Advances and Applications in Mathematical Sciences
Volume 21, Issue 6, April 2022, Pages 3315-3327

CROSS MODAL INFORMATION RETRIEVAL USING MULTILAYER BIDIRECTIONAL LSTM MODEL

SHAILY MALIK and POONAM BANSAL

Research Scholar
University School of Information
Communication and Technology
GGSIPU, New Delhi, India-110058
E-mail: shaily.singh99@gmail.com

Department of Computer Science and Engineering
Maharaja Surajmal Institute of Technology
GGSIPU, New Delhi, India-110058

Abstract

Cross-modal information retrieval employs methodology for building a common representation domain from different heterogeneous modalities, like text, audio, images, videos etc. The task gets particularly challenging for images and their captions because of multiplicity of correspondences. Inter-modality semantic correlations accentuate the understanding of how the data from different modalities can be mapped into a common latent semantic space which is also known as the heterogeneity gap. In this research we propose a cross modal retrieval system that leverages on image and text encoding to overcome this heterogeneity gap. For capturing of semantic relationship, many multimodal architecture employ different networks for every modality. In the proposed work we have used multilayer Bidirectional Long Short Term Memory Model (BiLSTM) to map the semantic representations into a common space. These Semantic representations can be easily retrieved from image and text modalities using the convolution deep neural network. We have made a cross-modal retrieval system for image and text data which responds for the image and text queries. The feature extracted from text and image modalities are mapped into a common latent semantic space by passing through the networks and similarity is measured in terms of cosine similarity measure. We have trained the model using Adam swap advancement calculation for stochastic slope plunge and Mathematical categorical cross entropy as the loss function. However, in our proposed work we are achieving comparable results in terms of cross modal retrieval with just single network for each modality by using image-text encodings. It is the power of deep learning models to map the image-text encodings on the common space. For this research we have evaluated our system on Flickr8k

2020 Mathematics Subject Classification: 68T07.
Keywords: Cross-modal retrieval, deep learning, Semantic Similarity, LSTM.
Received September 20, 2021; Accepted January 15, 2022.

dataset because it has well defined captions associated with each image. The result of the given query will be from the top K results which are retrieved from the dot product of the query vector performed with each vector from the common representation space. We have compared our multilayer BiLSTM model with the conventional LSTM and Bi-LSTM and our model outperformed both of them.

1. Introduction

The aim of cross-modal retrieval system is to retrieve relevant samples from different heterogeneous modalities, which is important in numerous multi-modal applications. The major challenge in implementation of Cross-Modal Retrieval is modality gap and the solution is to generate new representation domain from different modalities [1]. Information from various modalities might bless semantic connections, support cross-modular recovery that profits applicable consequences of one methodology in light of the inquiry of another methodology, e.g., recovery of image with text question and the other way around. An advantageous solution to cross modal retrieval is mapping inputs from multiple modalities to a space (called common representation space), which can then be utilized to retrieve results for an input from different modalities.

A few ongoing kinds of researches in the field of profound learning have taken extraordinary steps and have helped Artificial Intelligence to outperform human execution. The profound design fuses low-level elements into undeniable level non-visual provisions with non-direct change, permitting it to have the option to take in semantic representation from images [2]. Because of such headways in profound learning and handling power, there have been a developing number of explores in this field. With the new leap forwards that have been going on in information science, it is discovered that for practically these grouping expectation issues, Long Short-Term Memory organizations, or LSTMs have been seen as the best arrangement.

The problems which needs short term memories in order to identify a solution can use LSTM's as they are designed with different gates (Input gate, cell gate, output gate and forget gate) that help in regulating the flow of information. These gates are capable of learning that which information in a succession is essential to keep or to discard. Then it becomes easier to make

predictions on the basis of the pertinent information. The customary LSTM anyway accompanies a few weaknesses as a result of which there exist two varieties of LSTM, in particular multilayer LSTM, and BiLSTM. On account of multilayer LSTM (or stacked LSTM), there are different LSTM layers with intermittent connections in the units of the same layer, and feed-forward connection in the units of an LSTM layer and then layer above it. The multiple layers stacking structure of LSTM takes into account more noteworthy model intricacy. This arrangement is quite similar to the stacked layered architecture used in a simple feed forward neural network. Bidirectional LSTM have an essential plan to introduce each preparation succession advances and in reverse to two LSTM once from starting as far as possible and once from end to starting. By preparing the contributions from both directions, it uses both the past and future context for a longer range. For this work, we will join the force of these two varieties with the end goal of cross-modular information retrieval.

The section 2 of the paper discuss a variety of practices used in the area of cross modal information retrieval systems, Section 3 describes about our proposed methodology and the procedure to implement it while section 4 provides the summary of the results of our experiments and Section 5 concludes our work and provide insights of the future work which can be carried out in this research area.

2. Related Work

Cross modal Machine Learning models have progressed in such a way that they are capable for heterogeneous data analysis for multiple tasks such as scene construction using virtual reality principles, indexing of the images, image search, detection of emotions, collaborative learning, learning of cross modal representations of content-rich data types, task atheism, representations of image linguistic, visual dialog, learning in incremental ways and even automated categorization of dark web sites which have multimedia content in the area of cyber security.

Most of the researchers have investigated the task of multimodal retrieval on the scope of text and image areas [3-6]. Some of them also worked on dependency tree relations for equating the pieces of an image

objects with phrases and represented the fragments into a latent space which is common for both the modalities [7]. Huang et al. proposed a smLSTM where they have used modulated global attention scheme to make use of a cross modal framework and calculate the salient instance pairs by using LSTM [8]. Lately, many researchers introduced a model based on neural networks for image descriptor retrieval which consists of Recurrent Neural Networks, Convolutional Neural Networks, and additional layers for multimodal retrieval [9-15].

The main reason that contributes to the demanding growth of different deep learning approaches is the wide range availability of abundant data and information pool on the Web. A Modality-Consistent Embedding Network (MCEN) which focuses on the modality-invariant representations by protruding images and texts to the common embedding space for that is based on cross-modal retrieval for the data of food items images and different cooking recipes [16]. A deep multimodal convolutional recurrent network also has been introduced which is able to learn about match of the pairs (positive match) and mismatch of the pairs (negative match) by using a triplet ranking which is based on hinge. The model infer image-text similarity by representing both language and vision based simultaneous learning [17].

The Multi modal data retrieval framework can be used for both coupled and uncoupled samples. This framework is made up of two-part concept that focuses to present single modal depiction of high level with uncoupled samples. This alliance combines multiple modalities like text, image, audio etc through a few coupled training samples. This framework facilitates a cross modal retrieval method which is entirely based on the uniformity between the semantic structures of multiple modalities. To address each example's semantic comparability from the reference point, created from single-modular bunching, both the heterogeneous medium (pictures and message) are addressed with a semantic design based on normal portrayal area. Then, at that point the cross-modular similitude is estimated with the consistency between the semantic constructions [18]. A neural network architecture is used for cross modal retrieval when only one modality is present in the input and more than one modalities are present in the collection items. The proposed network design presents a cross breed LSTM CNN to portray the image modalities and to project the text modalities into a

typical subspace it uses skip-gram model, which contains embeddings of words in the text modalities and embeddings of words that depict the image modality. The proposed framework likewise incorporates a gating organization to change the data stream by considering idea level and point level coordinating with results.

3. Methodology

In the proposed work we have used multilayer BiLSTM Model to map the semantic embeddings into a common space. These Semantic representations can be easily retrieved from image and text modalities using the convolution deep neural network. We have made a cross-modal retrieval system for image and text data which responds for the image and text queries. The feature extracted from text and image modalities are mapped into a common latent semantic space by passing through the networks and similarity is measured in terms of cosine similarity measure. The Adam learning method is used for training of the network and categorical cross entropy has been used as the loss function.

3.1 Dataset Used. For this work we have used the Flickr8k dataset [20] which is easily available at the kaggle website. This dataset has 8000 distinguish image and 5 separate describing sentences are mapped with each image. The dataset is non specific because it also contains the images of unknown places and people. There are 6000 training images, 1000 development dataset images and 1000 test images. We have trained our model on these 6000 images which made our model usable for any other dataset used for retrieval purpose. To verify the same we have used our Flicker8K trained model to test the Wang's image dataset [21, 22]. This dataset comprises of 1000 images which are distributed over 10 classes having 100 images each.

3.2 Data Pre-processing. The transformation processes like cleaning, dimensionality reduction and transformation of raw data into significant and valuable format are applied on the dataset prior its use in the model. As discussed earlier, each image is defined by 5 descriptions associated with it, so these descriptors are cleaned properly before we fed them into the network. To do so we remove all the punctuations (coma, full stop etc) at the

very first place, then we convert the uppercase letters in the descriptors to lowercase for ensuring uniformity, then we append 'startseq' at the start of a sequence and 'endseq' flags to show the end of a sequence to LSTM.

3.3 Feature Extraction. Feature extraction is a technique which reduces the time and efforts required to process the image or text data without any loss of important and relevant information. This process discards the superfluous and ineffective information which results in reducing the analysis data amount by a significant ratio. Because of the above stated reason we use it in the area of dimensionality reduction. There are a wide variety of techniques are available to do Feature extraction [23]. For this work we have used deep convolution neural network (DCNN) VGG19 model to perform feature extraction on images. Then we will use these extracted image features to generate descriptions.

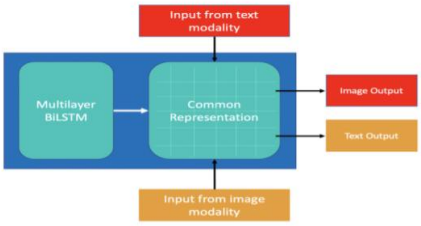


Figure 1. The proposed multilayer BiLSTM Architecture used for cross modal retrieval.

As shown in figure 1, we will used the extracted image features as the input from image. The VGG19 is a deep convolutional neural network consisting of 19 layers (16 convolution layers, 5 maxpool layers, 3 Fully connected layers and 1 SoftMax layer) [25]. The model has learned rich component portrayals for a wide scope of images.

The deep features of the image are acquired by the CNN part of VGG19 for which the pretrained image net weights are used. The extracted features dimension is 1×4096 after giving an input image of size 224×224. These features are fed to our used multilayer BiLSTM model. For text feature extraction, the Bag of Words (BoW) technique is used. First, we cleaned all the keywords by removing punctuations, digits and then all the unique words are extracted from the pool of keywords in the training dataset. The total

numbers of unique words are 8120 which represents the columns of BoW table. Now for every image in Wang dataset, there is one associated keyword representing the rows of BoW. For every keyword, we loop through the unique set of words and the respective column is being marked as 1 or 0 depending on whether the word is present in the keyword of not. The dimension of BoW table is 900 * 8120.

3.4 Generating Common Representation. For generating a common representation that describes the data, the output would be pre-processed in such a way it creates a mapping where each individual image/text data point in the dataset has its own set of descriptive vectors. When we receive an input for any modality (text, image) it is pre-processed to fit the representation space which our model created. Once this is achieved retrieval can be easily performed.

3.5 Training the model. As shown in figure 2, our architecture is working with two CNN networks simultaneously, one of them is for the textual features and another is for the images features. As discussed earlier our data set have 5 captions associated with each image, we created 5 distinguished sets of these captions and encapsulated the image in each one of them. Now, these sets are fed into the network. The image inside each set is gone through the pre-trained VGG19 model which gives us the profound features of the image. These extracted features are then gone through the additional layers. Close by the textual part of the set is cleaned, diminished, changed, and encoded, prior to giving to the additional layers. The organization dependable to prepare on the literary information likewise has an implanting layer which is instating by arbitrary loads for all words and later learns the embeddings for every one of the words in the dataset. The outcomes are then given to the multilayer BiLSTM model whose output is then joined with the output from the first model to get the preferred outcomes from the dense layers. We have trained the model using Adam swap advancement calculation for stochastic slope plunge and Mathematical categorical cross entropy as the loss function.

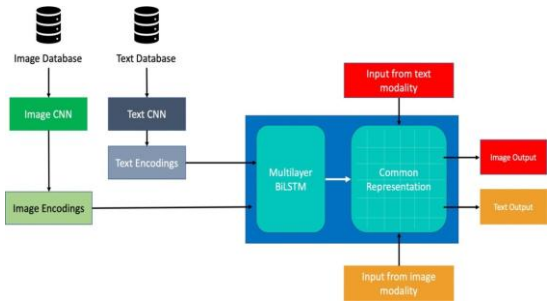


Figure 2. Multilayer BiLSTM Architecture used for generating a common representation space.

3.6 Retrieving Data. With this common representation space available for both modalities, we can now retrieve images and text modality outputs from it, corresponding to a query image from text or image modality respectively. The input shall follow a series of steps to obtain a form which can be mapped to the common representation space generated earlier during training. These are ideally vectors. A dot product of this vector is then performed with each vector from the common representation space and the images and text corresponding to the top K results is returned as the result set for the query.

3.7 Evaluation and Scoring. The precision and recall are the measures used to evaluate the performance of our cross modal retrieval system. Precision estimates the capacity of the framework to recover just applicable models while recall estimates the capacity of the framework to recover every significant model. In any model the precision tells us that what proportion of positive identifications were actually correct while recall tells us what proportion of actual positives was identified correctly. These are defined mathematically as:

$$\text{Precision} = \frac{\text{True Positive}}{\text{True Positive} + \text{false negative}} \tag{1}$$

$$\text{Precision} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Positive}} \tag{2}$$

To assess the effectiveness of a model, we must examine both recall and precision. Unfortunately, there is always trade-off between precision and recall. That is, increasing precision typically decreases recall and vice versa.

Hence, we might know that we want to maximize either recall or precision at the expense of the other metric. For our use case since a poor recall might not be a problem for the result set, but a poor precision would have. Thus, for our research we will target to maximize the precision for the outputs generates by our model.

4. Results and Discussion

Precision and Recall are the statistical and mathematical measures of the relevance of a retrieved image with respect to the query. For our work these are calculated as:

$$\text{Precision} = \frac{\text{Number of relevant images retrieved}}{\text{Number of total images retrieved}} \tag{4}$$

$$\text{Precision} = \frac{\text{Number of relevant images retrieved}}{\text{Number of total images Present}} \tag{5}$$

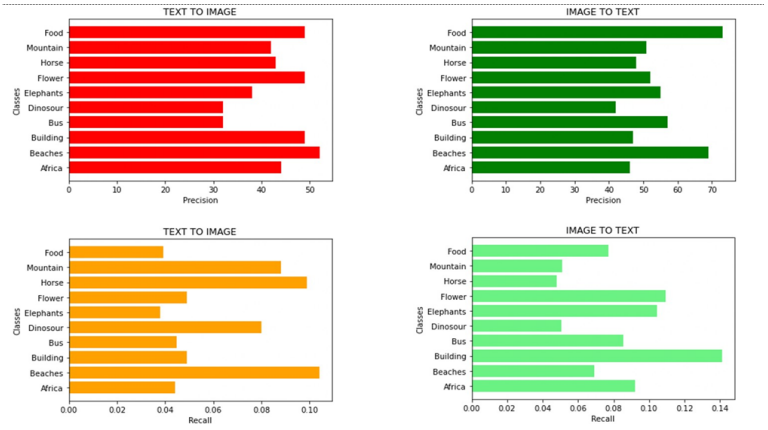
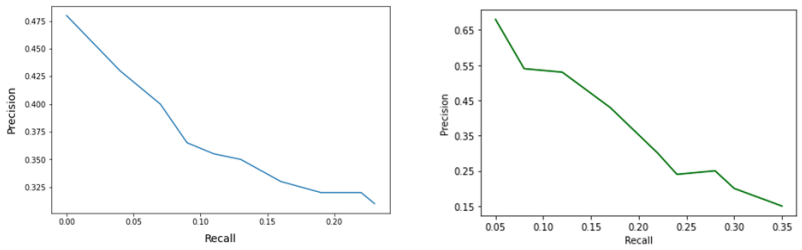


Figure 3. Class wise precision and recall values for $k = 10$ images retrieved per query.

Figure 3 show that when initially an image query is fed to the proposed architecture then we get multiple text captions for that query image. We have examined first 10 results for the given image query and calculated the class wise precision and recall values for those 10 text results corresponding to the input image. Similarly we have also used our model for the text query and examined 10 retrieved images and calculated the class wise precision and recall for these images corresponding to our input query.



(a) (b)

Figure 4. (a) Precision vs Recall graph for Text to Image Modality and 4(b) Precision vs Recall graph for image to Text Modality.

Figure 4 (a) and 4(b) shows the value of precision with its corresponding recall value. The curve lies almost on the main diagonal and hence this implies that the model performance is good in terms of precision and recall. However, Table 1 shows the precision calculated by the multilayer BiLSTM model as compared to the precision and recall values calculated by the Bi-LSTM and LSTM models.

Table 1. class wise precision of the proposed model with 10 image/text values per query.

Class Name	Precision by Multilayer BiLSTM		Precesion by BiLSTM		Precesion by LSTM	
	Image to Text	Text to Image	Image to Text	Text to Image	Image to Text	Text to Image
Africa	46	44	43	43	46	44
Beaches	69	52	68	55	65	54
Building	47	49	48	44	45	47
Bus	57	32	54	32	60	33
Dinosaur	42	32	46	33	44	33
Elephants	55	38	49	40	49	38
Flowers	52	49	49	49	51	49

Horse	48	43	50	43	46	41
Mountain	51	42	49	42	49	42
Food	73	49	69	44	69	49
Average	54	43	52	42	52	43

5. Conclusions and Future Work

With many application benefits, cross-modular information retrieval has acquired a great deal of exploration consideration in research. This work has presented a system for cross-modular information retrieval. The significant aim of the proposed architecture was the utilization of a multilayer BiLSTM model to map textual data and image data to a common latent space which could further be used for the cross modal retrieval task. The model was effectively prepared to fill the need for image retrieval from an advanced data set of images utilizing image encodings created with the assistance of VGG19 model and their corresponding captions accessible as a part of the used dataset. To prove its utility we trained our model on Flicker 8K dataset and then used this pretrained model for testing the Wang's image dataset, along with a custom dataset prepared manually over Wang's image dataset and a decent class wise precision was obtained. The multi-facet BiLSTM model is a very intricate organization and could give much preferred outcomes for datasets bigger than the Flickr8k dataset. As a future work the proposed model for cross modal retrieval could also be extended to several modalities with the condition that a representation space to map all the selected modalities could be created. We have worked with the text and image modality which can be extended to sound and videos data also.

References

[1] S. Malik and P. Bansal, Semantic space autoencoder for cross modal data retrieval, In proc. of International Conference on Innovative Computing and Communication (ICICC) 21-23 (2020), 509-516.

[2] S. Malik and P. Bansal, Multimodal semantic analysis with regularized semantic autoencoder, Journal of Intelligent and Fussy Systems (JIFS), IOS Press, preprint 1-9. DOI: 10.3233/JIFS-189759.

[3] H. Wang, Y. Cai, Y. Zhang, H. Pan, W. Lv and H. Han, Deep learning for image

retrieval: what works and what doesn't, IEEE International Conference on Data Mining Workshop (ICDMW), Atlantic City, NJ (2015), 1576-1583.

[4] Wang, Qingzhong and Antoni B. Chan, CNN+CNN: Convolutional Decoders for Image Captioning, ArXiv abs/1805.09019, 2018.

[5] A. Karpathy and Li Fei-Fei, Deep visual-semantic alignments for generating image descriptions, IEEE Transactions on Pattern Analysis and Machine Intelligence 39 (2017), 664-676.

[6] Z. Niu, M. Zhou, L. Wang, X. Gao and G. Hua, Hierarchical multimodal lstm for dense visual-semantic embedding, IEEE International Conference on Computer Vision (ICCV) (2017), pp. 1899-1907.

[7] Y. Liu, Y. Guo, E. M. Bakker and M. S. Lew, Learning a Recurrent Residual Fusion Network for Multimodal Matching, IEEE International Conference on Computer Vision (ICCV) (2017), 4127-4136.

[8] Andrej Karpathy, Armand Joulin, and Li Fei-Fei, Deep fragment embeddings for bidirectional image sentence mapping. International Conference on Neural Information Processing Systems (NIPS'14). MIT Press, Cambridge, MA, USA 2 (2014), 1889-1897.

[9] Yan Huang, Wei Wang and Liang Wang, Instance-Aware Image and Sentence Matching with Selective Multimodal LSTM, IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2017), 7254-7262.

[10] Yan Huang, Qi Wu and Liang Wang, Learning Semantic Concepts and Order for Image and Sentence Matching, IEEE/CVF Conference on Computer Vision and Pattern Recognition (2018), 6163-6171.

[11] F. Yan and K. Mikolajczyk, Deep correlation for matching images and text, IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2015), 3441-3450.

[12] Zhedong Zheng, Liang Zheng, Michael Garrett, Yi Yang, Mingliang Xu and Yi-Dong Shen, Dual-path Convolutional Image-Text Embeddings with Instance Loss, ACM Transactions on Multimedia Computing, Communications, and Applications 16(2) (2020), 1-23.

[13] Jeff Donahue, Lisa Anne Handrix, Marcus Rohebach et al., Long-term recurrent convolutional networks for visual recognition and description, IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2015), 2625-2634.

[14] Guy Lev, Gil Sadeh, Benjamin Klein and Lior Wol, RNN Fisher Vectors for Action Recognition and Image Annotation, Computer Vision and Pattern Recognition, ECCV, 2016.

[15] Mao, Junhua et al., Deep captioning with multimodal recurrent neural networks (m-RNN), arXiv: Computer Vision and Pattern Recognition, 2015.

[16] J. Gu, J. Cai, S. Joty, L. Niu and G. Wang, Look, Imagine and Match: Improving Textual-Visual Cross-Modal Retrieval with Generative Models, 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (2018), 7181-7189.

[17] H. Fu, R. Wu, C. Liu and J. Sun, MCEN: Bridging Cross-Modal Gap between Cooking

Recipes and Dish Images with Latent Variable Model, IEEE/CVF Conference on Computer Vision and Pattern Recognition, USA (2020), 14558-14568.

[18] H. A. Khojasteh, E. Ansari, P. Razzaghi and A. Karimi, Deep Multimodal Image-Text Embeddings for Automatic Cross-Media Retrieval, ArXiv, abs/2002.10016, 2020.

[19] Qibin Zheng, Xiaoguang Ren, Yi Liu and Wei Qin, Abstraction and association: cross-modal retrieval based on consistency between semantic structures, Mathematical Problems in Engineering, Article no. 2503137, 2020.

[20] Saeid Balaneshin-kordan and Alexander Kotov, Deep Neural Architecture for Multi-Modal Retrieval based on Joint Embedding Space for Text and Images. ACM International Conference on Web Search and Data Mining (WSDM'18) (2018), 28-36.

[21] Micah Hodosh, Peter Young, and Julia Hockenmaier, Framing image description as a ranking task: data, models, and evaluation metrics. Journal of Artificial Intelligence Research 47 (2013), 853-899.

[22] Jia Li, James Z. Wang, Automatic linguistic indexing of pictures by a statistical modelling approach, IEEE Transactions on Pattern Analysis and Machine Intelligence, 25(9) (2003), 1075-1088.

[23] James Z. Wang, Jia Li and Gio Wiederhold, SIMPLIcity: Semantics-sensitive Integrated Matching for Picture LIbraries, IEEE Transactions on Pattern Analysis and Machine Intelli- gence 23(9) (2001), 947-963.

[24] M. Turkoglu, D. Hanbay and A. Sengur, Multi-model LSTM-based convolutional neural networks for detection of apple diseases and pests, Journal of Ambient Intelligence and Humanized Computing, 2019.

[25] K. Simonyan and A. Zisserman, Very Deep Convolutional Networks for Large-Scale Image Recognition, CoRR. 2015;abs/1409.1556.

[26] J. Deng, W. Dong, R. Socher, L. Li, Kai Li and Li Fei-Fei, Image Net: A large-scale hierarchical image database, 2009 IEEE Conference on Computer Vision and Pattern Recog- nition, Miami, FL, (2009), 248-255.