceptual Captions dataset (Sharma et al., 2018), as a representative sample of human-authored image captions. The Conceptual Captions dataset is a collection of web-harvested images paired with their associated ALT-TEXT, created by human authors under various non-public guidelines (regarding style, objectivity, etc.) for over 111,000 web pages including news articles, advertisements, educational posts, blogs, etc.

A second subset of 5,000 image–caption pairs, to be used as a representative sample of machine-authored captions, is obtained from the outputs of 5 of the top models that participated in the image-captioning challenge for the Conceptual Caption Workshop at the 2019 Conference on Computer Vision and Pattern Recognition (CVPR). These machine-authored captions are over a set of 1,000 images from the Open Images Dataset (Kuznetsova et al., 2020), and are publicly available.[2]

Protocol Although specific inferences have been shown to be realizable by crowd workers (Alikhani et al., 2019), the results of our pilot studies for annotating these more general relations with the help of crowd workers were not satisfactory. We have found that expert raters' decisions, however, have high agreement on our discourse categories. The study has been approved by Rutgers's IRB; the annotators, two undergraduate linguistics students, were paid a rate of $20/h.

In our annotation protocol, we ask the annotators to label the main relations described in Section 3, as well as certain fine-grained sub-relations. The following briefly summarizes our guidelines; our GitHub repository includes an exact copy of what

[2] http://www.conceptualcaptions.com/winners-and-data

the annotators used.

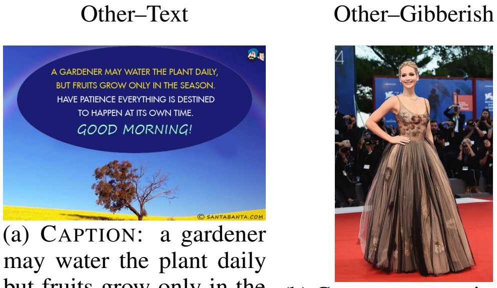
Annotations of *Visible* are given for captions that present information intended to recognizably characterize what is depicted in the image, while annotations of *Meta* indicate not only information about the scene depicted but also about the production and presentation of the image itself. The *Meta* labels have additional fine-grained labels such as *When*, *How*, and *Where*. A few details regarding these fine-grained labels are worth mentioning: location mentions such as "in the city" are labeled as *Meta-Where*, but generic states, e.g., "in the snow," are merely annotated as *Visible*. Captions considering the view or the photo angles, or a photos composition, i.e. "portrait" or "close-up", are annotated as *Meta-How*.

Annotations of *Subjective* are primarily given for captions that included phrases with no objective truth value, i.e. phrases using predicates of personal taste. For example, captions including noun phrases like "pretty garden" are annotated as *Subjective*: whether the garden is pretty or cannot be determined except by appeal to the opinions of an implicit judge. Note that first-person reports, like "I want ..." or "I need ..." are not annotated as *Subjective* but rather as *Story*, because they describe the speaker's definite state rather than an implicit judgment.

Captions annotated as *Story* cover a much wider range compared to captions in other categories, including *Meta* and *Subjective*. These captions range from those that read like instructions, i.e. "how to ...", to those that present speaker desires, i.e. "I want ..." or "I need ...", to those that give background information not captured in the image, i.e. "she is an actress and model", and more.

Other* and *Irrelevant Some of these image–caption pairs contain incomplete captions that are hard to understand. A number of these examples include images that contained text. The text in these cases is relevant to the image and the accompanying captions; in this cases, the coherence relations are marked as *Other–Text* (Figure 3). Some examples of such instances are images containing signs with text, greetings on cards, or text that does not affect the interpretation of the image or caption, such as city names or watermarks.

Other times, the caption text is irrelevant and indicate that the image and caption do not correlate. Some examples of these instances are captions of "digital art selected for" paired with an irrelevant image, and images that clearly do not match the caption, such as an image of a man walking with the caption "a field of strawberries". We have specifically labeled cases where the caption is almost true or almost relevant to the image at hand, such as the caption "horses in a field" with an image containing donkeys with "minor error". Other cases include images that look like powerpoint slides with bullets and text. Our GitHub repository includes detailed examples and explanations.


Other–Text
Other–Gibberish
A GARDENER MAY WATER THE PLANT DAILY, BUT FRUITS GROW ONLY IN THE SEASON. HAVE PATIENCE EVERYTHING IS DESTINED TO HAPPEN AT ITS OWN TIME.
GOOD MORNING!
© SANTABANTA.COM

(a) CAPTION: a gardener may water the plant daily but fruits grow only in the season.

(b) CAPTION: actor in retail at the mother.

Figure 3: Examples of image–caption pairs in the *Other* category. (Photo credit: santabanta.com; Mary Sollosi)

Experiment Interface We have developed software for annotating coherence relations in image–text presentations that can flexibly and easily accommodate various annotation schema. The annotators used this software for annotating the image–text pairs. They had the option of choosing multiple items and leaving comments.

Agreement To assess the inter-rater agreement, we determine Cohens κ. For this, we randomly chose 300 image–caption pairs from the Conceptual Caption ground-truth data and assigned them to two annotators. The resulting κ coefficient is 0.81, which indicates a high agreement on these categorical decisions.

3.2 Analysis

In this section we present the overall statistics of the dataset annotations, the limitations of the caption-generation models, and the correlation of the distribution of the coherence relations with genre.

Overall statistics The exact statistics over the resulting annotations are presented in Table 1 and Table 2. Overall, *Visible* captions constitute around 65% and 70% of captions for the ground-truth labels and the model outputs, respectively. The rate

	Visible	Subjective	Action	Story	Meta	Irrelevant
Ground-truth	64.97%	9.77%	18.77%	29.84%	24.59%	3.09%
Model output	69.72%	1.99%	11.22%	17.19%	58.94%	16.97%
Ground-truth + Model	66.91%	6.58%	15.68%	24.67%	38.65%	8.77%

Table 1: Distribution of coherence relations over the ground-truth and the model outputs.

	When	How	Where
Ground-truth	33.74%	64.40%	28.60%
Model output	21.75 %	72.84%	41.03%

Table 2: Distribution of fine-grain relations in the Meta category over the ground-truth and the model outputs.

of *Subjective* and *Story* captions decreases significantly for the model outputs (compared to ground-truth), indicating that the models learn to favor the *Visible* relation at the expense of *Subjective* and *Story*. However, the rate of *Meta* captions increases by around 25% in the model outputs, which points to potential context hallucination effects introduced by these models. As expected, the rate of *Irrelevant* captions increases to around 17% in the model-generated captions, compared to 3% in the ground-truth captions. Moreover, it appears that the models have some ability to learn to generate the locations that events take place; however, there is a drop in their ability to generate temporal information (see Table 2).

In terms of overlap, *Visible* and *Meta* overlap 22.49% of the time for the ground-truth captions, whereas this rate goes up to 54.55% in the model outputs. This "conflation" of these two relations is highly problematic, and one of the main motivations for building caption-generation models that have control over the type of discourse relation they create (see Section 5). Our GitHub page includes additional statistics about overlapping relations.

Coherence relations indicate Genre Coherence relations are indicative of the discourse type and its goals, and therefore our annotations correlate with the genre under which the captions have been produced. That is, image–caption pairs from different publication sources have different distributions of coherence relations. For instance, pairs from the Getty Images domain mostly come with the *Meta* and *Visible* relations. In contrast, from the Daily Mail domain are mostly story-like, and include very few captions that describe an action, compared with the Getty Images and picdn domains. Figure 4 shows the distribution of the coherence labels for the top four domains from the Conceptual Caption dataset.

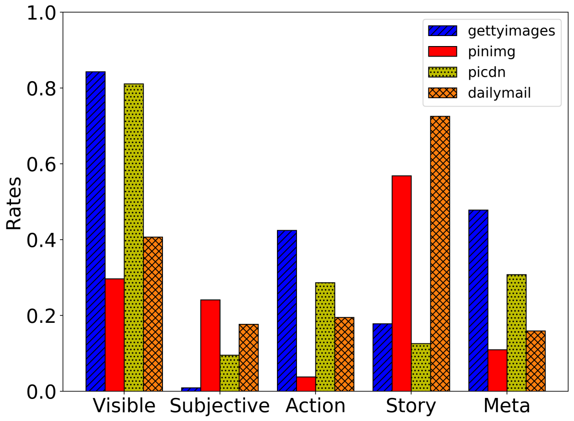


Figure 4: Different resources have different kinds image–caption pairs. The graph shows the distribution of labels in the top four domains present in the Conceptual Captions dataset.

4 Predicting Coherence Relations

In this section, we introduce the task of predicting cross-modal coherence relations. We describe a number of preliminary experiments that justify the potential of machine learning models in classifying coherence relations in text and imagery. To this end, we train and test different models on the Clue dataset to automatically predict the coherence labels given an image and its caption.

4.1 Multi-Label Prediction

We first treat the relation prediction problem in its original multi-label setting. The train–test split for all the models described in this section is 80%–20% and the numbers are reported using 5-fold cross validation.

As a baseline, we report the results of a SVM classifier that uses only the text to predict the relationship between image-caption pairs. We extract bag-of-words features by using N-grams (for N from 1 to 5), and pass them to the SVM classifier

	Visible	Subjective	Action	Story	Meta	Irrelevant	Weighted
SVM (text-only)	0.83	0.12	0.32	0.21	0.19	0.00	0.48
GloVe (text-only)	0.80	0.44	0.58	0.57	0.44	0.08	0.63
BERT (text-only)	0.82	0.35	0.62	0.62	0.44	0.06	0.65
GloVe + ResNet	0.81	0.36	0.58	0.60	0.45	0.07	0.64
BERT + ResNet	0.83	0.36	0.69	0.62	0.44	0.06	0.67

Table 3: The F_1 scores of the multi-class classification methods described in Section 4.1; 80-20 train-test split; 5-fold cross validation.

as input. Next, we discuss two multi-modal classifiers for predicting the image–caption coherence relations.

GloVe + ResNet-50 This model contains a text encoder for textual-feature extraction and an image encoder for image-feature extraction. For the image encoder, we use a ResNet-50 (He et al., 2016) pre-trained on ImageNet followed by a BatchNorm layer, a fully connected layer and a ReLU activation function. The text encoder takes as input word embeddings from the GloVe model (Pennington et al., 2014), and consists of an LSTM layer, a Batch-Norm layer, a fully connected layer with tanh activation function.

BERT + ResNet-50 To test the impact of the text encoder in this setup, we reuse the setup of the previous model with a different textual-feature extractor. We train and test using an encoder that takes sentence embeddings as input using the $\langle \text{CLS} \rangle$ representation produced by the BERT-base model (Devlin et al., 2018).

Results The results of all of our models are presented in Table 3, where we present the F_1 scores over each of the individual relations, as well as an overall weighted average. The BERT+ResNet model achieves the highest performance ($|t| > 9.54$, $p < 0.01$), with an overall F_1 score of 0.67. For the interested reader, we present in the GitHub page the top features of the Naive Bayes SVM classifier (Wang and Manning, 2012).

4.2 Single-Label Prediction

To achieve the goal of generating captions with a desired coherence relation to the image, it is important to clearly differentiate between often co-occurring label types (such as *Visible* and *Meta*). To this end, we introduce a label-mapping strategy for predicting coherence relations, such that each image–caption pair is assigned a single coherence label. We map the set of human-annotated coherence relations for an image–caption pair to a single label using the following heuristic:

1. If the set contains the *Meta* label, then the image–caption pair is assigned the *Meta* label.
2. If the set contains the *Visible* label and does not contain either *Meta* or *Subjective*, then the image–caption pair is set to *Visible*.
3. If none of the above rules are met for this image–caption pair, we randomly sample a label from its set of labels.

The distribution of labels after this mapping is given in the first row of Table 4. As opposed to the ground-truth label distribution in Table 1, these values add up to 100%.

Using the label mapping described above, we retrain and evaluate the BERT+ResNet classifier presented in Sec. 4.1. In addition, we perform additional experiments in which the caption text in encoded using the pre-trained Universal Sentence Encoder[3] (USE) (Cer et al., 2018), which returns a 512-dimensional embedding for the text. On the image encoding side, we also experiment with the pre-trained Graph-Regularized Image Semantic Embedding model (Juan et al., 2020), which is trained over ultra-fine–grained image labels over web-sized amounts of data – roughly 260M examples over roughly 40M labels; this model returns a compact, 64-dimensional representation for the image. We concatenate the text and image features into a single vector, and feed it to a fully-connected neural network with 3 hidden layers of 256 units each with ReLU activations (for all but the last one), followed by a softmax layer which computes the logits for the 6 target classes. We divide the 3910 labeled image–text pairs from the ground-truth split of our data into training and test sets, with 3400 and 510 samples, respectively. We use dropout with probability of 0.5, and tune the model

[3] tfhub.dev/google/universal-sentence-encoder-large/3

	Visible	Subjective	Action	Story	Meta	Irrelevant	Weighted
Ground-truth Distribution	46.65%	7.07%	1.31%	19.09%	23.42%	2.46%	
BERT + ResNet	0.64	0.26	0.02	0.52	0.46	0.07	0.52
BERT + GraphRise	0.59	0.15	0.00	0.42	0.34	0.00	0.45
USE + GraphRise	0.69	0.45	0.00	0.57	0.48	0.00	**0.57**

Table 4: The F_1 scores of coherence relation classifiers **with label mapping**. The aggregated Weighted scores use the numbers in the first row as weights.

parameters using the Adam optimizer (Kingma and Ba, 2014) with a learning rate of 10^{-6}.

Results Table 4 shows the results of the single-label prediction experiments, where we present the F_1 scores over each of the individual relations, as well as an overall weighted average. The USE+GraphRise model using the label mapping achieves the highest performance, with an overall F_1 score of 0.57. Next, we describe how we use this classifier's predictions to annotate the training and validation splits of the Conceptual Caption dataset (3.3 million image–captions pairs), in order to train a controllable caption-generation model.

5 Generating Coherent Captions

We use the coherence label predictions on the Conceptual Captions dataset (Section 4) to train a coherence-aware caption generation model.

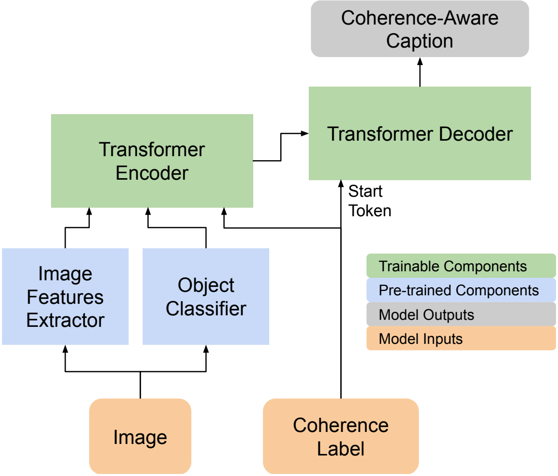


Figure 5: Coherence-aware image captioning model

Model We model the output caption using a sequence-generation approach based on Transformer Networks (Vaswani et al., 2017). The output is the sequence of sub-tokens comprising the target caption. The input is obtained by concatenating the following features.

Image Features We obtain a 64 dimensional representation for the image using the Graph-RISE (**?**) feature extractor, which employs a ResNet-101 network to classify images into some 40M classes. We do not fine tune this image encoder model. We use the 64-dimensional feature available immediately before the classification layer, and embed into the Transformer encoder embedding space using a trainable dense layer.

Detected Objects We obtain object labels for the image using Google Cloud Vision API.[4] We represent each label using pre-trained 512-dimensional vectors trained to predict co-occurring objects on web pages, in a similar fashion as the word2vec model (Mikolov et al., 2013). We embed each of these into the Transformer encoder embedding space using a trainable dense layer.

Coherence relation label This is an input label fed at training time, for which we use the inferred coherence relation for the image–caption pair; at inference time, the label input is used to control the information in the generated caption. Embeddings for the coherence labels are trainable model parameters. Additionally, the relationship label serves as the start token for the Transformer decoder (Figure 5), i.e., it is made available both for the encoder network and directly for the decoder network. When training and evaluating a coherence-agnostic model, this label is set to a special symbol, such as *NONE*, essentially running the model without coherence information. For all models described in this paper, the Transformer network has 6 encoder layers, 6 decoder layers, 8 attention heads, and a 512-dimensional embedding space.

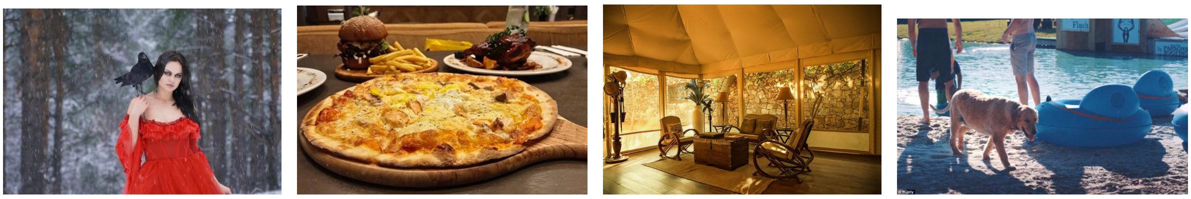
6 Results and Evaluation

In what follows, we discuss evidence for our hypotheses: (a) a coherence-aware model presents information that is aligned with the goal of the

[4] cloud.google.com/vision

	Coherence	Visible	Subjective	Story	Meta
	agnostic	coherence-aware	coherence-aware	coherence-aware	coherence-aware
Visible	**52.1%**	**79.9%**	31.7%	25.0%	42.80%
Subjective	11.4%	2.6%	24.4%	2.6%	1.9%
Action	10.7%	10.8%	6.3%	8.8%	11.4%
Story	**51.3%**	16.0%	**45.0%**	**58.8%**	17.34%
Meta	31.2%	32.8%	15.1%	17.7%	**46.5%**
Irrelevant	12.2%	12.3%	10.7%	9.9%	21.40%
When	9.5%	5.6%	4.1%	17.7%	9.6%
How	21.3%	21.3%	9.6%	25.0%	30.26%
Where	5.3%	8.6%	4.1%	8.8%	16.6%

Table 5: The distribution of coherence relations in image–caption pairs when captions are generated with the discourse–aware model vs the discourse agnostic model (the mode of the distribution in bold).



(a) coherence-aware *Meta*: A girl in the winter forest.
coherence–agnostic: beautiful girl in a red dress.

(b) coherence-aware *Visible*: the pizza at restaurant is seen.
coherence–agnostic: the best pizza in the world.

(c) coherence-aware *Subjective*: beautiful chairs in a room.
coherence–agnostic: the living room of the home.

(d) coherence-aware *Story*: how to spend a day.
coherence–agnostic: dogs playing on the beach.

Figure 6: Captions generated by the coherence-aware and coherence-agnostic models. (Photo credits: YesVideo; TinnaPong; Sok Chien Lim; GoPro)

discourse; and (b) a coherence-aware model can significantly improve caption quality.

Evaluation by expert annotators We train the model described above with the predicted discourse relation labels for image–caption pairs in the Conceptual Captions training and validation sets. The checkpoint with highest CIDEr (Vedantam et al., 2015) score on the validation set is selected for inference and human evaluations. We asked our annotators to annotate a subset of randomly selected image–caption pairs generated by this model. These evaluation images were selected from the Conceptual Captions evaluation set based on their predicted coherence label using the single-label classifier (Section 4) on the captions generated by the coherence-agnostic model (Section 5).

According to our sensitivity power analysis, with a sample size of 1500 image–text pairs, 300 in each category, we are able to detect effect sizes as small as 0.1650 with a power and significance level of 95%. Table 5 shows the result distributions for the coherence-agnostic and coherence-aware model. Differences greater than 3% are statistically significant with ($p < 0.05, t > 2.5$). The ability to control the generated caption using an input coherence relation is clear: when asking for *Visible* (the column under Visible), 79.85% of the captions are evaluated to fit the *Visible* label (non-overlapping), an absolute increase of 27.7% over the coherence-agnostic model (with only 52.09% Visible); at the same time, the rate of *Story* and *Subjective* captions reduces significantly. This reduction is particularly noteworthy in the light of eliminating potential context hallucinations, which are likely to be found under the *Story* and *Subjective* labels.

A similar trend is observed when asking for, e.g., *Meta*: 46.49% of the captions are evaluated to fit the *Meta* label (non-overlapping; the column under Meta), up 15.3% over the coherence-agnostic model (with 31.18% Story). A qualitative analysis of the generated captions shows that captions generated under the *Meta* label include terms such as “screenshot” and “view”, while *Subjective* captions come with adjectives such as “beautiful” or “favorite”. Figure 6 shows several examples.

Crowdsouring and Automatic Metrics For the following experiments, a subset of the Conceptual Captions validation data was selected where the ground-truth captions are labeled as *Visible*.

To compare the quality of the generated captions using our framework with other models, we follow the same crowdsourcing protocol that Sharma et al. (2018) employed for quality assessment. We asked subjects whether the generated captions are "good" or not. 86% of the captions generated by the coherence-aware model were selected as "good" captions, whereas only 74% of the captions generated by the coherence-agnostic model were selected as "good" captions. Note that, based on the human-evaluation data published[5] for the Conceptual Caption Workshop at CVPR 2019, this rate is on average 67% "good" captions for the participating state-of-the-art models in 2019. Furthermore, in a follow-up experiment we ask subjects to choose between a caption generated by the coherence-aware model and one generated by the coherence-agnostic model: 68.2% of the time subjects preferred the coherence-aware result, versus 31.8% for the coherence-agnostic one.

In addition, we study the *quality* and the *relevance* of the captions generated by our model as suggested by (van der Lee et al., 2019). On a scale of 0 to 5, the average scores of the *quality* of the captions generated by the coherence-aware and the coherence-agnostic model are, respectively, 3.44 and 2.83. The average score of the *relevance* for the coherence-aware and the coherence-agnostic conditions are, respectively, 4.43 and 4.40. Note that subjects rated the quality and the relevance of the captions while seeing the questions on the same page. Screenshots and code for the experiments can be found on our GitHub page.

With the exception of the *relevance* condition, the results of the other questions that we asked in the crowdsourcing experiments are statistically significantly different ($p < 0.05, t > |3.1|$), which indicates that subjects prefer captions generated by the coherence-aware model. We also mention here that this difference in quality, albeit significant from a human-rating perspective, is not reflected in the CIDEr score computed on the same data (against the available reference captions). The CIDEr score of the captions generated by the coherence-aware and the coherence-agnostic models are, respectively, 0.958 and 0.964. This is not surprising, as the reference captions used by CIDEr are subject to the same distribution over coherence relations as the rest of the data, and therefore generating caption outputs with a different coherence-relation distribution (Table 5) is unlikely to have a positive impact on reference-driven metrics such as CIDEr.

7 Conclusions and Future Work

Representing coherence in image–text presentations can provide a scaffold for organizing, disambiguating and integrating the interpretation of communication across modalities. We show that cross-modal coherence modeling significantly improves the consistency and quality of the generated text with respect to information needs. This is a step forward towards designing systems that learn commonsense inferences in images and text and use that to communicate naturally and effectively with the users. In addition, the presented dataset, Clue, provides opportunities for further theoretical and computational explorations. The experiments described for the coherence relation prediction task set the stage for designing better models for inferring coherence for images–text pairs.

The presented work has limitations that can be addressed in future research. According to the description of the Conceptual Captions dataset, its captions have been hypernymized. However, by studying the examples in the *Other* category, we discovered an additional coherence relation that exists between an image and caption, in which the caption identifies an object or entity in the image–*Identification*. Examples of this relation involves a caption that mentions the brand of a product or the name of the person in the image. *Identification* is easy to annotate but missing from this work due to the properties of the corpus we annotated. Future work should study this additional relation in the context of caption annotation and generation.

Acknowledgement

The research presented here is supported by NSF Awards IIS-1526723 and CCF-19349243 and through a fellowship from the Rutgers Discovery Informatics Institute. Thanks to Gabriel Greenberg and the anonymous reviewers for helpful comments. We would also like to thank the Mechanical Turk annotators for their contributions. We are grateful to our data annotators, Ilana Torres and Kathryn Slusarczyk for their dedicated work.

[5] http://www.conceptualcaptions.com/winners-and-data

References

Malihe Alikhani, Sreyasi Nag Chowdhury, Gerard de Melo, and Matthew Stone. 2019. Cite: A corpus of image-text discourse relations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 570–575.

Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. 2017. Guided open vocabulary image captioning with constrained beam search. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 936–945, Copenhagen, Denmark. Association for Computational Linguistics.

Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St. John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, Brian Strope, and Ray Kurzweil. 2018. Universal sentence encoder for English. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 169–174, Brussels, Belgium. Association for Computational Linguistics.

Cesc Chunseong Park, Byeongchang Kim, and Gunhee Kim. 2017. Attend to you: Personalized image captioning with context sequence memory networks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. 2019. Show, control and tell: A framework for generating controllable and grounded captions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8307–8316.

Samuel Cumming, Gabriel Greenberg, and Rory Kelly. 2017. Conventions of viewpoint coherence in film. *Philosophers' Imprint*, 17(1):1–29.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Longteng Guo, Jing Liu, Peng Yao, Jiangwei Li, and Hanqing Lu. 2019. Mscap: Multi-style image captioning with unpaired stylized text. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4204–4213.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 770–778.

Jerry R Hobbs. 1978. Why is discourse coherent. Technical report, SRI INTERNATIONAL MENLO PARK CA.

Jerry R. Hobbs. 1985. On the coherence and structure of discourse. Technical report, Center for the Study of Language and Information, Stanford University.

Julia Hunter, Nicholas Asher, and Alex Lascarides. 2018. A formal semantics for situated conversation. *Semantics and Pragmatics*.

Da-Cheng Juan, Chun-Ta Lu, Zhen Li, Futang Peng, Aleksei Timofeev, Yi-Ting Chen, Yaxi Gao, Tom Duerig, Andrew Tomkins, and Sujith Ravi. 2020. Ultra fine-grained image semantic embedding. In *Proceedings of the 13th International Conference on Web Search and Data Mining*, WSDM 20, page 277285, New York, NY, USA. Association for Computing Machinery.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Julia Kruk, Jonah Lubin, Karan Sikka, Xiao Lin, Dan Jurafsky, and Ajay Divakaran. 2019. Integrating text and image: Determining multimodal document intent in Instagram posts. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4614–4624, Hong Kong, China. Association for Computational Linguistics.

Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, et al. 2020. The open images dataset v4. *International Journal of Computer Vision*, pages 1–26.

Alex Lascarides and Matthew Stone. 2009. A formal semantic analysis of gesture. *Journal of Semantics*, 26(4):393–449.

Chris van der Lee, Albert Gatt, Emiel van Miltenburg, Sander Wubben, and Emiel Krahmer. 2019. Best practices for the human evaluation of automatically generated text. In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 355–368, Tokyo, Japan. Association for Computational Linguistics.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer.

Emily E Marsh and Marilyn Domas White. 2003. A taxonomy of relationships between images and text. *Journal of Documentation*, 59(6):647–672.

Scott McCloud. 1993. *Understanding comics: The invisible art*. William Morrow.

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

Christian Otto, Matthias Springstein, Avishek Anand, and Ralph Ewerth. 2019. Understanding, categorizing and predicting semantic image-text relations. In *Proceedings of the 2019 on International Conference on Multimedia Retrieval*, pages 168–176. ACM.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

B Phillips. 1977. A calculus of cohesion. In *Fourth LACUS Forum, Montreal, Canada*.

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind K Joshi, and Bonnie L Webber. 2008. The Penn discourse treebank 2.0. In *LREC*. Citeseer.

Craige Roberts. 2012. Information structure: Towards an integrated formal theory of pragmatics. *Semantics and Pragmatics*, 5:6–1.

Hannah Rohde, Alexander Johnson, Nathan Schneider, and Bonnie Webber. 2018. Discourse coherence: Concurrent explicit and implicit relations. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2257–2267, Melbourne, Australia. Association for Computational Linguistics.

Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. 2018. Object hallucination in image captioning. *arXiv preprint arXiv:1809.02156*.

Philipp Sadler, Tatjana Scheffler, and David Schlangen. 2019. Can neural image captioning be controlled via forced attention? In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 427–431, Tokyo, Japan. Association for Computational Linguistics.

Deborah Schiffrin. 1980. Meta-talk: Organizational and evaluative brackets in discourse. *Sociological Inquiry*, 50(34):199–236.

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 2556–2565.

Kurt Shuster, Samuel Humeau, Hexiang Hu, Antoine Bordes, and Jason Weston. 2019. Engaging image captioning via personality. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Una Stojnic, Matthew Stone, and Ernest Lepore. 2013. Deixis (even without pointing). *Philosophical Perspectives*, 26(1):502–525.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. 2015. CIDEr: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575.

Alakananda Vempala and Daniel Preotiuc-Pietro. 2019. Categorizing and inferring the relationship between the text and image of twitter posts. In *Proceedings of the 2019 Conference of the Association for Computational Linguistics*.

Sida Wang and Christopher D Manning. 2012. Baselines and bigrams: Simple, good sentiment and topic classification. In *Proceedings of the 50th annual meeting of the association for computational linguistics: Short papers-volume 2*, pages 90–94. Association for Computational Linguistics.

Bonnie Webber, Alistair Knott, Matthew Stone, and Aravind Joshi. 1999. Discourse relations: A structural and presuppositional account using lexicalised tag. In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*, pages 41–48. Association for Computational Linguistics.

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *TACL*, 2:67–78.