Cross-Modal Coherence for Text-to-Image Retrieval

Malihe Alikhani[1*] **Fangda Han**[2*] **Hareesh Ravi**[2*] **Mubbasir Kapadia**[2]
Vladimir Pavlovic[2] **Matthew Stone**[2]

[1]University of Pittsburgh [2]Rutgers University

malihe@pitt.edu, {fh199, mk1353, vladimir, mdstone}@cs.rutgers.edu, hr268@scarletmail.rutgers.edu

Abstract

Common image-text joint understanding techniques presume that images and the associated text can universally be characterized by a single implicit model. However, co-occurring images and text can be related in qualitatively different ways, and explicitly modeling it could improve the performance of current joint understanding models. In this paper, we train a *Cross-Modal Coherence Model* for text-to-image retrieval task. Our analysis shows that models trained with image–text coherence relations can retrieve images originally paired with target text more often than coherence-agnostic models. We also show via human evaluation that images retrieved by the proposed coherence-aware model are preferred over a coherence-agnostic baseline by a huge margin. Our findings provide insights into the ways that different modalities communicate and the role of coherence relations in capturing commonsense inferences in text and imagery.

Introduction

When using text to retrieve an image for it, humans often rely on commonsense inference. Text and imagery can be related with different kinds of implicit inferences. The content of a caption or sentence that accompanies an image may not necessarily have a lot of overlaps with the content of the image. The text, for instance, can describe quantities that compliment what is depicted in the image (e.g. add two cups of water) or react to what is depicted in an image (e.g. fantastic view). Retrieving imagery is therefore not just finding an image that portrays text content but finding an image that coherently fits with text to convey an integrated message.

These commonsense inferences can be modeled using representations and algorithms informed by approaches to natural language discourse, particularly coherence relations (Hobbs 1985; Asher, Asher, and Lascarides 2003; Taboada and Mann 2006). Coherence relations characterize the inferential links (such as temporal, causal, and logical) that connect the content of text and imagery.

Clues from text and from the typical relations of text and imagery provide important evidence about what kinds of visual content is coherent. Therefore, coherence agnostic methods don't necessarily deliver images that fit naturally with text. By modeling coherence in text and imagery, we can supply images to text that human raters prefer by a large margin. This paper describes models of these broader associations between text and imagery for the task of image retrieval.

Caption: The start of the race.

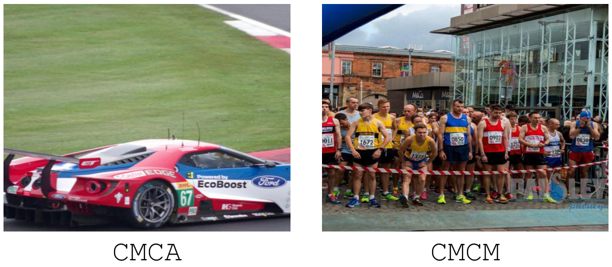


Figure 1: Example retrieved image by the proposed *Cross-Modal Coherence Model* (right) vs *Cross-Modal Coherence Agnostic model* (left) for input caption (top).

We hypothesize that bringing in coherence relations (Alikhani et al. 2020) into the retrieval process, in contrast to personalities defined in (Shuster et al. 2019), should better improve the performance of text-to-image retrieval in a more generalizable way. We build on Salvador et al. (2017) and Chen et al. (2018) and introduce a new framework that integrates coherence relations in text-to-image retrieval task by extracting features for each modality separately then building a lower-dimensional common representation space. Our proposed framework introduces a *Coherence Aware Module* that learns to predict coherence relations that characterize an input image–text pair during training, and predictions from the module are applied during testing through a *Selective Similarity Refinement* technique to further improve the retrieval performance. This module helps the retrieval model focus on the intent of the users and the potential effects of image–text combination.

The examples of Figure 1 illustrate our approach. They contrast the output of our baseline *Cross-Modal Coherence Agnostic model* (CMCA) taken from Han, Guerrero, and Pavlovic (2020) and that of the proposed *Cross-Modal Coherence Model* (CMCM) trained on image-text pairs with *Story* coherence relations. We observe that the proposed CMCM provides more importance to the words *the start* compared to

[*]These authors contributed equally. Author names ordered alphabetically based on last name.

CMCA that concentrates on visually grounded words like *race*. Thus, *Coherence Aware Module* provides more interpretable and robust results, by virtue of explicitly modelling image-text coherence. During inference, the model leverages this knowledge to retrieve relevant images. We evaluate our systems on two image-text coherence datasets namely **CITE++** (Alikhani et al. 2019) and **CLUE** (Alikhani et al. 2020). Each of these datasets correspond to different domains and are annotated with different coherence relations as shown in Table 1 and Table 2. We also analyze the effect of each coherence relation in the datasets by modifying the *Coherence Aware Module* in the proposed CMCM model to detect only the presence of a single relation. These models CMCM$_c$ show which coherence relations improve/reduce the performance when compared with the baselines.[1]

Related Work

Text-to-image retrieval models have been used in several multimodal NLP tasks and applications. Saggion, Pastra, and Wilks (2003) extract syntactic relations from captions for indexing and retrieving photographs of crime scenes. Elliott, Lavrenko, and Keller (2014) use image retrieval as a testbed for learning spatial relationships between image regions using Visual Dependency Representations. None of the previous works in this line have studied a discourse-aware approach for text-to-image retrieval which would best suit the context of the dialogue, inferences between text and imagery in multimodal documents, and the role of coherence in learning better models of image-text alignments.

Inspired by the success of coherence theory that has been applied to other forms of multimodal communication such as gesture (Lascarides and Stone 2009) and comics (McCloud 1993), Alikhani et al. (2019, 2020) characterized coherence relations in text and imagery. Examples of these relations include *elaboration*, when the text include information that is not depicted in the image (e.g. leave it in the oven for 30 minutes) or *subjective* when the text evaluates or reacts to the content of the image (e.g. a delicious pizza). They evaluated the effectiveness of coherence relations on a controlled caption generation task where an image and a coherence relation are given as input while the model generates a caption that adheres to both the image and the relation. We do not train a controllable model as we hypothesize that not all relations equally characterize the image and text in a pair. Though the relations are defined for joint image-text discourse, some coherence relations like "Subjective" in the Clue dataset characterize how the caption relates to the image and not the other way around. Hence, conditioning image retrieval on the relation is not reasonable. The proposed method evaluates the effectiveness of coherence relations by comparing CMCA with CMCM. Note that the proposed *Cross-Modal Coherence Model* is not the same as in Alikhani et al. (2020). Instead, our model learns to predict the coherence relation during training.

With the advent of the Transformer (Vaswani et al. 2017) architecture, there have been large pretrained multimodal transformers (Lu et al. 2019; Chen et al. 2020; Li et al. 2020) that train on large datasets like MSCOCO and others on muiltiple joint image–text learning tasks such as cross modal retrieval. Though they obtain state of the art performance, they do not directly support the addition of our proposed *Coherence Aware Module*. We hence leave the exploration of such architectures for the proposed setting as future work. To the best of our knowledge, this is the first work that comprehensively analyzes the effect of coherence relations for image retrieval and trains text-to-image retrieval models with cross-modal coherence.

Methodology

In this Section we describe the details of our proposed model. We argue that coherence relations characterize the data for multimodal discourse comprehension and hypothesize that a model with coherence (CMCM) will better retrieve relevant images compared to CMCA. Figure 2 shows our framework for CMCM that consists of Image and Text Encoders that project the two modalitied onto a common embedding space optimized over cosine similarity, followed by a *Coherence Aware Module* that predicts the image-text coherence relations that characterize the input image-text pair. We show that addition of *Coherence Aware Module* regularizes the latent space and improves the performance of text-to-image retrieval by modelling the different coherence relations that characterize an image-text pair. To further explicitly use the predictions from *Coherence Aware Module*, we propose a Selective Similarity Refinement technique to refine and rank the retrieval result.

To further analyze the performance of each coherence relation on the overall model, we train separate CMCM$_c$ models that are aware of only one relation. The *Coherence Aware Module* is modified to predict only the presence of a particular relation c (through binary classification in contrast to multi–label classification in the overall model) in these models. Further details about the model architecture are described in Sec. .

Model Architecture

In order to train CMCM for text-to-image retrieval,we write $\mathbf{S} = [w_1, w_2, ..., w_m]$ for the input natural language text composed of m words. (In principle w_i could be words, phrases, sentences or any other semantic unit of text.) Similarly, we write $\mathbf{I}$ for the corresponding image. Given text $\mathbf{S}_i$, the objective of an image retrieval model is to retrieve the paired image $\mathbf{I}_i$ from an image pool $\{\mathbf{S}_j\}, j \in [1, ..., N]$, where N is the number of images in pool.

Image Encoder The image encoder $\mathrm{E_I}$ is a pretrained Resnet-50 (He et al. 2016) followed by a bottleneck layer to transform image feature to the shared latent space. Image is first resized to 224×224, and then forwarded through $\mathrm{E_I}$ to get the image embedding $\mathrm{f_I} \in \mathbb{R}^{300}$.

Text Encoder The text encoder $\mathrm{E_S}$ starts from a pretrained word2vec model that embeds each word into a 300 dimensional vector. The word2vec model is trained using Gensim (Řehůřek and Sojka 2010). The maximum length of the text sequence considered is 200 for CITE++ and 40 for Clue based on the longest sentences in the dataset. Then, the word

[1] Our code and data are attached with this submission and will be made public on acceptance.

Relation	Question	Description	Positive rate
Expansion	Q2	The image gives visual information about the step described in the text.	0.821
ImageNeeded	Q3	You need to see the image in order to be able to carry out the step properly.	0.115
Elaboration$_t$	Q4	The text provides specific quantities (amounts, measurements, etc.) that you would not know just by looking at the picture.	0.329
Elaboration$_{i-tool}$	Q5	The image shows a tool used in the step but not mentioned in the text.	0.193
Temporal$_{i<t}$	Q6	The image shows how to prepare before carrying out the step.	0.158
Temporal$_{i>t}$	Q7	The image shows the results of the action that is described in the text.	0.588
Temporal$_{i=t}$	Q8	The image depicts an action in progress that is described in the text.	0.313

Table 1: Coherence relations, their distribution and entropy in CITE++ dataset. We use the question identifier and the relation name interchangeably in the paper. *Positive rate* is the percentage of samples that are labeled as ‘Yes’ for that question

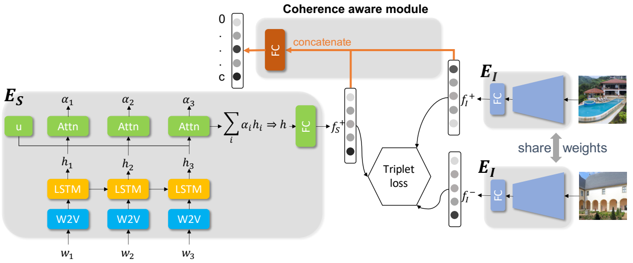


Figure 2: Framework of our proposed *Cross-Modal Coherence Model*. $\mathrm{E_S}$ stands for text encoder, $\mathrm{E_I}$ stands for image encoder, α_i stands for the attention of word embedding h_i

embeddings are given as input to a Long Short Term Memory (LSTM) network to get each word representation. We next apply an *attention mechanism* (Vaswani et al. 2017) to the LSTM representations, which learns the attention for each word and helps the model attend to key words that a re important to our task. Finally a fully-connected layer is applied to encode the joined representation of all words h into the shared latent space.

The outputs of the text and image encoders are then used with a triplet objective using cosine similarity trained with hard negative mining. Hard negative mining targets on the most difficult negative image for each query in a batch based on the similarities to improve performance (Hermans, Beyer, and Leibe 2017). Let $s(a,b) = a^T b / \sqrt{(a^T a)(b^T b)}$ measure the cosine similarity between two vectors a and b, then the objective for the retrieval task per sample is given by Equation 1,

$$\begin{aligned} trip\,(a,p,n) \; &= \; s(a,p) \; - \; s(a,n) - \; \alpha, \\ \mathcal{L}_{ret} \, &= \, \min\{0, \; trip(f_S^+, \; f_I^+, \; f_I^-)\} \qquad (1) \\ &\quad + \, \min\{0, \; trip(f_I^+, \; f_S^+, \; f_S^-)\}, \end{aligned}$$

where L_{ret} is the retrieval loss, f_S^+ and f_I^+ are outputs of text and image encoder for a pair of text and image while f_S^- is a text output that does not correspond to current image and f_I^- is an image output that does not correspond to current text. The margin α is set to 0.3 by cross-validation.

Coherence Aware Module Instead of relying only on the encoders, we also leverage coherence relations labelled by humans. We add a *Coherence Aware Module* that takes the normalized features from both text and image encoders as input and then passes them through a multi-layer perceptron to predict the relations.

The dimension of the final linear layer is equal to the number of relations in the dataset when trained with all relations (i.e. multi-label classification) and 1 when trained with a single relation (i.e. single-label classification). We use Binary Cross Entropy (BCE) as the loss function and the objective of *Coherence Aware Module* for one sample is,

$$\mathcal{L}_{cls} = \textstyle\sum_c w_c \left(y_c \log(x_c) + (1 - y_c) \log(1 - x_c) \right), \quad (2)$$

where x_c is the probability assigned to relation c by the model while y_c is the ground truth binary value. Since the relations are not equally distributed in the dataset, we balance the training of different relations by giving a weight w_c for each relation that is reciprocal to its proportion in the dataset. For $\mathtt{CMCM}_c$ models, the summation is removed as there is only one relation that is predicted.

The model is thus trained in a multi-task setting where the coherence predictor is the auxiliary task. The final objective over the entire batch with batch size N is given in Equation 3,

$$\mathcal{L}_{total} = \frac{1}{N} \sum_{n=1}^{N} \left(\mathcal{L}_{ret}^n + \lambda_{cls} \mathcal{L}_{cls}^n \right), \qquad (3)$$

where λ_{cls} is the weight associated with the coherence aware module and is chosen empirically as described later.

Selective Similarity Refinement

The performance of the retrieval model depends on the similarities between a query caption S_i and all possible images $\{I_j\}, j \in [1, ..., N]$ (including the ground-truth image). We use cosine similarity (though any other valid similarity metric can be used) and notate the similarity between query S_i and one image I_j as $\theta_{i,j} = cosine(S_i, I_j)$.

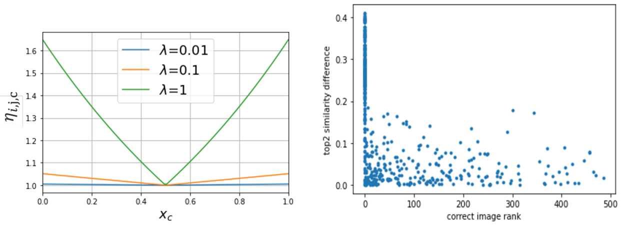
Leveraging Confidence Score We use the coherence prediction from *Coherence Aware Module* to refine the similarity between an image–text pair for retrieval during inference. Note that we do not know the coherence relation characterizing a ground truth image–text pair. However, a well trained *Coherence Aware Module* is expected to predict coherence for a ground truth image–text pair with high confidence. We define a confidence function for a query caption S_i and one possible image I_k as

$$\eta_{i,j,c} = e^{\lambda|x_{i,j,c}-0.5|}, \tag{4}$$

$$\eta_{i,j} = \sum_{c} \eta_{i,j,c}, \tag{5}$$

where x_c is defined in Equation 2, and λ is a hyperparameter decided by cross validation. Confidence function with different λ are shown in Figure 3 (a). We can see that lower λ decreases the impact of the confidence function. We set $\lambda = 0.13$ for CITE++ and $\lambda = 0.12$ for Clue datasets empirically. The refined similarity is defined as,

$$\bar{\theta}_{i,j} = \theta_{i,j} * \eta_{i,j} \tag{6}$$



(a) (b)

Figure 3: (a) Confidence function $\eta_{i,j,c}$ with different λ. (b) Correct image rank vs. the difference between the similarities of the top 2 retrieved images on CITE++ validation set

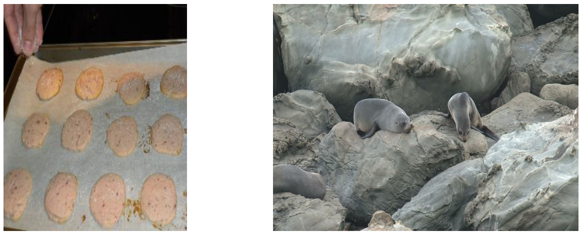
Selective Refinement Though confidence score helps, by itself the score is a weak indicator performing only slightly better than random. We hence limit the use of confidence score to difficult examples. We hypothesize that similarity between a correct image–text pair should on average be "α" larger than that of a wrong image–text pair because of Equation 1. In Figure 3 (b), we verify this hypothesis by plotting the rank of ground truth image vs. the difference between the similarities of the top 2 retrieved images with the query caption. We observe that when the difference between the similarities of the top 2 images (Δ) is large enough (*e.g.* ≥ 0.2), the retrieval is always successful (*e.g.* ground truth image rank = 1). Based on this analysis, we select difficult query captions as those with $\Delta < T$, where T is a hyperparameter chosen as 0.1 empirically. We use the refined similarity Equation 6 for ”difficult” examples during inference.

Image-Text Coherence Datasets

We study the efficacy of `CMCM` for image-retrieval by leveraging two image-text datasets CITE++ and Clue (Alikhani et al. 2020) that are annotated with image-text coherence relations. CITE++ is extended by us from CITE (Alikhani et al. 2019) adding 2242 image-text pairs annotated with coherence relations.

CITE++

We extend the CITE dataset which is a subset of a popular recipe dataset RecipeQA (Yagcioglu et al. 2018). The RecipeQA dataset consists of multimodal recipes that contains textual instructions accompanied by one or more images. CITE leveraged recipes that have one-to-one correspondence between instruction and image, *e.g.* every instruction in the text has one image that visualizes it. Using Amazon Mechanical Turk, the authors obtained answers to 10 questions that help characterize the relationship between image and text. We choose the questions that are best suited to train `CMCM` as described in Table 1. The original dataset has 2057 image-text pairs annotated with True/False answers to these questions indicating presence/absence of the coherence relation. To perform a more comprehensive experiment, we collected 2242 more pairs using the same annotation protocol, giving us a total of 4299 image-text pairs. The distribution of relations in the entire dataset is given in Table 1. Figure 4 [a] shows an example from CITE++ dataset.



(a) Once they have baked remove them from the oven and sprinkle lightly with sugar. After you have dressed them allow them to cool for about 5 minutes and serve

(b) Seals fighting for a spot to sleep on the rocks

Figure 4: Example image-text pairs from CITE++ (a) and Clue (b) datasets. Image-text pair on the left has relations *Expansion*, *Elaboration* and *Temporal*$_{i>t}$ while the one on the right has relations *Action* as *Visible*

Relation	Visible	Subj.	Action	Story	Meta	Irr.
% Positive	67.4	6.6	15.7	24.3	39.1	08.7

Table 2: Coherence relations (Subj. is Subjective and Irr. is Irrelevant) and their distribution in Clue dataset (Alikhani et al. 2020)

CLUE

The Clue dataset (Alikhani et al. 2020) is constructed using the much larger Conceptual Captions dataset (Sharma et al. 2018) which is primarily an image captioning dataset like COCO (Lin et al. 2014). Clue annotated 7559 image-caption pairs with six coherence relations to summarize the structural, logical and purposeful relationships between the contributions of texts and images. Example image-caption pair with coherence relations are shown in Figure 4 (b).

Experimental Setup

Network Details. The backbone of image encoder E_i is ResNet-50, with one additional batch normalization layer and one fully-connected layer to transform the feature into the shared space ($\mathbb{R}^{1024}$). The word2vec model encodes each word into a vector of $\mathbb{R}^{300}$, text encoder E_t takes the vector as input and forwards it through a bidirectional, one-layer

LSTM module following an attention layer (Vaswani et al. 2017), and finally the attention-weighted summation of word features is also transformed into the shared space ($\mathbb{R}^{1024}$) by a batch normalization layer and a fully-connected layer. *Coherence Aware Module* contains one fully-connected layer, adding more layers does not improve performance.

Evaluation Metrics. We evaluate the retrieval performance of all the models using the median retrieval rank (MedR) and the recall at K (R@K) metrics following existing works on text–to–image retrieval (Han, Guerrero, and Pavlovic 2020; Frome et al. 2013). The retrieval range is set to be 500. Since CITE++ and Clue have image-text pairs that exhibit complex relationships, we also perform comprehensive user study to evaluate the performance of the model. **MedR** ($0 \leq$ MedR ≤ 1) is computed as the median rank of the true positive over all queries, a lower MedR suggests better performance. **R@K** ($0 \leq$ R@K ≤ 100) computes the percentage of true positives recalled among the top-K retrieved candidates, higher indicates better performance. Here we only report the results of retrieving image by using the caption as query.

Dataset Partition. In our experiments, we evaluate the model and the coherence relations on CITE++ and Clue datasets independently. We split the CITE++ dataset as $3439/860$ for training/testing while the Clue dataset as $6047/1512$ for training/testing. 10% of the training data is used as validation. Further training and hyperparameter details are given in the appendix.

Comparative Evaluation. For both the datasets, we train the proposed model and compare with various baselines as shown in Table 3. The baseline CMCA is similar to existing CNN-RNN architectures such as (Xu et al. 2015; Ravi et al. 2018; Yang et al. 2020). Note we also compare with CMCM$_c$, which only uses one specific relation to train the system. We perform this experiment primarily to analyze the effect of each relation as not all relations contribute equally to the retrieval system. This also helps us better understand the influence of different relations on the proposed *Cross-Modal Coherence Model* model. Though it is possible to develop transformer based models for the proposed setting, we use GRUs and CNNs because of the low cardinality of the datasets and the necessity of large datasets for transformer based models (Inan et al. 2021; Ganesh et al. 2021; Crawford 2021).

Model	*Coherence Aware Module*	Relations	Attention
Base	✗	-	✗
CMCA	✗	-	✓
CMCM-NoAttn	✓	All	✗
CMCM	✓	All	✓
CMCM$_c$	✓	c	✓

Table 3: Description of the models used for comparison. -NoAttn means removing the attention module from the proposed model.'All' relations indicate that the *Coherence Aware Module* is trained with all the relations in a multi-label multi-class setting. c indicates only one relation is used with the *Coherence Aware Module* in a binary classification setting.

Results and Discussion

CMCM vs CMCA

The results on CITE++ dataset are shown in Table 4. As can be seen, having attention over the text clearly improves retrieval performance. This can be attributed to the lengthy texts in CITE++ dataset. Moreover, we observe that CMCM model performs better than CMCA and $Base$ across all metrics though with variable significance. For example, MedR for CMCA model is 5.4 but all CMCM models achieve average MedR of less than 5.0. Moreover the standard deviation is also lower indicating more robust performance. The results on the Clue dataset are given in Table 5. We observe that both the attention mechanism and the coherence-aware module improve the performance. We use the example in Fig. 1 to intuitively explain the effect of *Coherence Aware Module*. Note CMCA retrieves the incorrect image as there are more images of "races" in general than there are of "start of a race" in the dataset. Also, the text "start of a race" communicates a story rather than factually describe elements in an image, CMCA ignores the different characteristics by which an image-text pair can be related thereby producing the most commonly found semantically similar image. CMCM resolves this concern by considering coherence relations between the two modalities and retrieves the correct image. We observe that all per-relation CMCM$_c$ models perform better than CMCA. In some instants, per-relation models perform better than CMCM, confirming the conjecture that not all relations contribute in increasing the performance of the retrieval model. We perform additional analysis on per-relation contribution to CMCMs performance in the Appendix.

Impact of Similarity Refinement. To evaluate the contribution of selective similarity refinement, we compare MedR based on θ_i and $\bar{\theta}_i$ of the same model in Figure 5. The CMCM (except 'NoAttn') variants clearly outperform CMCA and baseline. Moreover, the selective refinement technique improves the result of almost all the CMCM models even further by a large margin as can be seen by the difference between the blue and orange bars. In Clue dataset, in most cases, model using selective similarity refinement performs better than the same model without refinement, proving the effectiveness of the refinement technique. For CMCM$_{Irrelevant}$ model on Clue (last two bars on Figure 5 right), applying the refinement severely degrades the performance. We believe that 'Irrelevant' relation does not effectively characterize the relationship between an image–text pair on top of its low positivity score.

Human Evaluation

Both CITE++ and Clue have image-text pairs with complex coherence relations in contrast to datasets like MSCOCO that have predominantly just *Visible* relation. Hence, considering the ground truth as gold standard is not reasonable. Given the wide distribution of different relations in the datasets, the quantitative metrics (*e.g.* MedR and Recalls) are unreliable for the proposed setting. Therefore, we perform human evaluation where the top 1 retrieved images by CMCA and CMCM models are shown for pairwise comparison.

We recruit 250 participants through Amazon Mechanical

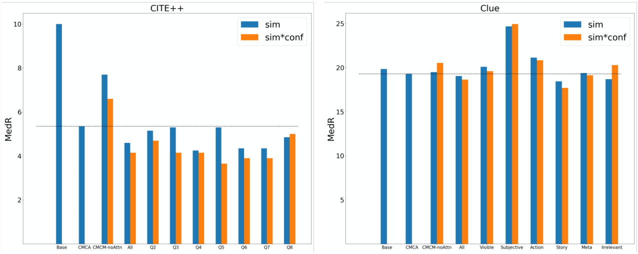


Figure 5: Comparison MedR between baseline, CMCA and different CMCM variants; as well as the comparison between the same model with and without selective similarity refinement. Left: CITE++ dataset. Right: Clue dataset

	MedR↓	R@1↑	R@5↑	R@10↑
Base	$10.0^{\pm 3.7}$	45.7	48.4	50.6
CMCA	$5.4^{\pm 2.3}$	46.0	50.1	53.8
CMCM-NoAttn	$6.6^{\pm 2.5}$	46.0	49.5	52.0
CMCM	$\mathbf{4.2^{\pm 1.2}}$	**46.5**	**51.4**	**53.9**
CMCM$_{Q_2}$	$4.7^{\pm 2.0}$	46.4	50.6	53.4
CMCM$_{Q_3}$	$4.2^{\pm 1.3}$	46.2	51.1	54.2
CMCM$_{Q_4}$	$4.2^{\pm 1.3}$	46.2	51.2	54.2
CMCM$_{Q_5}$	$\mathbf{3.7^{\pm 1.3}}$	46.6	**51.5**	**54.4**
CMCM$_{Q_6}$	$4.6^{\pm 1.4}$	45.9	50.8	53.4
CMCM$_{Q_7}$	$3.9^{\pm 1.7}$	**46.9**	51.2	54.1
CMCM$_{Q_8}$	$5.0^{\pm 1.7}$	46.4	50.8	53.8

Table 4: Quantitative comparison in CITE++ dataset. The relations corresponding to each Q_i are shown in Table 1. ↓ indicates that lower the better and ↑ indicates that higher the better.

Turk. All subjects were US citizens, agreed to a consent form approved by *anonymized institution* review board, and were compensated at an estimated rate of USD 15 an hour. We showed subjects the caption, the top image retrieved by the coherence aware and the coherence agnostic model for five relations from both the datasets and asked them to choose one of the following options:
(1) I prefer image A (2) I prefer image B (3) The images are exactly the same (4) Neither of the images is a good match for this text. The order of images is random and each example was ranked by three workers and the final rank is decided via majority voting. The results are shown in Table 6. It can be seen that the images retrieved by the proposed model are preferred by humans. More importantly, the difference in preference is significant in contrast to the quantitative metrics. We can also see that the difference in preference between CMCM and CMCA is higher when the relation is *Subjective* or *Story* when compared to regular captions (see *Visible*), indicating the importance of explicitly modeling coherence relations for cross-modal understanding. The results of the t-test shows that the differences observed in CMCM and CMCA category are all statistically significant ($p < 0.01, \text{t} > 14.1$). The results of the sensitivity power analysis shows that our experiment detects effect sizes as small as 0.17 with a power and significance level of 95%. These results effectively show that the quantitative metrics such as MedR and Recall can not solely measure performance of the model especially given the nature of the dataset and coherence relations.

	MedR↓	R@1↑	R@5↑	R@10↑
Base	$19.8^{\pm 1.9}$	11.6	28.6	38.3
CMCA	$19.3^{\pm 2.0}$	13.2	30.6	40.0
CMCM-NoAttn	$20.6^{\pm 2.6}$	12.4	28.9	38.8
CMCM	$\mathbf{18.7^{\pm 1.6}}$	**13.8**	**31.6**	**40.6**
CMCM$_{Visible}$	$19.6^{\pm 3.1}$	13.4	**31.7**	**41.1**
CMCM$_{Subjective}$	$25.0^{\pm 3.1}$	12.9	29.4	38.0
CMCM$_{Action}$	$20.9^{\pm 2.1}$	11.7	28.4	38.0
CMCM$_{Story}$	$\mathbf{17.7^{\pm 1.7}}$	13.0	30.7	41.5
CMCM$_{Meta}$	$19.2^{\pm 1.5}$	13.1	31.0	40.7
CMCM$_{Irrelevant}$	$20.3^{\pm 1.9}$	12.6	31.1	40.9

Table 5: Quantitative comparison of the models trained and evaluated on Clue dataset.

	Better	Worse	Both Good	Both Bad
CMCM$_{Visible}$	**24%**	17%	46%	13%
CMCM$_{Subjective}$	**53%**	10%	7%	30%
CMCM$_{Story}$	**40%**	10%	33%	17%
CMCM$_{Meta}$	**56%**	9%	25%	10%
CMCM$_{Q_7}$	**43%**	17%	27%	13%

Table 6: Human evaluation results. Values indicate the percentage of samples for which humans voted the output of CMCM as **Better**, **Worse**, **Both Good**, **Both Bad** when compared with CMCA.

Qualitative Analysis

To further understand the behavior of the model, we investigate the attention weights over input text for CMCM and CMCA models. In example Figure 6 (a), the proposed coherence-aware model retrieves the ground truth within the top 5 images. We can see from Figure 7 (left) that adding *Coherence Aware Module* increases the weight on words *horse* and *grazing* relative to the agnostic model. This can be attributed to the model's ability to predict the associated coherence relation to help retrieve the right image. The CMCA model, however, attends more to commonly visualized words like *forest* and *outdoors*. Similarly, in Figure 6 (right), CMCM shows improved attention weights for words like *male* and *golfer*. The result is the model being able to retrieve the correct image in top 1 though both models retrieve images of *Vector illustration* in the top 5 Figure 6 [b]. More examples for other relations are provided in the Appendix.

In CITE++ dataset, we observe similar behavior as shown in Figure 6 [c]. The relation *Temporal*$_{i>t}$ characterizes the temporal correlation between an image and text where the image visualizes the result of the process described in the corresponding text. These relations are difficult to implicitly understand as the text is no different from any other step in the recipe. Training with *Coherence Aware Module* that explicitly models temporal relation improves the performance of image

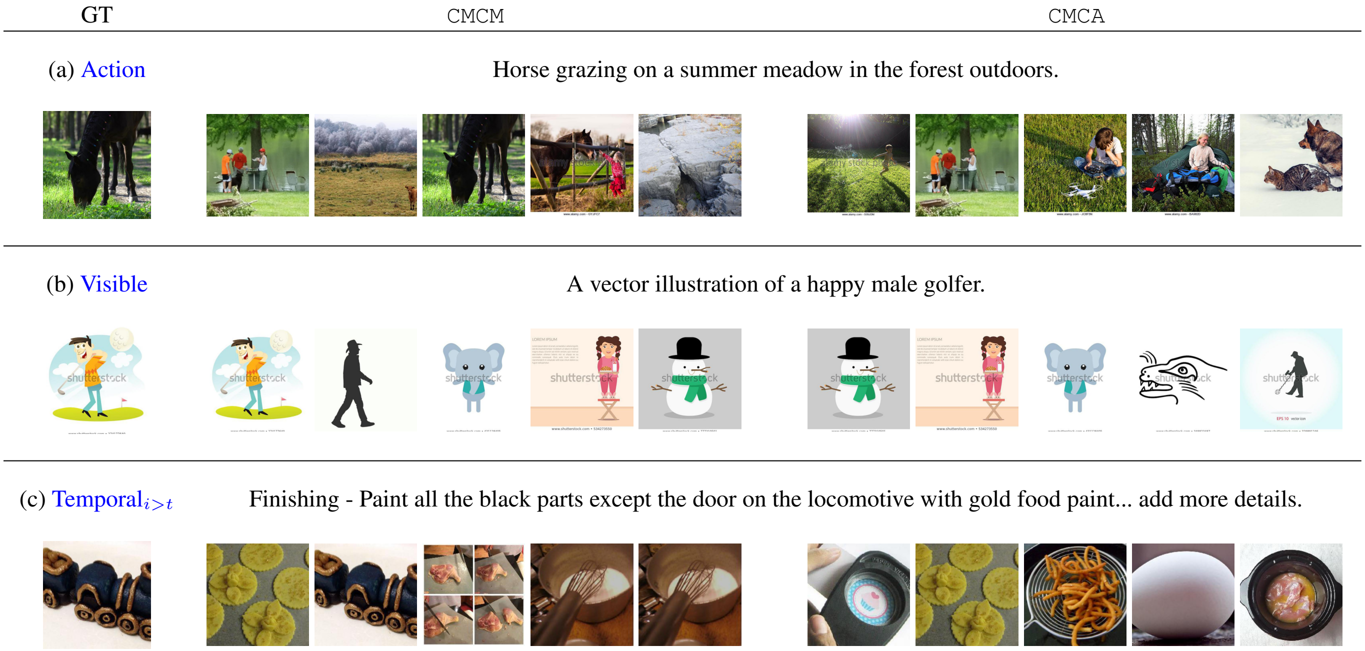

GT
CMCM
CMCA
(a) Action
Horse grazing on a summer meadow in the forest outdoors.
(b) Visible
A vector illustration of a happy male golfer.
(c) Temporal$_{i>t}$
Finishing - Paint all the black parts except the door on the locomotive with gold food paint... add more details.

Figure 6: The ground truth image (Left) and the top 5 retrieved images by the CMCM and the CMCA models for two examples. The coherence relation (in blue) and caption are given above the images. The image-text pair in example (a) has *Action* relation while in example (b) has *Visible* relation. In example (a) the CMCM model leverages the *Action* coherence relation to retrieve images that depict some action in the top 5. Similarly in example (c) images retrieved by proposed our model with CAM retrieves images that depict the result of a process as given by the *Temporal*$_{i>4}$ relation, whereas the agnostic model shows images that depict action in progress.

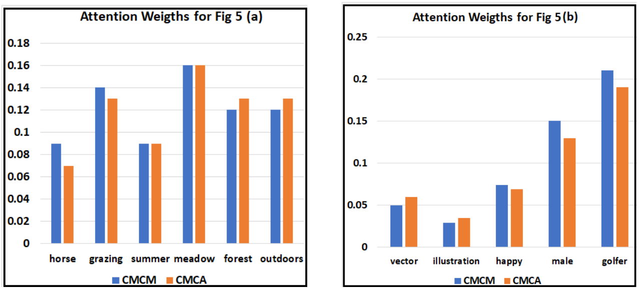


Figure 7: Attention weights for CMCM and CMCA models for example (a) and (b) in Figure 6.

retrieval. For example, we can see in Figure 6 that all top 5 images retrieved by the CMCM are images that visualize the *result of a process*, in contrast to CMCA model that shows images of *the step being carried out* as well.

Predicting Coherence Relations

To further understand the effect and importance of coherence relations, we analyze the model's ability to predict the presence of a coherence relation given the ground truth image and text. For this, we use the models trained using the original objective in Equation 3. We provide the ground truth image and text as input and calculate the Average Precision (AP) of coherence relation prediction. The results are provided in the Appendix. We can see that in most cases as expected, the CMCM$_c$ algorithms perform reasonably well. We haven't provided the results for CMCA models as they were not trained with coherence relations. These results are comparable to similar experiments performed in (Alikhani et al. 2020) though in their experiment, classification was the only objective. Interestingly *Subjective* relation has very low AP (cf. appendix) similar to retrieval performance but the proposed model obtains significance gain in performance in the human evaluation. This reinforces the unreasonableness of the Recall and Rank metrics for this task.

Conclusion

Automating the understanding and generation of multimodal discourse requires a joint understanding of co-occurring images and text. Our study shows the effectiveness of cross-modal coherence modeling for text-to-image retrieval tasks. Our evaluation shows that the performance of the coherence-aware model is significantly better compared to the agnostic models. We also observe that the existing Recall based quantitative metrics for text-to-image retrieval are unreliable and fail to meaningfully evaluate retrieval systems especially when image-text pairs can be characterized by different coherent relations. Future work involves developing new transformer-based coherence-aware metrics that can better measure the performance of retrieval models. Based on the evidence shown in this paper, an important extension is to annotate existing datasets with coherence relations to further improve semantic joint understanding of image and text.

References

Alikhani, M.; Chowdhury, S. N.; de Melo, G.; and Stone, M. 2019. CITE: A Corpus of Image–Text Discourse Relations. *arXiv preprint arXiv:1904.06286*.

Alikhani, M.; Sharma, P.; Li, S.; Soricut, R.; and Stone, M. 2020. Cross-modal Coherence Modeling for Caption Generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 6525–6535.

Asher, N.; Asher, N. M.; and Lascarides, A. 2003. *Logics of conversation*. Cambridge University Press.

Chen, J.-J.; Ngo, C.-W.; Feng, F.-L.; and Chua, T.-S. 2018. Deep understanding of cooking procedure for cross-modal recipe retrieval. In *Proceedings of the 26th ACM international conference on Multimedia*, 1020–1028.

Chen, Y.-C.; Li, L.; Yu, L.; Kholy, A. E.; Ahmed, F.; Gan, Z.; Cheng, Y.; and Liu, J. 2020. Uniter: Universal image-text representation learning. In *European Conference on Computer Vision*.

Crawford, K. 2021. *The Atlas of AI*. Yale University Press.

Elliott, D.; Lavrenko, V.; and Keller, F. 2014. Query-by-Example Image Retrieval using Visual Dependency Representations. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, 109–120. Dublin, Ireland: Dublin City University and Association for Computational Linguistics.

Frome, A.; Corrado, G. S.; Shlens, J.; Bengio, S.; Dean, J.; Ranzato, M.; and Mikolov, T. 2013. Devise: A deep visual-semantic embedding model. In *Advances in neural information processing systems*, 2121–2129.

Ganesh, P.; Chen, Y.; Lou, X.; Khan, M. A.; Yang, Y.; Sajjad, H.; Nakov, P.; Chen, D.; and Winslett, M. 2021. Compressing large-scale transformer-based models: A case study on bert. *Transactions of the Association for Computational Linguistics*, 9: 1061–1080.

Han, F.; Guerrero, R.; and Pavlovic, V. 2020. CookGAN: Meal Image Synthesis from Ingredients. In *The IEEE Winter Conference on Applications of Computer Vision*, 1450–1458.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Hermans, A.; Beyer, L.; and Leibe, B. 2017. In defense of the triplet loss for person re-identification. *arXiv preprint arXiv:1703.07737*.

Hobbs, J. R. 1985. On the coherence and structure of discourse.

Inan, M.; Sharma, P.; Khalid, B.; Soricut, R.; Stone, M.; and Alikhani, M. 2021. COSMic: A Coherence-Aware Generation Metric for Image Descriptions. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, 3419–3430.

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

Li, X.; Yin, X.; Li, C.; Zhang, P.; Hu, X.; Zhang, L.; Wang, L.; Hu, H.; Dong, L.; Wei, F.; et al. 2020. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *European Conference on Computer Vision*, 121–137. Springer.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*, 740–755. Springer.

Lu, J.; Batra, D.; Parikh, D.; and Lee, S. 2019. ViLBERT: pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, 13–23.

McCloud, S. 1993. Understanding comics: The invisible art. *Northampton, Mass*.

Ravi, H.; Wang, L.; Muniz, C.; Sigal, L.; Metaxas, D.; and Kapadia, M. 2018. Show me a story: Towards coherent neural story illustration. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 7613–7621.

Řehůřek, R.; and Sojka, P. 2010. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, 45–50. Valletta, Malta: ELRA.

Saggion, H.; Pastra, K.; and Wilks, Y. 2003. Nlp for indexing and retrieval of captioned photographs. In *10th Conference of the European Chapter of the Association for Computational Linguistics*.

Salvador, A.; Hynes, N.; Aytar, Y.; Marin, J.; Ofli, F.; Weber, I.; and Torralba, A. 2017. Learning cross-modal embeddings for cooking recipes and food images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3020–3028.

Sharma, P.; Ding, N.; Goodman, S.; and Soricut, R. 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*, 2556–2565.

Shuster, K.; Humeau, S.; Hu, H.; Bordes, A.; and Weston, J. 2019. Engaging image captioning via personality. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 12516–12526.

Taboada, M.; and Mann, W. C. 2006. Applications of rhetorical structure theory. *Discourse studies*, 8(4): 567–588.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. In *Advances in neural information processing systems*, 5998–6008.

Xu, K.; Ba, J.; Kiros, R.; Cho, K.; Courville, A.; Salakhudinov, R.; Zemel, R.; and Bengio, Y. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, 2048–2057.

Yagcioglu, S.; Erdem, A.; Erdem, E.; and Ikizler-Cinbis, N. 2018. RecipeQA: A Challenge Dataset for Multimodal Comprehension of Cooking Recipes. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 1358–1368.

Yang, L.; Hu, H.; Xing, S.; and Lu, X. 2020. Constrained LSTM and Residual Attention for Image Captioning. *ACM Trans. Multimedia Comput. Commun. Appl.*, 16(3).