JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

A DEEP LEARNING APPROACH TO IMAGE-TEXT EMBEDDING

[1]V. Vishnu Vardhan, [2] **V. Veera Nagendra,** [3] *V.Vaishnavi,[4] V.Varshitha Reddy,[5] DR P Venkateswara Rao*
[1]Student, [2]Student, [3]Student,[4]Student,[5]Proffesor
AIML
Malla Reddy University,Hyderabad,India

Abstract : Image-text matching is an important task in computer vision and natural language processing that aims to find the semantic relationship between an image and its corresponding text description. One approach to image-text matching is deep learning, which has shown promising results in various computer vision and natural language processing tasks. The first step is to pre-process the image and text data to make them suitable for deep learning models. For images, this may involve resizing, cropping, and normalization. For text, this may involve tokenization, stop-word removal, and stemming. In image processing, this may involve using convolutional neural networks (CNNs) to extract features from images. In natural language processing, this may involve using pre-trained language models such as BERT or GPT to extract features from text.
Once the features are extracted, the next step is to create a common embedding space for both the image and text features. This can be done by training a deep learning model that learns to map both the image and text features into a shared embedding space. The final step is to measure the similarity between the embedded image and text features. This can be done using various similarity measures such as cosine similarity or Euclidean distance.

I. INTRODUCTION

In recent years, the amount of visual data available on the internet has increased dramatically, and it has become a challenging task to organize and search this data effectively. One approach to this problem is image-text matching, which involves matching an image with its corresponding textual description. Image-text matching has numerous applications, such as image search engines, image captioning, and visual question answering.

Deep learning has shown great promise in various computer vision and natural language processing tasks, including image-text matching. Deep learning models can extract high-level features from both images and text and learn a common representation of the two modalities, enabling accurate image-text matching.

OBJECTIVE OF THE PROJECT

- Collecting and pre-processing a dataset of images and their corresponding text descriptions.
- Training and evaluating a deep learning model for image-text matching using various architectures and techniques.
- Comparing the performance of the model with state-of-the-art methods in the field.
- Analyzing the model's performance and identifying areas for improvement.

EXISTING SYSTEM

Traditional machine learning-based methods usually involve feature engineering, where handcrafted features are extracted from both the image and text modalities and then combined using various similarity measures.

Some examples of traditional machine learning-based methods for image-text matching include bag-of-words models, visual-semantic embedding, and cross-modal hashing.

PROPOSED SYSTEM

Dual Path CNN (DP-CNN) is a specific architecture that can be utilized for image-text embedding. It involves the parallel processing of images and text using separate CNN pathways and then fusing their representations in a joint embedding space

RELEATED WORKS

There have been several works on deep learning approaches for image-text embedding using dual-path CNN architectures. Here are a few notable ones:
"Deep Visual-Semantic Alignments for Generating Image Descriptions" by Karpathy et al. (2015): This work introduced the concept of using a dual-path CNN architecture for image-text embedding. The image pathway used a pretrained CNN to extract visual features, and the text pathway utilized an LSTM to encode the textual descriptions. The fusion of image and text representations was performed using a multimodal embedding layer. The model was trained on a large dataset of image-caption pairs and demonstrated promising results in generating descriptive captions for images.

"Image-Text Cross-Modal Retrieval with Dual-Attention Multi-Layer Fusion Network" by Huang et al. (2019): This work proposed a dual-attention multi-layer fusion network for image-text cross-modal retrieval. The image pathway used a CNN to extract visual features, and the text pathway employed a combination of recurrent and convolutional layers to encode textual descriptions.
Dual-attention mechanisms were applied to capture fine-grained interactions between image and text modalities. The fusion of image and text features was performed through multi-layer fusion blocks. The model was evaluated on several benchmark datasets and achieved competitive performance in cross-modal retrieval tasks.

"Unifying Vision-Language Tasks via Text-Image Multi-Modal Transformers" by Lu et al. (2019): This work presented a unified framework for various vision-language tasks, including image-text embedding. The image pathway used a CNN to extract visual features, and the text pathway employed a transformer-based architecture to model textual descriptions. A cross-modal transformer was introduced to capture interactions between image and text representations. The model was trained on multiple vision-language datasets and demonstrated state-of-the-art performance in tasks such as image captioning, visual question answering, and image-text retrieval.

DATASET DESCRIPTIONS

In addition to object recognition, the COCO dataset also includes annotations for object segmentation, where each pixel of the object is labelled, and for image captioning, where each image is annotated with five natural language descriptions.

GoogleNews-vectors-negative300.bin is a pre-trained word embedding model created by Google. It was trained on a massive corpus of Google News articles, which contains over 100 billion words. The model contains 300-dimensional vectors for around 3 million words and phrases, where each vector represents the distributional similarity of the corresponding word with respect to other words in the vocabulary. The vectors were trained using the word2vec algorithm, specifically the skip-gram model with negative sampling.

DATA PRE-PROCESSING

- Image resizing
- Data augmentation
- Word2Vec Corpus

METHODS AND ALGORITHMS

Dual-Path Image Text Embedding (DPITE) is a neural network architecture that jointly learns image and text representations by combining convolutional neural networks (CNNs) of Image and Text by the following are some of the methods and algorithms used in DPITE:

CNNs for Image processing: The image input is processed by a CNN, which is designed to extract image features. In DPITE, the ResNet-101 CNN architecture is used for image feature extraction.

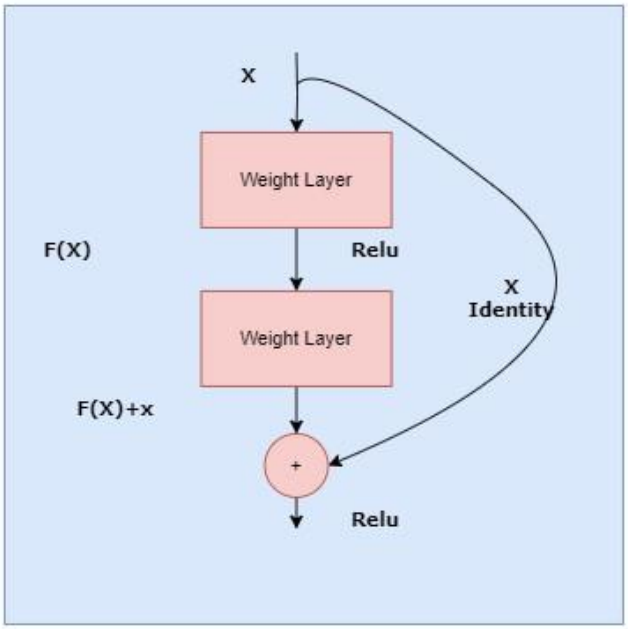

Figure-1: Resnet-101 for Image

CNNs with Embeddings: Another way to use CNNs for text processing is to use CNNs with word embeddings. Word embeddings are a type of dense vector representation that capture the semantic meaning of words. By using word embeddings as input to the CNN, the network can learn to extract relevant features from the text data that are important for the task at hand.

IMPLEMENTATION

Implementing a deep learning approach for image-text embedding with a dual-path CNN involves several steps. Here's a high-level outline of the implementation process

1.Data Preparation :

Obtain a dataset that consists of image-text pairs, where each pair is semantically related. This dataset can be created by collecting images and their corresponding textual descriptions or by using pre-existing dataset MSCOCO. Pre-process the images by resizing them to a consistent size and applying any necessary normalization or data augmentation techniques. Pre-process the textual descriptions by tokenizing the sentences, converting words to numerical representations (e.g., word embeddings), and padding the sequences to a fixed length.

2.Model Architecture:

Design the architecture of the dual-path CNN model. This typically involves defining the image pathway, text pathway, and fusion mechanism.

Choose appropriate CNN architectures for the image pathway, such as VGG16,ResNet, or Efficient Net, and pretrain them on a large image dataset like ImageNet if desired. Select suitable text encoding techniques for the text pathway, such as word embeddings like Word2Vec or GloVe, and recurrent or transformer-based models like LSTMs or Transformers. Decide on the fusion mechanism to combine the image and text representations. This can involve concatenation, element-wise multiplication, or more complex attention mechanisms.

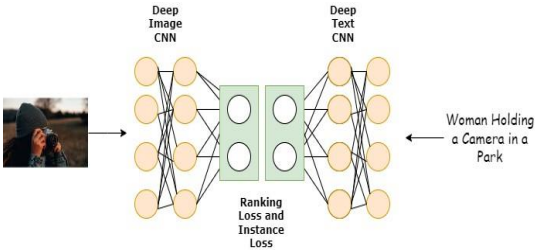

Figure-3:Embedding Image and Text CNN into Common Embedding Space

3. Model Training:

Split the dataset into training and validation sets.
Define a suitable loss function for training the model[12,11]. Common choices include contrastive loss or triplet loss, which encourage similar image-text pairs to be closer in the embedding space. Train the model using the training dataset. This involves feeding image-text pairs into the dual-path CNN model and optimizing the loss function using backpropagation and gradient descent. Monitor the model's performance on the validation set and adjust hyperparameters as needed, such as learning rate, batch size, or regularization techniques.

4. Loss Function

At Last We Use Cross Entropy Criterion as a Loss Function and We Update the Weights of the Network of the Weights for Every Epoch[12]

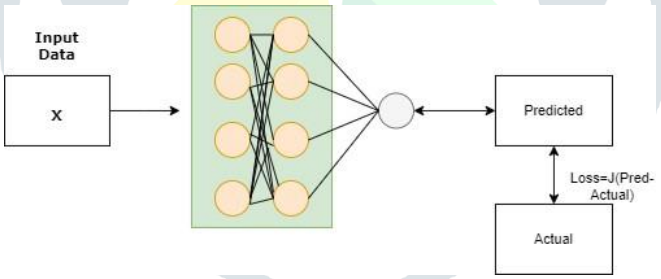

Figure : 4

RESULTS

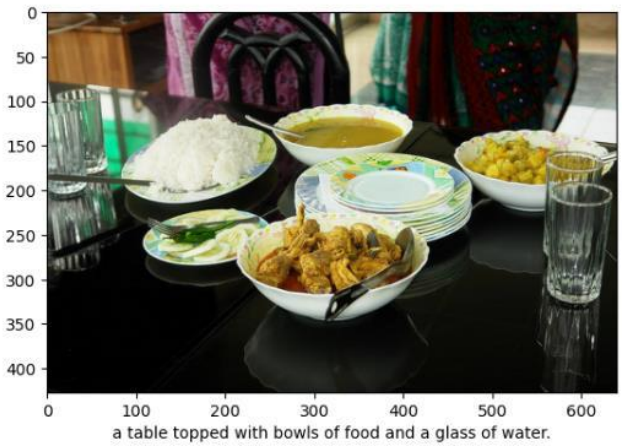

Figure-a

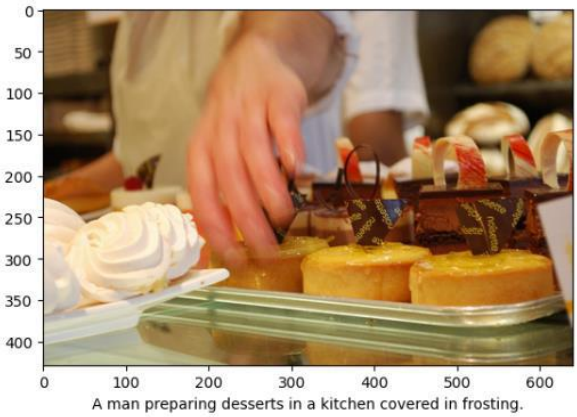

Figure-b

Figure-c

Figure-d
Figure- a, b, c, d represents the Results Created By the model.

CONCLUSION

In conclusion, image-text embedding with a deep learning approach using a dual-path CNN is a powerful technique for representing images and text in a shared latent space. By leveraging convolutional neural networks (CNNs) to process images and encoding textual descriptions with techniques like recurrent neural networks (RNNs) or transformers, this approach enables the creation of semantically meaningful embeddings for image-text pairs. The dual-path CNN architecture allows parallel processing of images and text, followed by fusion of their representations in a joint embedding space. The fusion can be achieved through concatenation, element-wise multiplication, attention mechanisms, or other fusion techniques. The implementation of image-text embedding with a dual-path CNN involves data preparation, designing the model architecture, training the model, evaluating its performance, and applying it to specific applications. It requires careful selection of pre-trained CNNs, text encoding techniques, fusion mechanisms, and optimization strategies. The field of image-text embedding continues to evolve, and researchers are constantly exploring new architectures and techniques to improve the performance and capabilities of these models. By leveraging the power of deep learning and the rich representation capabilities of CNNs, image-text embedding enables a wide range of multimodal applications that bridge the gap between visual and textual data.

REFERENCES

[1]"Deep Visual-Semantic Alignments for Generating Image Descriptions" by Karpathy et al. (2015)
[2] "Image-Text Cross-Modal Retrieval with Dual-Attention Multi-Layer Fusion Network" by Huang et al. (2019)
[3]"Unifying Vision-Language Tasks via Text-Image Multi-Modal Transformers" by Lu et al. (2019)
[4]"Dual Learning for Machine Translation" by He et al. (2016)
[5]"Learning Cross-Modal Embeddings for Cooking Recipes and Food Images" by Salvador et al. (2017)
[6]"VSE++: Improving Visual-Semantic Embeddings with Hard Negatives" by Faghri et al. (2018)
[7]"Show, Attend and Tell: Neural Image Caption Generation with Visual Attention" by Xu et al. (2015)
[8]"SCAN: Learning to Classify Images without Labels" by Reed et al. (2016)
[9]"Aligning Books and Movies: Towards Story-Like Visual Explanations by Watching Movies and Reading Books" by Kiros et al. (2015)
[10]"Learning Deep Representations of Fine-Grained Visual Descriptions" by Ma et al. (2016)
[11] Code Reference: https://github.com/pshroff04/Dual_Path_CNN
[12] Research Paper Reference:
http://arxiv.org/abs/1711.05535
[13] Data Set Reference
https://cocodataset.org/