**Clue: Cross-modal Coherence Modeling for Caption Generation**

Content Hallucination

* Crowd annotations 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.
* Conceptual Captions dataset, Sharma et al. (2018) show that this dataset is large enough, at 3.3M examples, to significantly alleviate content hallucination

1. along with descriptive captions (e.g.,“this is a person in a suit”), the dataset also includes texts that provide
2. contextual background (e.g., “this is the new general manger of the team”) and subjective evaluations (e.g., “this is stylish”).

* Current captioning models trained on Conceptual Captions avoid content hallucination but also introduce different, more subtle and harderto-detect issues related to p

Approach

* coherence-aware approach inspired by the framework of discourse coherence theory (Hobbs, 1978; Phillips, 1977)
* the framework characterizes the inferences that give discourse units a coherent joint interpretation using a constrained inventory of coherence relations
* new image–text coherence relations capture the structural, logical, and purposeful relationships between the contributions of the visual modality and the contributions of the textual modality
* Created a coherence relation annotation protocol for image-caption pairs
* Create Dataset, named Clue with 10,000 image caption pairs over images coming from the Conceptual Captions (Sharma et al., 2018) and Open Images (Kuznetsova et al., 2020) datasets
* training models to automatically **induce coherence-relation annotations** and to build models for **coherence-aware image captionin**g.

Prior Work

Coherence in Images and Captions

* **overlapping set of high-level relations**, inspired both by theoretical work linking discourse coherence to discourse structure and discourse goals

Graphical user interface, application

Description automatically generated

* **Visible**: where text presents information that is intended to recognizably characterize what is depicted in the image, analogous to Restatement relations in text
* **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.
* **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;
* **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.

Data

* 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 imagecaptioning 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.
* 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 not 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”.
* **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

Text

Description automatically generated

* 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.

Table

Description automatically generated

Table

Description automatically generated

Chart, bar chart

Description automatically generated

Predicting Coherence Relations

Baseline

* **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 as input.

Multi-modal classifiers

* **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 mode**l (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 hCLSi representation produced by the BERT-base model (Devlin et al., 2018)

Label Mapping

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%.

Training with Label Mapping

Additional experiments performed in which the caption text in encoded using the pre-trained Universal Sentence Encoder3 (USE), 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 groundtruth 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 parameters using the Adam optimizer (Kingma and Ba, 2014) with a learning rate of 10−6 .

Table

Description automatically generated

Generating Coherent Captions

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

Diagram

Description automatically generated

**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:

1. **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.
2. **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
3. **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.

**Results**

Table

Description automatically generated

A picture containing timeline

Description automatically generated

In what follows, we discuss evidence for our hypotheses: (a) a coherence-aware model presents information that is aligned with the goal of the discourse; and (b) a coherence-aware model can significantly improve caption quality.

**Integrating Text and Image: Determining Multimodal Document Intent in Instagram Posts**

Introduction

Rather, determining author intent with text+image content requires a richer kind of meaning composition that has been called **meaning multiplication** (Bateman, 2014): the creation of new meaning through integrating image and text.

Meaning multiplication includes simple meaning **intersection** or **concatenation** (a picture of a dog with the label “dog”, or the label “Rufus”). But it also includes more sophisticated kinds of composition, such as **irony** or **indirection**, where the text+image integration requires **inference that creates a new meaning**.

To better understand author intent given such meaning multiplication, we create three **novel taxonomies** related to the relationship **between text and image and their combination/multiplication** in Instagram posts

Our taxonomies measure the authorial intent behind the image-caption pair and two kinds of text-image relations: the **contextual relationship** between the literal meanings of the image and caption, and the **semiotic relationship** between the signified meanings of the image and caption.

Previous Work

A wide variety of work in multiple fields has explored the relationship between text and image and extracting meaning, although often assigning a subordinate role to either text or images, rather than the symmetric relationship in media such as Instagram posts.

Taxonomies

A picture containing website

Description automatically generated

Intent Taxonomy

Developed a set of eight illocutionary intents from our examination and clustering of a large body of representative Instagram content, informed by previous studies of intent in Instagram posts. There is some overlap between categories:

1. advocative: advocate for a figure, idea, movement, etc.
2. promotive: promote events, products, organizations etc.
3. exhibitionist: create a self-image reflecting the person, state etc. for the user using selfies, pictures of belongings (e.g. pets, clothes) etc.
4. expressive: express emotion, attachment, or admiration at an external entity or group.
5. informative: relay information regarding a subject or event using factual language.
6. entertainment: entertain using art, humor, memes, etc.
7. provocative/discrimination: directly attack an individual or group.
8. provocative/controversial: be shocking.

Contextual Taxonomy

The contextual relationship taxonomy captures the relationship between the literal meanings of the image and text.

1. **Minimal Relationship:** The literal meanings of the caption and image overlap very little. For example, a selfie of a person at a waterfall with the caption “selfie”. While such a terse caption does nevertheless convey a lot of information, it still leaves out details such as the location, description of the scene, etc. that are found in typical loquacious Instagram captions.
2. **Close Relationship:** The literal meanings of the caption and the image overlap considerably. For example, a selfie of a person at a crowded waterfall, with the caption “Selfie at Hemlock falls on a crowded sunny day”.
3. **Transcendent Relationship:** The literal meaning of one modality picks up and expands on the literal meaning of the other. For example, a selfie of a person at a crowded waterfall with the caption “Selfie at Hemlock Falls on a sunny and crowded day. Hemlock falls is a popular picnic spot. There are hiking and biking trails, and a great restaurant 3 miles down the road ...”.

Semiotic Taxonomy

Dataset

Computational Model