An Ensemble Approach Towards Correlating Articles and their Corresponding Images

Sudhanva Rajesh[1,†], Ashwath Krishnan[1,†] and Bhaskarjyoti Das[1,*,†]

[1]PES University, Bengaluru, India

Abstract

This paper presents a novel approach for the ranking of images for news articles as part of the the "NewsImages: Relating news articles and images" task, in order to better understand the relationship between the textual and visual content of news articles. The proposed approach combines the entity relationship and the contextual similarity between text and image by summarising both modalities into text annotations. The text annotations are generated from the news articles using Named Entity Recognition and Part of Speech tagging. Image annotations comprise of objects and labels generated from the image using the Google Vision API. The text and image annotations were further expanded by generating synsets and enriched using the wikipedia API. The analysis of the results has been carried out on the training dataset provided, and the five different results generated using five configurations of the proposed approach are compared.

1. Introduction

News articles are associated with a visual representation, and the relationship between the articles and the Images is complicated. With the image being important to understand the context of the article, the MediaEval challenge 2022 organized NewsImages task[1] to investigate these intricacies in more depth. The training dataset provided consists of two parts, a collection of images and a set of news articles. Each article has a corresponding image, and the goal of this task was to retrieve the images based on the semantic relationship between them, and derive insights on the same. Our approach aimed at deriving the semantic gap or distance between the two modalities. We generate annotations for the news articles and the images and analyze them. There are five kinds of relationships[2] observed:

1. **Instance of itself:** This is a case of pure synonyms observed in the text and image annotations.
2. **Member of relation:** This is the case where the image has an entity which belongs to a class suggested by the text article or vice versa.
3. **Part of relation:** This is the case where the image has an entity which is part of a larger entity suggested by the text article or vice versa.
4. **Semantically related:** This is the case where the image has an entity that is closely related to an entity present in the text.
5. **Closely Related:** This is the case where the image entity and text entity are not only closely related, but are also similar in meaning.

MediaEval'22: Multimedia Evaluation Workshop, January 13–15, 2023, Bergen, Norway and Online

*Corresponding author.

†These authors contributed equally.

sudhanva0001@gmail.com (S. Rajesh); ashwathkrishnan45@gmail.com (A. Krishnan); Bhaskarjyoti01@gmail.com (B. Das)

© 2022 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR Workshop Proceedings (CEUR-WS.org)

Our proposed approach aims at covering these semantic gaps, for retrieving the corresponding images for each of the text articles.

2. Related Work

Chee Wee Leong and Rada Mihalcea[2] presented a methodology for quantifying cross-modal semantic links between random pairings of words and visuals. The efficiency of a corpus-based technique for automatically determining semantic relatedness between words and images was investigated, and empirical assessments were carried out by comparing its correlation with human annotators. The semantic relatedness of words and images was assessed by developing a shared semantic vector space using visual code words and textual words.

Eric Müller-Budack et al.[3] provided a multimodal method for quantifying entity coherence between picture and text in real-world news. Named entity linking was used in their suggested approach to extract people, places, and events from news texts. Using cutting-edge computer vision techniques, the authors used numerous metrics to calculate the cross-modal similarity of things in text and pictures.

To recommend photos for news items, Pontus Svensson[4] presented a retrieval technique based on canonical correlation. The effects of several dense text representations created by Word2vec and Doc2vec, as well as picture representations produced by pre-trained convolutional neural networks were investigated. Word2Vec outperformed Doc2Vec in the task, indicating that the meaning of article texts was not as significant as the individual words that comprised them.

Nelleke Oostdijk et al.[5] in their analysis of 1000 news articles with images proved the inadequacy of simplistic correlation between modalities i.e. apart from visually describing text, image can describe the entity inside the text, can describe the discordance between past, present or present and future.

3. Approach

With respect to the images, the Google image annotator is used to generate captions for an image. Post this, the Wikipedia API is employed to improve the semantics of these captions and to expand the context of the image. Named entity recognition is first performed on the article to extract entities such as places, objects etc. Once this has been done, we proceed to generate WordNet synsets, hyponyms and hypernyms for each of the extracted entities. The synset of a word is a group of data elements that are considered semantically equivalent for the purposes of information retrieval. A hyponym is a word of more specific meaning than a general or superordinate term applicable to it and a hypernym is a word with a broad meaning constituting a category into which words with more specific meanings fall. The Wikipedia API is then used to generate annotations for the synsets to improve the semantics of the article. To rank the images for a given article, the following two approaches are followed:

1. **WordNet distance (D1):** The Wu-Palmer Similarity is used to calculate the WordNet distance. The WuP similarity returns a score denoting how similar two word senses are, based on the depth of the two senses in the taxonomy and that of their Least Common Subsumer (most specific ancestor node).
2. **BERT Similarity (D2):** The extracted entities for the articles and the images are embedded into a vector of 1x768. Since a given article/image has multiple entities, the mean of these embeddings are calculated to find the embedding for an image and an article.

Cosine similarity is then calculated between an article and every image, based on which the images are ranked. Higher the cosine similarity, higher the rank of that image for an article.

4. Results and Discussion

To rank and retrieve the top 'K' images for every article, five different approaches were tested out to find the optimum ranking system. For testing, 100 articles from the training set were used to evaluate different approaches. The weighted sum was calculated to rank the images, with varying weights for each of the below approaches. The formula to calculate the similarity is as follows:

$$Score = \alpha * D1 + (1 - \alpha) * D2$$

1. **WordNet Distance only:** In this approach, $\alpha = 1$ i.e, only the Wu-Palmer similarity was considered for ranking the images for each of the news articles. This approach retrieved the corresponding image for the news articles within the top-5 results for 40% of the articles.
2. **BERT Similarity only:** In this approach, $\alpha = 0$ i.e, only the Cosine similarity between the embeddings generated using S-BERT for the image annotations and the text annotations was considered for ranking the images for each of the news articles. This approach retrieved the corresponding image for the news articles within the top-5 results for 80% of the articles.
3. **Equal weights for WordNet Distance and BERT Similarity:** In this approach, the weighted sum Wu-Palmer similarity and the cosine similarity was considered for ranking the images. In this case, $\alpha = 0.5$, which implied that the weightage towards both the similarity measures were the same. This approach retrieved the corresponding image for the news articles within the top-5 results for 60% of the articles.
4. **Increased weightage for WordNet Distance:** In this approach, the weighted sum Wu-Palmer similarity and the cosine similarity was considered for ranking the images. In this case, $\alpha = 0.8$, which implied that the weightage towards the Wu-palmer similarity was higher. This approach retrieved the corresponding image for the news articles within the top-5 results 40% of the time.
5. **Increased weightage for BERT Similarity:** In this approach, the weighted sum Wu-Palmer similarity and the cosine similarity was considered for ranking the images. In this case, $\alpha = 0.2$, which implied that the weightage towards the Cosine similarity was higher. This approach retrieved the corresponding image for the news articles within the top-5 results 60% of the time.

Table 1 discusses the results obtained on the test set containing 1500 articles, with the top 100 images ranked for each article.

Table 2 compares the five approaches listed in the previous section and cites examples for which each approach works best. Overall, the best results are obtained when both the WordNet Distance and BERT similarity are considered, with a higher weightage assigned to the latter for ranking images but this can be due to the nature of the dataset.

Since our aim is to determine the semantic gap or distance between the modalities, A vector of both modalities and correlation analysis would not be a good approach. For example, when given an article that talks about "Lockdown due to COVID-19" and an image that depicts an empty city with abandoned road, only considering the BERT similarity based ranking would fail to identify the relation. Such kinds of news articles where there are multiple interpretations

Table 1
Quantitative comparison of the five approaches

Approach	Matches	Mean Reciprocal Rank	Mean Recall @100
WordNet Distance only	114	0.00422	0.07600
BERT Similarity only	128	0.00892	0.08533
Equal weights for WordNet Distance and BERT Similarity	124	0.00662	0.08267
Increased weightage for WordNet Distance	123	0.00505	0.08200
Increased weightage for BERT Similarity	135	0.00940	0.09000

Table 2
Comparison of the five approaches

Approach	Idea of Article	Image	Type of Relation	Explanation
WordNet Distance only	EU-China relationship	The flags of EU and China	Member of Relation, Part of Relation	In this case the image consists of "flags" which is an entity that belongs to the "country" class present in text
BERT Similarity only	Lack of water	Desert	Semantically Related	A desert is associated with a lack of water making them semantically related
Equal weights for WordNet Distance and BERT Similarity	Gun violence	Armoury	Closely Related	The entity 'Gun' synonymous with gun violence and an armoury
Increased weightage for WordNet Distance	Deforestation	Tree	Closely Related, Member of Relation, Part of Relation	The entity 'Tree' is part of a larger entity i.e, 'Forest', both of which are closely related to deforestation
Increased weightage for BERT Similarity	Opening ceremony of a building	Group of people cutting a ribbon	Closely Related, Semantically Related	Cutting of a ribbon is associated with opening ceremony making the two entities closely and semantically related

of the article, and a direct correlation between the article and the image is not possible are the hardest to find the linked images. In such cases the discordance between the modalities convey the semantics[5] of the implied message and will be hardest to distinguish from pairs of completely unrelated images and texts. It is this category that should generate maximum false negative or false positive in any automated approach. The category of news for which it is easy to find linked images are the ones where the articles contain objects and a sufficient number of named entities which could also be present in an image. For instance, the training dataset consisted of an article that discussed the relations between EU and China, while the image depicted the two flags of the same. This image was easily retrieved using all five approaches. Out of the five different kinds of relationships[2] observed between the text and image annotations, the WordNet distance based similarity is better suited for Instance-of-Self, Member-of-Relation, and Part-of-Relation, since the WordNet synsets, hyponyms and hypernyms correlate directly with the above mentioned classes. BERT similarity using SBERT[6] on the other hand is better suited for Semantically-Related and Closely-Related text and image annotations.

References

[1] B. Kille, A. Lommatzsch, Özgöbek, M. Elahi, D.-T. Dang-Nguyen, News images in mediaeval 2022, 2022.

[2] C. W. Leong, R. Mihalcea, Measuring the semantic relatedness between words and images, in: Proceedings of the Ninth International Conference on Computational Semantics (IWCS 2011), 2011.

[3] E. Müller-Budack, J. Theiner, S. Diering, M. Idahl, S. Hakimov, R. Ewerth, Multimodal news analytics using measures of cross-modal entity and context consistency, International Journal of Multimedia Information Retrieval 10 (2021) 111–125.

[4] P. Svensson, Automated image suggestions for news articles: An evaluation of text and image representations in an image retrieval system, 2020.

[5] N. Oostdijk, H. v. Halteren, E. Basar, M. A. Larson, The connection between the text and images of news articles: New insights for multimedia analysis (2020).

[6] N. Reimers, I. Gurevych, Sentence-bert: Sentence embeddings using siamese bert-networks, arXiv preprint arXiv:1908.10084 (2019).