Image-text discourse coherence relation discoveries on multi-image and multi-text documents

Hanzhong Zheng¹, Zhexiong Liu¹, Dejia Shi²*

¹Department of Computer Science, The University of Pittsburgh, Pittsburgh, PA, 15213, USA.
²School of Computer and Information Engineering, Hunan University of Technology and Business, Changsha, 410205, China.

*Corresponding Author: Dejia Shi. Email: 1832@hutb.edu.cn

Abstract. The occurrences of the image-text pairs are common in social media such as image captions, image annotations, and cooking instructions. However, the links between images and texts as well as the coherence relationship among them are always implicit. We developed a multi-modal approach by firstly establishing the links between image and text in an unsupervised way, then discovered their coherence relations through computational models of discourse to improve the consistency and quality of the images and their corresponding texts and also explore a dynamic and flexible image-text coherence relation.

1. Introduction
Image-text pairs commonly co-exist together within a collection of documents, especially on the social medias, e.g., news stories, advertisements, and food recipes. The links among the images and the texts are not fully illustrative. A solid link between the images and descriptive texts greatly improves the amount of information provided by the image-text pairs. The analysis of the image-text pairs requires a multi-modal model.

However, one of the biggest challenges is that the corresponding texts contain many noises that cannot capture the salient content of the images and impair people’s reading experience, especially for food recipe dataset. Recipes are intentionally required to be informative and clear, but those which are posted on the social media are written by the nonprofessional-writers. Another challenging aspect is to bridge the semantic gap. Since the image information extraction usually is limited to low-level features. However, multimodal analysis, especially for linking the image and text, involves object recognition and scene understanding, which need semantic and high-level abstractions of visual and textual features. This is commonly known as the semantic gap. Without connecting the textual contents, the interpretation of the image may go in other directions. Many current researches have been done to bridge the semantic gap on the visual side in concept recognition and image captioning [1,2]. The limitation of these approaches is only focusing on one perspective of the visual contents e.g. objects, but it is lacking of scene interpretation that is above the visible scene content. Especially for analyzing multimodal information, the modality gap is the reason for causing semantic gap. A deeper understanding of the multimodal interplay between image and text pairs is required to eliminate the modality gap, but the imagetext pair interplay is usually complicated and versatile.

Establishing explicit links between the visual and textual contents can be the solution to eliminate the semantic gap. Otto C et. al. [3], Henning CA and Ewerth R [4,5] proposed two evaluation metrics
for measuring the shared contents between image-text pairs: crossmodal mutual information (CMI) and semantic correlation, which are based on the relation between image and text depending on their depicted contents and semantic context. Those two metrics can also be used for evaluation of the links between image-text pairs.

Modeling the coherence relations in image-text pairs allow you to explicitly explore the relationships in content of discourse units, this can be used for commonsense inference prediction in texts. The study on the discourse coherence framework proposed by Jerry R Hobbs [6] and Phillips B [7] developed a constrained inventory of coherence relation to assign the discourse units of image captions/texts with a coherent joint interpretation. Malihe A et. al. [8] presented a coherence-aware approach that can scale image-caption consistency to overcome the limitations of lacking enough data for coherence model training. Moreover, they are interested in the image-text coherence discourse structural, logical, and purposeful relationships between image modularity and text modularity. The current applications of discourse coherence are machine comprehension [9], text summarization [10] and sentiment analysis [11].

In this paper, we firstly develop an unsupervised model for establishing explicit links between image-text pairs among intra-documents. The formation links will be based on the cosine similarity. Then, we use the coherence-aware model to assign discourse units to textual contents. The coherence relations can improve the consistency and quality of textual information such as the image captions, cooking instructions. For coherence relation was trained on the well-labeled dataset as the supervised classification problem. After validating our model and the results, we evaluate the generalization of our model on other multimodal datasets for coherence relation detection.

2. Related work
The study of image-text relations has been well-known as a multimodal problem and numerous research in recent years have studied the multimodal information retrieval task. Otto C et. a. [3] focus on the semantic relations between image and text. In order to model the image-text complex relation, they propose an additional evaluation metric named Status to classify the image-text relation into eight semantic classes. In their work, they developed a deep learning-based framework for semantic classes prediction to not only bridge the semantic gap in image-text pairs, but also form a better connection between the linguistics domain and the communication science domain. Using the unsupervised approach of discovering the relations between image-text pairs is another approach to bridge the gap. Jack Hessel proposed to utilize the unsupervised models to identify multimodal interdocument links without accessing to supervision at the image-text level during model training time. Their results showed that the number of unsupervised image-text links can grow as the increase length of sentence and the sentences may have multiple corresponding images at the document level.

The coherence theory becomes essentially important in multimodal communication, but it has been left behind. Alikhani et. al. [12] study the discourse coherence in recipe dataset. In their results, most corresponding texts fall into the instructional discourse category and a more general coherence relations are required. We assume that even there are multiple links that could exist within image-text pairs at the document level, but only one image-text pair has solid links. The modelling coherence in images and texts can contribute image-text pair convey the information by improving the text generation for associating images, especially the image caption generation.

3. Problem statement
Give a set of documents \( D = \{ d_1, d_2, \ldots, d_n \} \), where \( D = \{ S_1, S_2, \ldots, S_m, I_1, I_2, \ldots, I_k \} \), here, each document \( d_i \) contains multiple sentences and images. The number of sentences and images could be different. We aim to link the sentences and images in order to form image-sentences pairs such as \( D = \{ (S_1, I_1), (S_1, I_2), \ldots, (S_m, I_k) \} \), where \( S \) represents a group of sentences or single sentence. Then, we use the discourse coherence model to predict the coherence relation into six categories: Visible, Subjective, Action, Story, Meta, Irrelevant followed by the definition in literature [8]. The Figure 1 illustrates the image-text links and discourse coherence discovery in 6 categories: **Visible**: texts are able to depict the
scene of the image, similar to the relations restatement. **Subjective**: texts reflect the state’s reaction, evaluation, and response of the image, similar to relation evaluation. **Action**: texts are the descriptions of the extended, dynamic motion that is captured by the image, similar to relation Elaboration. **Story**: texts give a free-standing description of certain events and circumstances in the images. **Meta**: readers of the texts can draw inferences about the scene in the image and production of the image, similar to meta-talk relations. **Irreverent**: texts are not in any of the above categories.

## 4. Proposed approaches

We developed a GRU and DenseNet-169 (Figure 2) for sentences and image based deep learning model for sentences and images encoding, then the full connected layer for creating image-text internal features that aim to fuse the information from texts and images. The final soft-max layer outputs the coherence relation class, the word embedding utilizes the pretrained Bert base, uncased model as the input of sentences.

![Figure 1. The discovery of the image-text links and discourse coherence.](image1.png)

![Figure 2. The GRU and DenseNet-169 based network for image-text pair encoding and coherence relation detection.](image2.png)

### Dataset

Clue is a collection of 10,000 well-labeled coherence relation image-caption pairs. The first 5,000 image-caption pairs are randomly selected from the Conceptual Captions dataset [13], which contains over 111,000 web pages related to news, articles, advertisements, social media, and other areas. The second 5,000 image-caption pairs are the outputs of machine generated captions from top 5 models in the task of image captioning at the 2019 Conference on Computer Vision and Pattern Recognition (CVPR) [14]. This dataset is also publicly available. The discourse coherence relation label is followed by inter-annotator agreement protocol studied by Alikhani [8].

### RQA RecipeQA is a well-known and comprehensive multimodal data for cooking recipes. It consists of approximate 20,000 step-by-step instructions and images. We treated each cooking step as the document associated with the images, which represents that the image associated with textual information has multiple sentences. The most important aspects of the RecipeQA are the data resources from real natural language found online and contributed by normal people in an open environment. It can commonly be used for questioning and answering, but we are only interested in the step-by-step cooking part.

### Evaluation

We first convert the image-text alignment links as a bipartite matching problem, where for a given set of images and a set of texts are treated as nodes and then chose a set of ages in such a way that no two edges have a shared endpoint. The bipartite matching commonly applied on the job applicant chosen problem. In the multi-node graph space, we utilize the cosine similarity measurement of the image features and textual features, its calculation formula is as follows:

\[
\cos(I, S) = \frac{IS}{||I|| ||S||} = \frac{\sum_{i=1}^{n} I_i S_i}{\sqrt{\sum_{i=1}^{n} (I_i)^2} \sqrt{\sum_{i=1}^{n} (S_i)^2}}
\]  

(1)
Here $I, S$ are the image and sentences features encoded by our network. Then we adopt the 0-1 linear programming in as a Top-k assignment program, which eliminates the effect of non-visual sentences to being aligned to no image [15]. This allows the existence of situation of image has no sentence and sentence has no images. The objective function is defined as follows:

$$\max \sum_{i,j} x_{ij} M_{ij}$$  \hspace{1cm} (2)

$$\forall j \sum_{i} x_{ij} \leq K; \forall i \sum_{j} x_{ij} \leq K$$  \hspace{1cm} (3)

Where $x_{ij}$ is the solution of the linear programing problem for each image-text pair and is a matrix that captures the similarities between image-sentences pairs.

5. Result and discussion

To study the modalities information in the multimodal datasets, we proposed several baseline models: text-only, image-only, text and image, multimodal text and image. The multimodal text and image model first extracts the feature embeddings from textual and visual contents, then fuse the information from textural and visual to create multi-embeddings, and then the fully connected layer for coherence relation classification. We split out train/test/validation in the size of 8:1:1 with respect to the size of the Clue dataset. We first evaluate the multimodal approach that we are mostly interested in regarding Clue dataset. In Table 1, we exam the effectiveness of 6 categories in Clue dataset and obtain that the visible has the highest 0.84 F1-score. This also indicates that visible, subjective, action, and meta can be well linked using our multimodal model fusion. Since the story and irrelevant categories are challenging to ground compared to visible or action, their performances have a reasonable drop. This finding demonstrates that our multimodal can significantly capture the image and sentence semantics.

Table 2 shows our four experimental models. The Text-only, Image-only, and Textimage models serve as strong baseline compared to our proposed multimodal mode. Regarding the visible, story, and meta, the multimodal has significantly outperformed all the baselines, while the subjective and action also gain considerable results on the Clue dataset. To exam the generalization of the proposed model, we first trained a multimodal model on the receipeQ&A dataset, then we applied learned parameters to further finetune and test on the Clue dataset. Figure 3 shows the performance of 6 category predictions, of which the visible substantially outperforms the rest categories, which illustrates that our proposed model has strong generalization in capturing visible category.

| Class      | Precision | Recall | F1-score |
|------------|-----------|--------|----------|
| visible    | 0.82      | 0.87   | 0.84     |
| subjective | 0.80      | 0.87   | 0.83     |
| action     | 0.87      | 0.80   | 0.83     |
| story      | 0.76      | 0.77   | 0.76     |
| meta       | 0.79      | 0.83   | 0.81     |
| irrelevant | 0.69      | 0.67   | 0.68     |
To visualize the image-text links on the discourse coherence, we compute image-text similarity matrix, having the links with highest weights as the most confident. Figure 4 shows example story coherence predictions tested on Clue. It is interesting to note that the sentence *the eye of the beholder* has strong links to an image depicting a bedroom with a *light*, since the *light* in somewhat serves as an *eye* of the bedroom in the linked image. In addition, we also visualized the image-text links on the ReceipeQ&A dataset shown in Figure 5. From the observation, all the image-text links have strong confidence in predictions. The first three images and texts discuss about the preparation of the mango jalapeno jam, while the last two are in the process of cooking. Apparently, the first three images and texts and the last two have few cross-links, which demonstrates the effectiveness of the multimodal since the preparation and the cooking procedures should have different semantics.
Figure 5. The predicted image-text links tested on ReceipeQ&A dataset. All the discovered links have strong confidence in predictions.

Table 2. The F-1 score on the Clue dataset for the baselines and the proposed models

|                     | visible | subjective | action | story | meta | irrelevant |
|---------------------|---------|------------|--------|-------|------|------------|
| Text Only           | 0.83    | 0.67       | 0.78   | 0.57  | 0.77 | 0.72       |
| Image Only          | 0.58    | 0.59       | 0.66   | 0.55  | 0.54 | 0.69       |
| Text-Image          | 0.82    | 0.87       | 0.87   | 0.75  | 0.77 | 0.66       |
| Multimodal          | 0.84    | 0.83       | 0.83   | 0.76  | 0.87 | 0.68       |

6. Result and discussion
We explored an unsupervised way of detecting the links between image-text pairs on the document level through multi-modal data resources. Without a determinant approach of only one image links to sentences or one sentence links to one image, we found that one image or sentence is able to have multiple links with sentences or images. We found that explicit image-text links can discover a rich image-text relation. To solidify the information that the image-text pair tries to convey to the readers, we then investigate the discourse coherence relation for descriptive textual information, especially for the image-caption task. The coherence aware model can contribute image caption generation to be more efficient in the way of improving readers’ experience. We mainly conducted experiment results on Clue and RecipeQA datasets. From our experiments, the visible, subjective, action, and meta coherence links can be well discovered. In comparison with three strong baselines, our proposed model successfully outperforms in the visible, story and meta categories. In RecipeQA, we automatically classify the links of image-text pairs mainly to have visible coherence relation since the major cooking images focus on the objects such as food, cooking tools, intermediate food preparation, etc.

In our further work, we would like to explore more specific image embedding techniques, such as object detection, instead of embedding the whole images. On the other hand, we plan to further apply our model on more general datasets, such as story-DII, MSCOCO.

Acknowledgments
This research is supported by Hunan Natural Science Foundation(2019JJ40145), Scientific Research Key Project of Hunan Education Department (19A273).

References
[1] He K, Zhang X, Ren S and Sun J 2016 Deep residual learning for image recognition Proc. Int. Conf. on Computer Vision and Pattern Recognition (Las Vegas) (IEEE) p 770-8
[2] Anderson P, He X, Buehler C, Teney D, Johnson M, Gould S and Zhang L 2018 Bottom-up and
top-down attention for image captioning and visual question answering

Proc. Int. Conf. on Computer Vision and Pattern Recognition (Salt Lake) (IEEE) p 6077-86

[3] Otto C, Springstein M, Anand A and Ewerth R 2019 Understanding, categorizing and predicting semantic image-text relations Proc. Int. Conf. on Multimedia Retrieval (Ottawa) (ACM) p 168-176

[4] Henning CA and Ewerth R 2017 Estimating the information gap between textual and visual representations. Proc. Int. Conf. on multimedia retrieval (Bucharest) (ACM) p 14–22

[5] Henning CA, Ewerth R 2018 Estimating the information gap between textual and visual representations. J. LIMIR 7(1) 43–56

[6] Jerry R H 1978 Why is discourse coherent SRI International report (Menlo Park)

[7] Phillips B 1977 A calculus of cohesion Fourth LACUS Forum (Montreal Canada)

[8] Malihe A, Piyush S, Li S, Radu S and Matthew S 2020 Clue: Cross-modal Coherence Modeling for Caption Generation Preprint arXiv:2005.00908

[9] Narasimhan K and Barzilay R 2015 Machine comprehension with discourse relations Proc. 53th annual meeting of the Association for Computational Linguistics and the7th International Joint Conference on Natural Language Processing (Beijing) vol 1 (ACL) p 1253-62

[10] Xu J, Gan Z, Cheng Y and Liu J. 2020. Discourse-aware neural extractive text summarization Proc. 58th annual meeting of the Association for Computational Linguistics (Seattle) (ACL) p 5021-31

[11] Wang F, Wu Y and Qiu L 2012 Exploiting discourse relations for sentiment analysis Proc. of COLING (Mumbai) (ICCL) p 1311–20

[12] Alikhani M, Chowdhury S N, Melo G D and Stone M 2019 Cite: A corpus of image-text discourse relations. Proc. Conf. of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Minneapolis) vol 1 (ACL) p 570–575

[13] Piyush S, Ding N, Goodman S and Soricut R 2018 Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning Proc. 56th Annual Meeting of the Association for Computational Linguistics (Melbourne) vol 1 (ACL) p 2556-65

[14] Levinboim T, Thapliyal A, Sharma P and Soricut R 2019 Quality estimation for image captions based on large-scale human evaluations Preprint arXiv:1909.03396

[15] Jack H, Lee L and David M 2019 Unsupervised discovery of multimodal links in multi-image, multi-sentence documents Proc. Int. Conf. on empirical methodsin natural language processing (Hong Kong) (ACL)