IJFANS INTERNATIONAL JOURNAL OF FOOD AND NUTRITIONAL SCIENCES
ISSN PRINT 2319 1775 Online 2320 7876
 UGC CARE Listed (Group -I) Journal Volume 11, Iss 13, 2022

CROSS MODEL PRODUCT SEARCH SYSTEM

Mr. K.L.V.G.Krishna Murthy[1], Dr V V Nagaraju Goriparthi[2], Mrs R Pushpalatha[3]
[1,2]Department of CSE, Vignan's Lara Institute of Technology & Science, Guntur, AP, India.
[3]Department of EEE, Acharya Nagarjuna University, Guntur, AP, India
vngplkrishnamurthy@gmail.com[1], nagaraju006@gmail.com
rayalapushpalatha30@gmail.com

ABSTRACT:
The goal of multi-modal search is to obtain pertinent results from a database by utilizing a variety of modalities, including text and images. The multi-modal search aims to integrate several data kinds to deliver more thorough and accurate search results. This technique is widely used in e-commerce, healthcare, social media, and entertainment, among other fields. Machine learning methods, similarity metrics, feature extraction, and other approaches are needed for the multi-modal search. Due to the increasing availability of multi-modal data, multi-modal search has become a more important research topic in the domains of computer vision and information retrieval.

1. INTRODUCTION:

When a user queries a system, the multimodal search technique allows the system to provide results in many modalities—text, image, audio, and video—based on the query. In order to produce more precise and pertinent search results, it blends several modalities. An input query in a multimodal search system often consists of text and/or image. The technology gathers pertinent features from every modality and merges them into a joint feature representation. Next, based on the user's inquiry, the joint representation is employed to retrieve the pertinent results. An e-commerce website, for instance, allows users to search for products by entering queries, and the system returns results based both the product's image and written description.

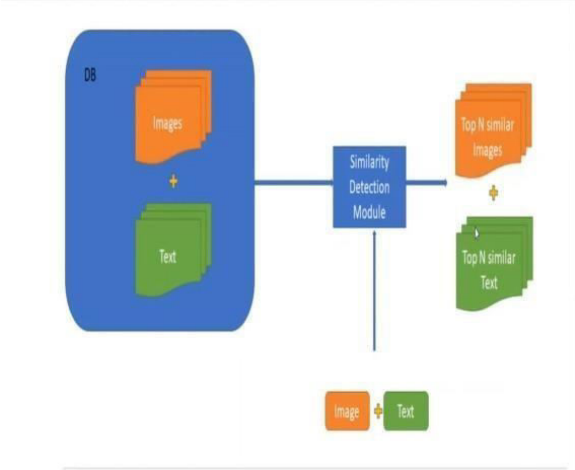


Fig 1(a) Architecture of the MultiModel Search

2. LITERATURE SURVEY:

The authors of the study "Learning Multi-modal Similarity" are Trevor Darrell, Marcus Rohrbach, Huazhe Xu, Ronghang Hu, Jiashi Feng, and Kate Saenko. In this paper, a technique for learning a similarity metric that can compare object representations in text and image is proposed. To learn a combined embedding space for the two modalities, the method combines cross-modal matching with triplet loss.

"Multi-Modal Search with Image and Text Queries" by Yushi Jing and Katja Hofmann. This paper proposes a multi-modal search

ISSN PRINT 2319 1775 Online 2320 7876
 UGC CARE Listed (Group -I) Journal Volume 11, Iss 13, 2022

system that allows users to search for images using text queries and vice versa. The system uses a combination of deep learning and information retrieval techniques to perform the search.

"Multi-Modal Deep Learning for Image and Text Recognition" by Hyunjung Shin and Dongsuk Yook. This study suggests a deep learning paradigm for text and picture recognition that mixes recurrent and convolutional neural networks.

3. PROPOSED SYSTEM:

1. The Basic Mechanism:

Identify the Text and Image Datasets: The first step in building a multi-model search system for text and images is to identify the datasets that will be included in the search. These could include databases of text documents, image repositories, or other sources of information.

Data Preprocessing: Once the datasets have been identified, the data needs to be preprocessed. This involves cleaning and standardizing the data so that it can be easily searched.

Text Embedding: The text data needs to be Embedded so that it can be searched quickly and efficiently. This involves creating the vectors for the text data that can be used to search for specific terms or phrases.

Image Embedding: The image data needs to be Embedded as well. In order to do this, features from the photos, such as colours, forms, and textures, must be extracted using computer vision algorithms. Afterwards, you may use these characteristics to look for related photographs.

Query Processing: When a user enters a search query, the system will process the query and search across both the text and image datasets. This involves using the text index to identify relevant text results and using the image features to identify relevant image results. The results can then be ranked based on relevance.

Machine Learning: Machine learning techniques can be used to increase the precision and applicability of search results. Natural language processing (NLP) models,for example, may be used to improve text search results, while image recognition models could be used to improve image search results. These models can be trained on vast amounts of data in order to identify relationships and patterns that can be utilized to improve search results.

3. 2. Working Mechanism:

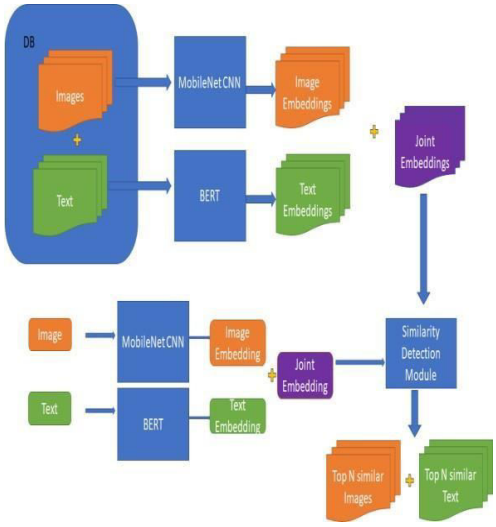


Fig 3.2(a) Working Mechanism Architecture

Here the Database contains both images and texts. Now we are using MobileNet CNN for the classification of Images and BERT for the Text Representation.

3. 2. 1. MobileNet CNN for Image Classification:

MobileNet is an architecture for convolutional neural networks (CNNs) designed to be efficient on embedded and mobile devices. Combining pointwise convolutions with depthwise separable convolutions allows for this.

ISSN PRINT 2319 1775 Online 2320 7876
 UGC CARE Listed (Group -I) Journal Volume 11, Iss 13, 2022

Depthwise separable convolutions split a standard convolution into two independent processes: a pointwise convolution that uses a 1x1 convolution to combine the output of the depthwise convolution, and a depthwise convolution that independently applies a single filter to each input channel.

By applying a distinct filter to each input channel instead of utilizing a single filter for all of the input channels, depthwise convolution lowers the number of parameters in the network. This can significantly lower the convolution's computational cost.

The pointwise convolution essentially performs a linear transformation on the result of the depthwise convolution by combining it with a 1x1 convolution. As a result, the network can learn more intricate correlations between the input and output, which increases its expressive capacity. Overall, the MobileNet design combines pointwise and depthwise separable convolutions to lower the amount of network parameters and boost embedded and mobile device performance.

Despite its efficacy, MobileNet is still able to achieve high accuracy on a range of image recognition tasks. **A typical convolutional neural network (CNN) for image classification consists of several layers, including:**

Input layer: The input image, which is commonly shown as a matrix of pixel values, is fed into this layer.

Convolutional layer: This layer applies several filters to the input image, each of which is capable of identifying specific features, such corners or edges. This layer produces a set of feature maps that indicate the locations in the image where each filter was applied.

Activation layer: Using the output of the convolutional layer and a non-linear activation function, this layer gives the model non-linearity and helps it learn more complex features.

Layer for pooling: By down sampling the feature maps, this layer lowers their spatial dimensionality, preventing overfitting and assisting in the model's parameter reduction.

Dropout layer: This layer helps prevent overfitting and improves generalisation by randomly removing a subset of the model's neurons during training.

Fully linked layer: In this layer, all of the neurons from the layer before and the layer following are connected, allowing complex non-linear correlations between the features and the output to be learned by the model.

Model's ultimate output, which is often a probability distribution over dataset classes, is produced by the output layer.

3. 2. 2. BERT for Text Classification: In 2018, Google developed a pre-trainedlanguage model dubbed Transformer-Based Bidirectional Encoder Representations, or BERT. It is comparable to a neural network design for natural language processing (NLP) that is transformer-based applications like text classification, question answering, andlanguage translation.

BERT is trained on both the left and right context of each word in a phrase, it is said to as bidirectional, in contrast to traditional NLP models that are taught on only one direction. This makes it possible for BERT to produce more accurate predictions and comprehend the context of a statement completely.

Massive amounts of text data are used to train a huge transformer-based neural network as

part of the BERT pre-training process, such as Wikipedia and other web sources. BERT can acquire a general knowledge of natural language and the relationships between words through this method, which can then be tailored for certain NLP tasks.

When using BERT for a particular NLP task, such as sentiment analysis or question answering, the pre-trained model is adjusted using a smaller dataset created especially for the task. The pre-trained model's last layers are added or removed during fine-tuning, and back propagation is utilized to train the model on the task-specific dataset.

BERT has attained cutting-edge results in a variety of NLP tasks, and it is simple for academics and developers to utilize it for theirown NLP applications because its pre-trained weights are freely available.

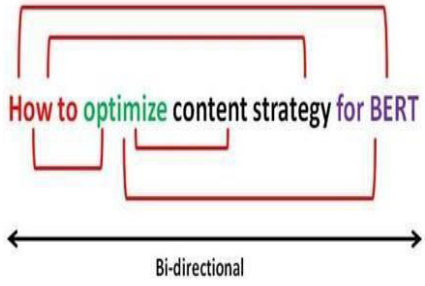


Fig 3.2.2(a) BERT Architecture

3. 2. 3. EMBEDDINGS:

Image Embeddings:

The process of turning an image into a vector representation that can be fed into machinelearning models is known as image embedding. Convolutional neural networks (CNNs) that have already been trained to extract high-level features from images, like VGG, ResNet, or MobileNet, can be used to create image embedding.

During the image embedding process, a pre- trained CNN is used to extract high-level characteristics from the image. These properties are then put into a fully linked layer, where they are flattened to produce an image embedding.

Text Embedding:

Text embedding is the process of converting text into a numerical vector representation that may be fed into machine learning models. Language models that have already been trained and are intended to extract semantic information from text, such BERT, GPT, or Word2Vec, can be used to do text embedding. Text embedding is the process of running the text through a language model that has already been trained, which generates a sequence of contextualized embeddings for each word in the text. These embeddings capture the meaning and context of each word, as well as the relationships between words in the sentence.

Joint Embedding:

Joint embedding refers to the process of creating a shared vector space that can represent both images and text in a way that captures their relationships. This is achieved by combining image embedding and text embedding into a single vector space, where each image and text has a corresponding vector representation that Is close to similar images and text in terms of their meaning.

2. 4. Similarity Detection Module: After performing all the above operations we store the joint embeddings for all the data. Now we follow the same strategy for the Inputtext and image which is given by the user. Now it's time to compare the similarity between the joint embedding from our dataset to the user provided text and image joint embedding. To find the similarity there are many different algorithms like Cosine similarity, Jaccard

ISSN PRINT 2319 1775 Online 2320 7876
 UGC CARE Listed (Group -I) Journal Volume 11, Iss 13, 2022

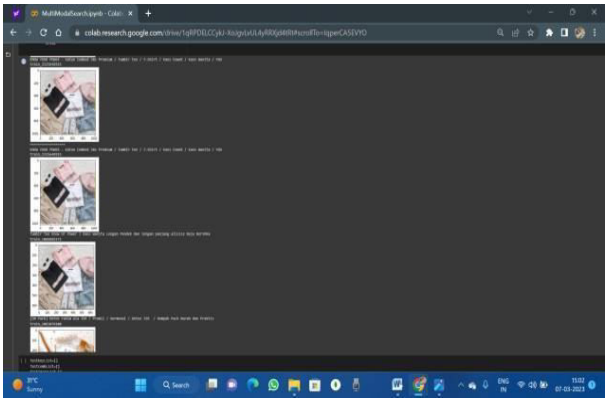
similarity, Levenshtein distance, Sequence alignment.

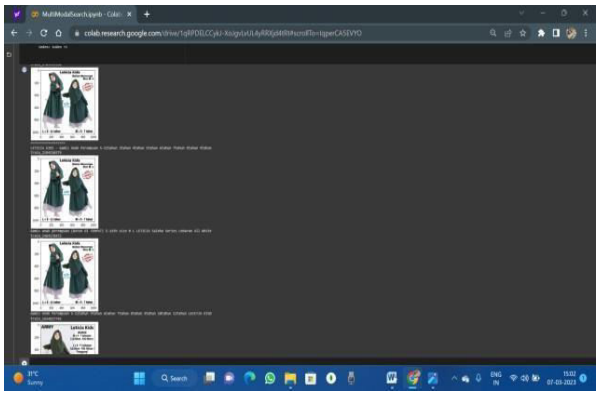
In this case, cosine similarity is used to find similarities. In an inner product space, two non-zero vectors can be compared using the cosine similarity measure. Between -1 and 1, it computes the cosine of the angle formed by the two vectors. In terms of cosine similarity, two vectors are considered equal when their cosine similarity is 1, orthogonal when their cosine similarity is 0, and diametrically opposing when it is 1.

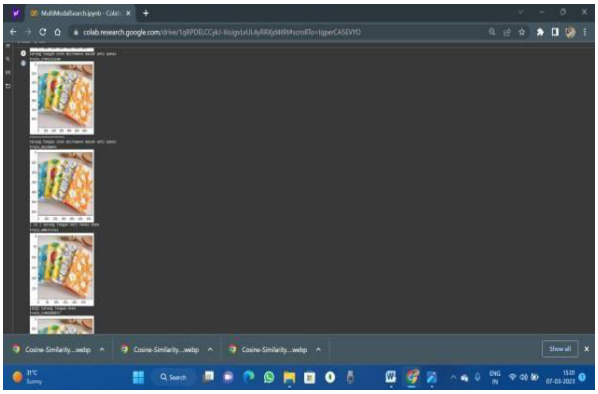
In machine learning and natural language processing, cosine similarity is frequently used to assess how similar two texts or text segments are to one another. Here, each document is transformed into a vector of word embeddings or a bag of words in order tocreate the vectors.

As the angles of the papers approach eachother, the Cosine Similarity rises (Cos theta).

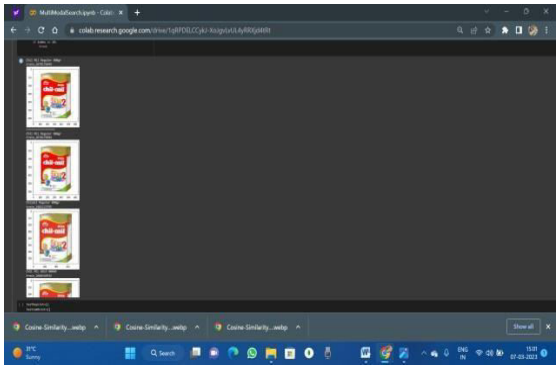
By checking the similarity between both joint embeddings from our database to user inputs we get top 'N' nearest results as output. For that we are using nearest neighbour 'kd-tree' algorithm.







3. Results:



4. Conclusion:

Multi-model search combining text and image is a powerful technique that can enhance the accuracy and relevance of search results.By analyzing both visual and textual features, the multi-model search can provide a more comprehensive understanding of the content and context of a given query. Applications for multi-model text and picture search include visual search, image captioning, and recommendation systems. The quality of the individual models used to handle textual and visual data, as well as the efficacy of the methods used to combine them, are key factors in the success of multi-model search for text and images. All things considered, multi-model search for text and images is an exciting field of research with the potential to significantly improve search engine efficiency and relevancy while also creating new avenues for application across numerous domains.

5. References:

A. Rubio, LongLong Yu, E. Simo-Serra, F. Moreno-Noguer(2017). Method for multi- modal product retrieval in the fashion e- commerce domain, combining images and textual metadata into a shared latent space.

Karpathy, A., and L. Fei-Fei (2015). deep visual-semantic relationships for creating image captions. chosen works from the 3128– 3137 pages of the IEEE Conference on the Recognition of Patterns and Computer Vision.Kiela, D., and Bottou, L. (2014). To improve multi-modal semantics, convolutional neuralnetworks are employed to learn picture embeddings. in 2014 Conference Proceedings on Empirical Methods in Natural Language Processing (EMNLP), pp. 36-45.

Sheikh, H. R., Simoncelli, E. P., Wang, Z., and Bovik, A. C. (2004). Image quality evaluation: from error visibility to structural similarity. 13(4), IEEE Image Processing, 600–612.

Li (2009), Li (2009), Li (2009), Deng (2009),

Dong (2009), Socher (2009), and Li (2009). ImageNet is a large-scale hierarchical image database. 2009 IEEE Conference on Computer Vision and Pattern Recognition, pages 248–255.IEEE

Yingbo Zhou and colleagues' "Multi-Objective Bayesian Optimization for Neural Architecture Search" (AAAI 2021) - In order to optimize many objectives at once for neural architecture search, this study presents a multi-objective Bayesian optimization technique.