Reconciling Image Captioning and User's Comments for Urban Tourism

Yazid Bounab
Faculty of ITEE, CMVS
University of Oulu
yazid.bounab@oulu.fi

Mourad Oussalah
Faculty of ITEE, CMVS
University of Oulu
Oulu, Finland
mourad.oussalah@oulu.fi

Ahlam Ferdenache
Chadli Bendjedid University, LASE
Badji Mokhtar University, Algeria
ferdenache.ahlam@gmail.com

***Abstract*—Image captioning as a process of assigning textual description to an image has gained momentum nowadays thanks to recent advances in deep learning related architectures and availability of associated tools. In the era of digital tourism, this offers a valuable framework to reconcile the widely available tourism images and user's generated content. This paper presents a novel approach to perform such reconciliation in a way to benefit tourism industry. Especially, IMGUR online image sharing services has been employed to construct a novel database, referred Tourism48, which contains gallery tourism images from 48 countries together with their associated user's comments. Google Cloud Vision API has been employed to perform image captioning of the underlined images, while similarity analysis has been employed to match user's comments to the obtained captioning results. The outcomes can trigger the development of subsequent policy research in tourism industry and behavior analysis. Besides, the Tourism48 dataset has been made available for research community.** [1]

***Index Terms*—Image captioning, social media, image description, string matching, google vision API.**

I. Introduction

In the era of social media, we witness the proliferation of open access structured and unstructured image and textual content generated by either random users or professional organizations. Such data becomes a valuable source for users who seek experience of their peers before deciding on any purchasing action. For instance, the rapid development of Trip advisor sites in tourism industry where customers pay more attention to their peer comments over professional advice demonstrates the usefulness and the importance of such user generated content. Similarly, the availability of such data provides a valuable source of information for professionals and researchers to perform analytics research and develop marketing strategy accordingly [1], [2]. Nevertheless, the unstructured nature of many of the available repositories containing such visual contents (images/videos), including blogs and online databases, renders the automatic exploitation of such resources rather difficult. The recent advances in automatic image captioning [3] provide enormous boost to utilize such raw materials in subsequent analytics modules. More specifically, image captioning as a process of assigning textual description to raw images, where a descriptive sentence or a set of keywords are generated for each image, has seen a growing interest in image processing community, especially with the advances in deep learning technologies and the multiplicity of open source toolkits in the field. Indeed, deep learning models have demonstrated their ability to achieve state-of-art and optimal results in many caption generation problems [3], [4]. In particular, deep image captioning architectures have shown excellent results in discovering the mapping between visual descriptors and words [5]. The authors in [5] employed Convolutional Neural Networks (CNNs) to extract an image representation and a Recurrent Neural Networks (RNNs) to build the corresponding sentence. Although, the progress in such techniques is encouraging, the human ability in the construction and formulation of a sentence is still far from being adequately emulated in todays image captioning systems as pointed out in [6]. Nevertheless, this should not undermine the usefulness of the generated keywords identified by state-of-art captioning systems. Although, this mainly questions only the lexical and semantic structure of the generated sentence. On the other hand, the availability of user generated textual content that comments, ranks or labels the posted visual content offers nice opportunities to leverage the two data types (visual and textual) that can boost subsequent data analytics task. This paper aims to tackle this aspect of analysis in tourism case study. Especially, the paper contributes to this field by providing two main novelties. First, a new Tourism48 database that gathers both image gallery collected from IMGUR online sharing images pertaining to 48 distinct countries and textual content is put forward. It has 480 images with 37,360 different comments and 4542 labels. The latter is constituted of both user generated contents as part of users' comments and perception on each image of the gallery as well as textual information issued from image captioning system. The latter is performed by the Google Cloud Vision API [7]. Second, a methodology for evaluating the extent of overlapping between user's generated content and visual content. This makes use of the well established Fuzzy string matching between collection set of users's comment of a given image and the Google Could Vision API output of that particular image. Section 2 highlights related work in the field. The methodology is detailed in Section 3 while database construction and proposed

[1]Tourism48:$https://www.kaggle.com/yazidbounab/tourism48-for-image-captioning-and-nlp-process$

methodology with preliminary results are described in Section 3 and Section 4, respectively.

II. Related work

A. image captioning and textual-visual matching

Related work associated to the proposal advocated in this paper can be split into two parts: image captioning and the matching process between the visual and the textual content. For the former, one distinguishes three streams of research in this field:

1) ***Template-based approaches:*** These approaches utilize the concept of fixed templates with several empty slots to produce captions, where they first identify different objects, attributes and actions and then, they fill the empty spaces in the templates. For image captions, [8] uses a triplet of scene elements to fill the template slots. Similarly, [9] extracts phrases related to detected objects, attributes and their relationships. A Conditional Random Field (CRF) is used by [10] to infer the objects, attributes, and prepositions before filling in the gaps. Especially, template-based approaches can produce grammatically correct captions. However, predefined templates can generate captions with a fixed length [4].
2) ***Retrieval-based approaches:*** In these approaches, captions are retrieved from existing caption training set by looking for visual similarities between target image and images in training set with their captions (candidate captions). The captions of any query image in test set are then assigned matched captions from the pool [11]. However, the main weakness of those approaches is that they cannot infer any new caption outside those in training set, which restricts the semantically correct captions [4].
3) ***Novel caption generation:*** This approach is based on both visual and multi-modal spaces. First, it extracts the visual content of an image then passes it to some pretrained language model to generate caption [12], [13]. This approach uses the relation between detected entities in an image through the semantic captured by the language model between entities. These methods can generate new captions for each image that are semantically more accurate than previous approaches. Most of the novel caption generation methods use deep machine learning based techniques [3], [4].

Usually, captions are generated for a whole scene in the image or different regions of an image (dense captioning). However, it can also generate a description for unseen objects. Deep learning-based image captioning methods can use either simple Encoder-Decoder architecture or compositional architecture [14], [15]. Some of these methods use attention mechanism, semantic concept, and different styles in image descriptions. Many of the image captioning methods use Long short-term memory (LSTM) as a language model [16]. However, several methods use other language models such as convolutional neural network (CNN, or ConvNet) and recurrent neural network (RNN) [3], [4]. Regarding the issue of matching visual content to textual content, possibly a close work is pointed out by Hexiang et al. [17] where a new dataset for visual and textual evaluation has been put forward. Although, their approach cast the matching problem from information retrieval perspective, it provides a nice setting for subsequent evaluation and scrutinizing analysis. In contrast, our approach does not confine to information retrieval based approach. Besides, the handling of social media like information raises extra challenges of unstructured and highly noisy data that require care in pre-and post-processing analysis.

Ultimately linked to system evaluation as well as methodological development, especially for constructing comprehensive training database, we provide below some of the popular dataset in the field. These dataset differ in many perspectives such as the number of images, the number of captions per image, the format of the captions, and image size. Finally, ones highlights below some relevant dataset for the purpose of image captioning research.

1) ***Flickr8k dataset:*** [18] has 8000 images collected from Flickr. The training data consists of 6000 images, the test and development data, each consists of 1,000 images. Each image in the dataset has 5 reference captions annotated by humans. A number of image captioning methods [19], [20] have been tested using the dataset.

2) ***Flickr30k dataset:*** [21] is a dataset for automatic image description and grounded language understanding. It contains 30k images collected from Flickr with 158k captions provided by human annotators. It does not provide any fixed split of images for training, testing and validation, which are then open for researchers to perform as they wish. The dataset also contains detectors for common objects, a color classifier, and a bias towards selecting larger objects. Image captioning methods such as [6], [22]–[25] used this dataset.

3) ***MS COCO dataset:*** Microsoft COCO dataset [26] is a very large dataset for image recognition, segmentation, and captioning tasks. There are more than 300,000 images with 2 million instances. It has 80 different object categories and 5 captions per image. Many image captioning methods [27]–[31] use this dataset in their experiments.

Especially, the above review showed the limitation of the existing dataset, particularly in the context of rural tourism in a way to benefit tourism professional industry. This motivates the development of new database in the field.

III. Dataset preparation

In order to explore the mapping between visual characters of an image to its surrounding text (comments/replies, title), a new database is constructed. For this purpose, we have

chosen the **IMGUR** web service, an online image sharing community and image host founded by Alan Schaaf, as a starting point to make the database. This is motivated by the API availability and the abundance of tourism galleries of sites' images / videos. The collected dataset contains data from 48 different countries. For each country, we selected 10 galleries. Each gallery contains a single image together with a set of users' comments and replies regarding the attractiveness /content of such picture as a destination target for urban tourism customers. We named this dataset **Tourism48**. See Figure 1 and Figure 2 for a conceptual description of the database construction.

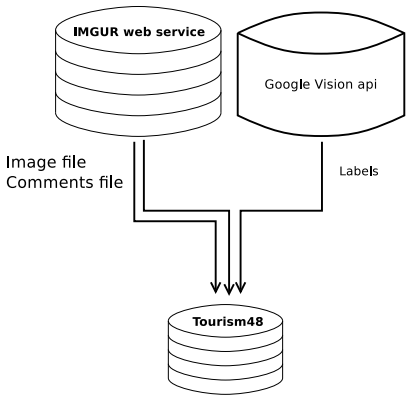

Fig. 1. Tourism48 data-set.

To use the IMGUR api, we need to get a key to download both the images and all the users comments/replies. Then we save this collected data into two separate files **gallery_id.jpg** that represent the image and **Comments.json**.

Finally, we feed each downloaded image to Google Cloud Vision Service to get what we call visual characters of that image (annotations/labels) and save them in a third file **google.json** as highlighted in Figure 2.

Fig. 2. structure of gallery.

An example of the database content and data structure is shown in Figure 3

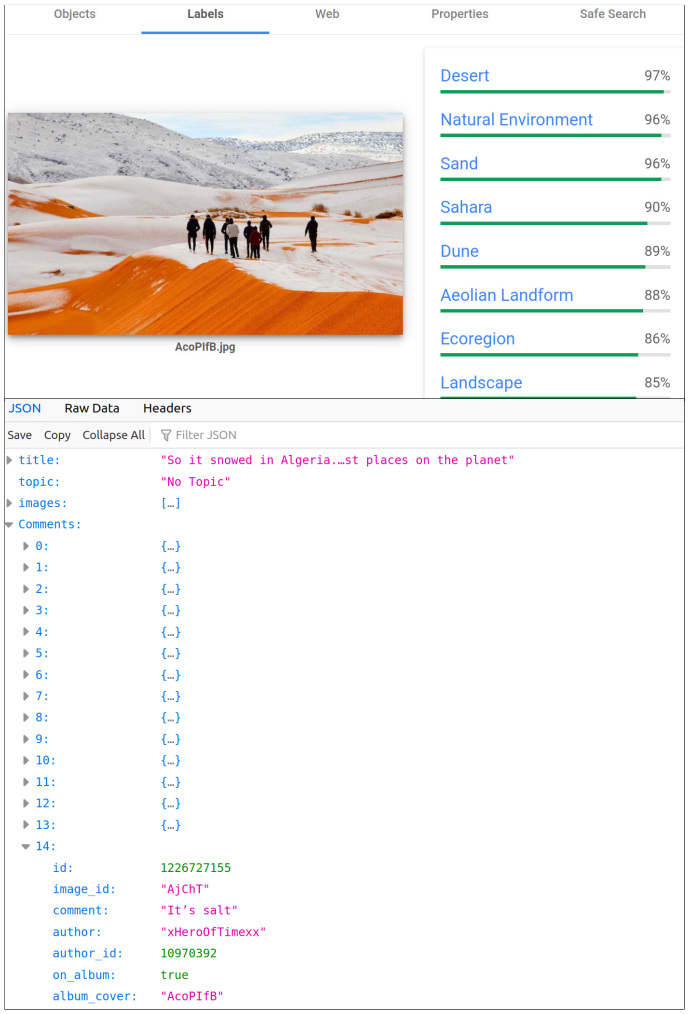

Fig. 3. Tourism48 dataset gallery sample

Next, the similarity between the user's generated content associated to a given and the image annotations generated by the Google Vision API is evaluated and quantified. This process is described in the next section.

IV. Proposed Method

A. Pre-processing

The high level generic pipeline of the data processing architecture is described in Figure 4. First, we download the image and the corresponding comments/replies. Second, for each image of the gallery, we take the whole comments as one single textual document, Then, we apply the following preprocessing stage. This consists of i) Remove URLs; ii) Remove emails; iii) Remove emojis; iv) Remove punctuation; v) Lowercase all characters; vi) Remove stopwords; vii) Remove swear expressions; ix) Remove extra-spaces; x) True case Annotation. This processed aims to filter out the vague amount of noise available in user's generated content and ensure, to some extent, a clean text.

Next, we apply named entity recognition (NER) on our cleaned text in order to determine the main relevant parts

that could exist within the labels of an image. However, we have excluded some of the entities that represent organisations, numbers, dates and periods. Finally, we end up with a distinct set of named entities from all user's comments which will be referred as set *A*. The motivation for doing so is that the labels outputted by Google Vision API are restricted to named entities that are identified in the image. Therefore, there is no need to use the whole syntax structure of the user's generated content in the matching process. Indeed, Google Cloud Vision API classifies images into thousands of categories, detects individual objects and faces within images. It has different output entities (labels, web, properties, Safe Search, JSON) [7]. The JSON format contains all these mentioned entities. We are only interested in labels and web entities, which they have the structure of a dictionary, where the keys are the labels and the values are the confidence scores. Thereafter, we merge both dictionaries to create one single dictionary that will be used in the matching phase after we apply the following preprocessing on each label of the dictionary: i) Lowercase; ii) Remove stopwords.

We will use only the list of keys that represent the distinct labels from the cleaned dictionary, referred as set *B*.
To measure the similarity between labels and entities sets, first, we need to compute the number of similar matched elements from both sets using some approximate string matching technique. For this purpose, the FuzzyWuzzy algorithm is employed. The use of FuzzyWuzzy enables us to account for minor spelling errors and allow for some flexibility in this respect. This is described below.

B. Approximate string matching:

The approximate string matching or fuzzy string matching is a technique for finding strings that match a pattern approximately (rather than exactly). This aims to match a string with a pattern approximately. Typically, the approach requires to set a threshold beyond which the two sets are deemed similar. More formally, let **M** be the approximate match between pattern **P** and text **T**. This is expressed as:

$$M(P,T) = \begin{cases} 1 & if\ M \geq Threshold \\ 0 & otherwise \end{cases}$$

In essence, FuzzyWuzzy algorithm uses Levenshtein distance to check weather two sequences (text) are somehow equivalent with respect to a specific threshold. In our case, we have set the matching threshold to 0.75 (75%), which means that we have tolerated the marge of the error to be 25% in case we have some typos in the user's comments.

C. Overall approach

Algorithm IV-C shows the entire process of computing similarity between image labels and users comments.

Algorithm 1 Similarity(A,B)

$Overlap \leftarrow 0$;
for $a\ \ in\ \ A$ **do**
  for $b\ \ in\ \ B$ **do**
    if $FuzzyWuzzy.ratio(a,b) \geq 0.75$ **then**
      $Overlap \leftarrow Overlap + 1$;
    end if
  end for
end for
$Sim \leftarrow Overlap/|A|$;
return Sim;

Next, we divide the number of matched elements by the number of the labels. Equation 1 shows the percentage of overlap between image labels and extracted entities from users comments.

$$Sim(Labels, Entities) = \frac{\mid Labels \bigcap Entities \mid}{\mid Labels \mid} \quad (1)$$

One notices from the preceding that we have not used the commonly employed Jaccard similarity formula entirely. Indeed, instead of the cardinality of the union of the two sets, we took only the cardinality of the label set. The reason behind that is when we collected the dataset we observed a lot of comments that have no relationship with the posted image in terms of mentioning entities. So if the original Jaccard similarity formula was used, it would yield a quite small value which indicates that there is no relationship at all between image labels and users comments.

Figure 4 shows the summary of our system.

D. Preliminary analysis results

Figure 5 summarizes the findings and contributions made in the experiment and identifies clear support and evidence of matching between user's generated textual inputs and visual content.

From Figure 5, it is clear that there are variations in terms of overlap values between the image labels and the entities extracted from comments that partly motivate the use of a modification of Jaccard similarity formula. Intuitively, discussion on social media can be wide and diverse, and often users pay more attention to other users' replies rather than the content of the image itself. This also provides another source of information bias when attempting to reconcile visual content to the available visual information. Nevertheless, the compilation of all users' comments and the use of set-based representation instead of a vector-based representation trivially contribute to the emergence of such overlapping. Indeed, the use of vector representation would yield a large sparse matrix with little to zero overlapping. Similarly, the compilation of all user's comments highlights a generic perception of the underlined image by the community, which likely includes a set of image descriptors and main patterns that will increase the matching with image captioning system. On the other hands, one may also mention other factors that do influence the users' textual inputs related to personal and sociological

factors [32], [33]. Among these factors, which are found to impact the structure of the comment, and thereby, the overlap value, one shall mention: the user culture, the age and gender of the user, te travel experience of the user, the mood of the user while he was writing his comment, the nature of image itself weather it is know or unknown (famous status). In other words, the topics carried by comments not only express what is in the image itself but it often conveys a deeper details that the image captioning system will never capture. Although, detailed analysis requires more statistical investigation to quantify the correlation of such factors with the current mapping result, which is left for further studies.

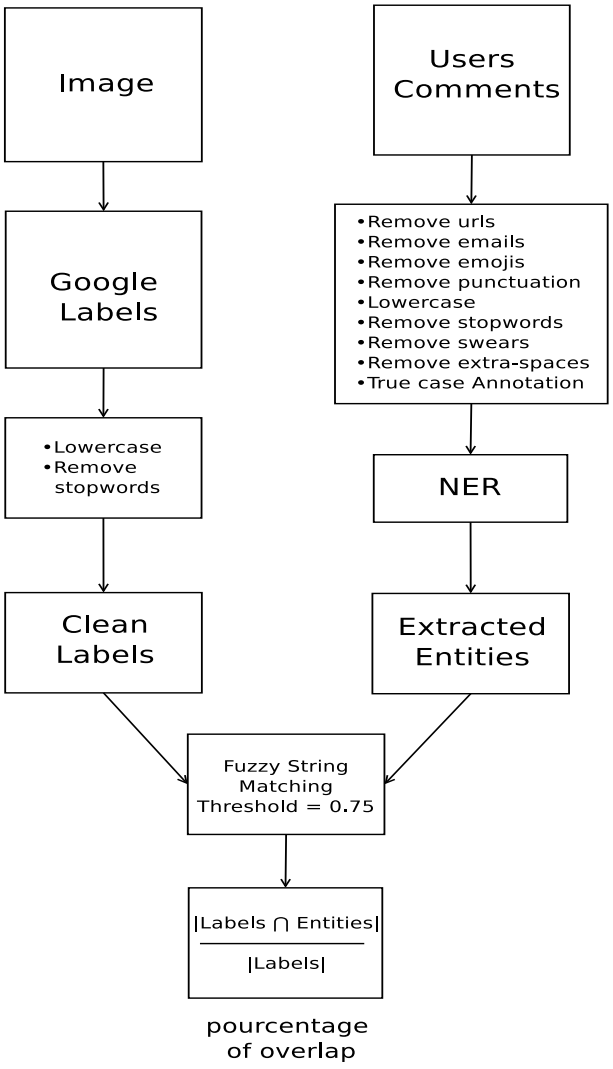

Fig. 4. Pipeline of our system

Figure 5 summarize our findings

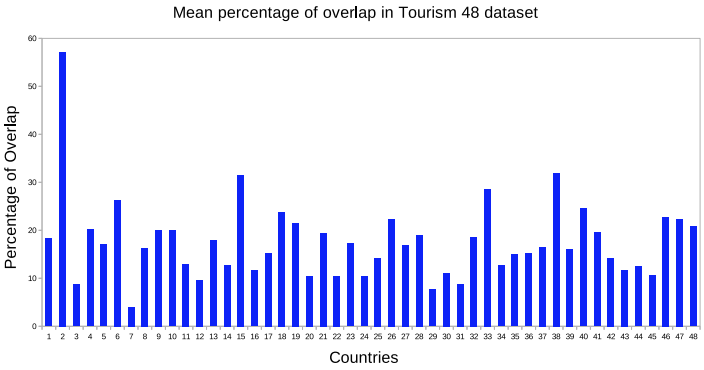

Fig. 5. Mean overlap percentage of Tourism 48 dataset countries

V. Conclusion

This paper investigates the issue of reconciling visual content and users' textual inputs in the context of urban tourism. In this respect, techniques from automatic image captioning have been employed to turn raw images into descriptive sentences and a set of keywords that best describe the content of the image. For this purpose, Google Cloud Vision API was utilized. A new database Tourism48 is made available for the research community. The database is made of three main attributes: i) raw images selected from IMGUR online image sharing services in 48 countries where for each country, 10 galleries of images are selected; ii) the outcomes of the Google Vision API in terms of generated image captioning for each image; iii) the third attribute bears the users' comments and replies associated to each individual image. The matching between the image captioning outcomes is carried out as part of natural language processing pipeline, which involves pre-processing stage, approximate string matching using Fuzzy-Wuzzy algorithm and a modified Jaccard's similarity index. Preliminary results indicates the feasibility of such approach and the possibility to carry out a large scale analytical analysis in order to derive policy change, comprehend users' behavior and provide useful insights to tourism professionals that guide future planning and actions. On the other hand, the authors concluded that when users engage with their comments, they help to provide different information from different angles of the same image in a very abstract way, which can, in turn, improve the textual description made by image captioning system. This opens up new directions for enhancing existing current captioning systems as well by using inputs from users' generated content.

Acknowledgment

This work is partly supported by the CBC KA 4033 on Business Creation and EU YoungRes #823701 projects, which are gratefully acknowledged.

References

[1] S. Hasani, M. Bahaghighat, and M. Mirfatahia, "The mediating effect of the brand on the relationship between social network marketing and consumer behavior," *Acta Technica Napocensis*, vol. 60, no. 2, pp. 1–6, 2019.

[2] M. A. Shareef, B. Mukerji, Y. K. Dwivedi, N. P. Rana, and R. Islam, "Social media marketing: Comparative effect of advertisement sources," *Journal of Retailing and Consumer Services*, vol. 46, pp. 58–69, 2019.

[3] S. Kalra and A. Leekha, "Survey of convolutional neural networks for image captioning," *Journal of Information and Optimization Sciences*, vol. 41, no. 1, pp. 239–260, 2020.

[4] M. Z. Hossain, F. Sohel, M. F. Shiratuddin, and H. Laga, "A comprehensive survey of deep learning for image captioning," *ACM Computing Surveys (CSUR)*, vol. 51, no. 6, pp. 1–36, 2019.

[5] Z. M. Griffin and K. Bock, "What the eyes say about speaking," *Psychological science*, vol. 11, no. 4, pp. 274–279, 2000.

[6] A. Karpathy and L. Fei-Fei, "Deep visual-semantic alignments for generating image descriptions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3128–3137, 2015.

[7] Google, "Web service, cloud computing," 2008.

[8] A. Farhadi, M. Hejrati, M. A. Sadeghi, P. Young, C. Rashtchian, J. Hockenmaier, and D. Forsyth, "Every picture tells a story: Generating sentences from images," in *European conference on computer vision*, pp. 15–29, Springer, 2010.

[9] S. Li, G. Kulkarni, T. L. Berg, A. C. Berg, and Y. Choi, "Composing simple image descriptions using web-scale n-grams," in *Proceedings of the Fifteenth Conference on Computational Natural Language Learning*, pp. 220–228, Association for Computational Linguistics, 2011.

[10] G. Kulkarni, V. Premraj, S. Dhar, S. Li, Y. Choi, A. C. Berg, and T. L. Berg, "Baby talk: Understanding and generating image descriptions proceedings of the 24th cvpr," *Citeseer. Google Scholar Google Scholar Digital Library Digital Library*, 2011.

[11] Y. Gong, L. Wang, M. Hodosh, J. Hockenmaier, and S. Lazebnik, "Improving image-sentence embeddings using large weakly annotated photo collections," in *European conference on computer vision*, pp. 529–545, Springer, 2014.

[12] R. Kiros, R. Salakhutdinov, and R. S. Zemel, "Unifying visual-semantic embeddings with multimodal neural language models," *arXiv preprint arXiv:1411.2539*, 2014.

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

[14] T. Yao, Y. Pan, Y. Li, and T. Mei, "Exploring visual relationship for image captioning," in *Proceedings of the European conference on computer vision (ECCV)*, pp. 684–699, 2018.

[15] L. Huang, W. Wang, J. Chen, and X.-Y. Wei, "Attention on attention for image captioning," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 4634–4643, 2019.

[16] L. Yang, H. Hu, S. Xing, and X. Lu, "Constrained lstm and residual attention for image captioning," *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, vol. 16, no. 3, pp. 1–18, 2020.

[17] H. Hex, I. Hu, L. Misra, and van-der Maaten, "Evaluating text-to-image matching using binary image selection," *arXiv preprint arxiv:1901.06595*, 2019.

[18] M. Hodosh, P. Young, and J. Hockenmaier, "Framing image description as a ranking task: Data, models and evaluation metrics," *Journal of Artificial Intelligence Research*, vol. 47, pp. 853–899, 2013.

[19] L. Chen, H. Zhang, J. Xiao, L. Nie, J. Shao, W. Liu, and T.-S. Chua, "Sca-cnn: Spatial and channel-wise attention in convolutional networks for image captioning," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5659–5667, 2017.

[20] X. Jia, E. Gavves, B. Fernando, and T. Tuytelaars, "Guiding the long-short term memory model for image caption generation," in *Proceedings of the IEEE international conference on computer vision*, pp. 2407–2415, 2015.

[21] B. A. Plummer, L. Wang, C. M. Cervantes, J. C. Caicedo, J. Hockenmaier, and S. Lazebnik, "Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models," in *Proceedings of the IEEE international conference on computer vision*, pp. 2641–2649, 2015.

[22] T.-H. Chen, Y.-H. Liao, C.-Y. Chuang, W.-T. Hsu, J. Fu, and M. Sun, "Show, adapt and tell: Adversarial training of cross-domain image captioner," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 521–530, 2017.

[23] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: A neural image caption generator," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3156–3164, 2015.

[24] C. Wang, H. Yang, C. Bartz, and C. Meinel, "Image captioning with deep bidirectional lstms," in *Proceedings of the 24th ACM international conference on Multimedia*, pp. 988–997, 2016.

[25] Q. Wu, C. Shen, P. Wang, A. Dick, and A. van den Hengel, "Image captioning and visual question answering based on attributes and external knowledge," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 6, pp. 1367–1381, 2018.

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

[27] B. Dai, S. Fidler, R. Urtasun, and D. Lin, "Towards diverse and natural image descriptions via a conditional gan," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2970–2979, 2017.

[28] Q. You, H. Jin, Z. Wang, C. Fang, and J. Luo, "Image captioning with semantic attention," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4651–4659, 2016.

[29] C. Gan, Z. Gan, X. He, J. Gao, and L. Deng, "Stylenet: Generating attractive visual captions with styles," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3137–3146, 2017.

[30] J. Jin, K. Fu, R. Cui, F. Sha, and C. Zhang, "Aligning where to see and what to tell: image caption with region-based attention and scene factorization," *arXiv preprint arXiv:1506.06272*, 2015.

[31] Z. Yang, Y. Yuan, Y. Wu, W. W. Cohen, and R. R. Salakhutdinov, "Review networks for caption generation," in *Advances in neural information processing systems*, pp. 2361–2369, 2016.

[32] X.-L. Shen, Y.-J. Li, Y. Sun, Z. Chen, and F. Wang, "Understanding the role of technology attractiveness in promoting social commerce engagement: Moderating effect of personal interest," *Information & Management*, vol. 56, no. 2, pp. 294–305, 2019.

[33] H. Yin, Q. Wang, K. Zheng, Z. Li, J. Yang, and X. Zhou, "Social influence-based group representation learning for group recommendation," in *2019 IEEE 35th International Conference on Data Engineering (ICDE)*, pp. 566–577, IEEE, 2019.