Deep Learning System for Image Retrieval

# Sumanth S Rao

*Department of Computer Science PES University* Bangalore, India

[sraosumanth@gmail.com](mailto:sraosumanth@gmail.com)

# Shahid Ikram

*Department of Computer Science PES University* Bangalore, India

[shahidikram0701@gmail.com](mailto:shahidikram0701@gmail.com)

# Parashara R

*Department of Computer Science PES University* Bangalore, India

[parasharar@gmail.com](mailto:parasharar@gmail.com)

***Abstract*—In the modern era of digital photography and advent of smartphones, millions of images are generated every day and they represent precious moments and events of our lives. As we continue to add images to our digital storehouse, the management and access handling of the images becomes a daunting task and we lose track unless properly managed. We are in essential need of a tool that can fetch images based on a word or a description. In this paper, we try to build a solution that retrieves relevant images from a pool, based on the description by looking at the content of the image. The model is based on deep neural network architecture and attending to relevant parts of the image. The algorithm takes a sentence or word as input and obtains the top images which are relevant to the caption. We obtain the representation of the sentence and image in a higher dimension, which enables us to compare the two and find the similarity level of both to decide on the relevance. We have conducted various experiments to improve the representation of the image and the caption obtained in the latent space for better correlation, for e.g use of Bidirectional sequence models for better textual representation, use of various baseline convolution-based stacks for better image representation. We have also tried to incorporate the attention mechanism to focus on only the relevant parts of the image and the sentence, thereby enhancing the correlation between the two spaces.**

1. INTRODUCTION

The issue of image retrieval can be seen dependent on a great deal of justification for relating a picture with a specific setting or portrayal. It depends on what we are looking for in the picture and need not really be what the real pictorial portrayal delineates. content-based image retrieval is the image retrieval technique widely popular today because it is different from the conventional meta-data or tag-based image retrieval. Content-based implies that the pursuit breaks down the substance of the image instead of the metadata like human annotated labels etc. The term content in this setting may allude to hues, shapes, surfaces or whatever other data that can be obtained from the image itself. Having people physically explain images by entering watchwords or metadata in an extensive database can be tedious and may not capture the catchphrases needed to portray the image.

Deep learning techniques in the computer vision domain were investigated upon to solve the problem of content-based image retrieval. The image retrieval model being proposed is a proof of concept that the model can be used on any image data to retrieve them using their description. The successful

validation of the model will lead to the proof of the concept and hence the model can be used to train any arbitrary image data and can be utilized to retrieve the images.

In this paper we propose a method that builds on the base- line model for content-based image retrieval. In the baseline architecture, the main idea is to reduce the distance between the feature representations of the image and its corresponding caption in the latent space. This distance is nothing but the similarity score which remains high for the actual pairs than the other ones. Various experiments were tried upon to obtain the best baseline model on which further improvements could be made. Further improvements to this model mainly focused on improving the feature representations of the image and the caption in the latent space. This was observed to happen on investigating into incorporating this deep learning technique called self attention while generating the feature vectors.

Various experiments were carried out and a set of metrics were recorded for comparative analysis. The metrics recorded for each experiment are the ones typically used in recom- mender systems namely: Recall@k and Mean Rank.

The Microsoft COCO dataset [1] containing an extensive set of images and corresponding captions was used for the experimentation.

1. RELATED WORK

In this section we provide background on existing work on image captioning, image retrieval and attention. The first approach to have achieved state of the art results in the multimodal machine learning domain for image captioning is shown by Show, Attend and Tell by Kelvin Xu et al. [2], where the concept of attention was successfully implemented on images to fix the gaze on a particular object in the image while predicting a particular word. The approach used was to shrink the focus area on the image to a region where the object is located on the image so that the prediction can be made to a more accurate level. Ashish Naswani et al. in their paper: Attention is all you need [3], prove the power of attention models by demonstrating that just using a sophisticated attention mechanism can eliminate the need for traditional sequence based model and show that these models to be superior in quality while being more parallelizable and requiring significantly less time to train. The transformer networks used in the model have multiple levels through which the input is processed and the number of units in these layers

can be handcrafted based on the problem domain and it’s flexible in addition to being efficient. Jason West et al. [4] in their work show that the joint feature representation of image and text is the core idea of multi-modal learning and they also use a sampling trick that leads to faster convergence and retrieval even on large datasets. The solution prove the scalability of the theory to large vocabulary datasets by using the above mentioned sampling trick and a unique loss func- tion.Another key idea for the multi-modal training of images and text is the need to preserve the visual semantic hierarchy. This paper lays the foundation for multi-modal approach on a large scale and proves its applicability on real world data. Ivan Vendrov et al. [5] in their work developed a rather interesting loss function which preserves the partial order as opposed to euclidean distance between similar objects of different modalities in a shared embedding space. This was done by learning an approximate order embedding which violates the order-embedding condition as little as possible by introducing a penalty. In doing so, they were able to show that using such a loss function gave them an improvement over state of the art results in hypernym prediction, caption-image retrieval and natural language inferencing. A very interesting take on cross-modal retrieval using generative models was proposed by Jiuxiang Gu et al. in their work. Here, the authors use two generative models(image-text, text-image) combined with an order preserving loss function which essentially tries to look at the data in one mode(image or text) and imagine what this data will be like in the other mode using either of generative models and finally match the pairs of input and generated output with what is supposed to be ground truth and learn from it [6]. This approach currently gives the state of the art results in the field of neural image retrieval.

1. MODEL ARCHITECTURE
2. *Overview*

We propose a multi-head architecture based model which can learn the hidden mapping between the two distinct modalities of images and text effectively. Past answers for this issue additionally utilize a comparable multi-head input model design where the caption is transformed into a higher dimension vector which the decoder at that point interprets and emits some sort of expectation. There are additionally a few arrangements proposed which fuses some Natural Language Processing based Grammatical-form-models in unison with Computer Vision based object recognition to make sense of a mapping between the terms in the sentence to the items found in the image. Our methodology basically revolves around the use of multi-head based designs without the association of any scene comprehension or object detection models or Natural Language Processing based methodologies. In this task we have explored different areas regarding a few variations of best in class segments which when combined together makes up our novel architecture. We have taken a stab at trying different things with a few variations of recurrent systems, word embedding models, convolutional neural network stacks

for transfer learning and furthermore fused the idea of self- attention into our models.

1. *High Level System Architecture*

As stated above, our model primarily revolves around the multi-head architectural design pattern. The figure below shows a high level abstract model architecture where each component can be realised with several implementations. This way we have carried out experiments by trying several implementations for a specific component. Each experiment carried out, has been designed in such a way so as to over- come the limitations of previously carried out experiments by incrementally replacing some existing components with more state of the art components which serve the same functionality.

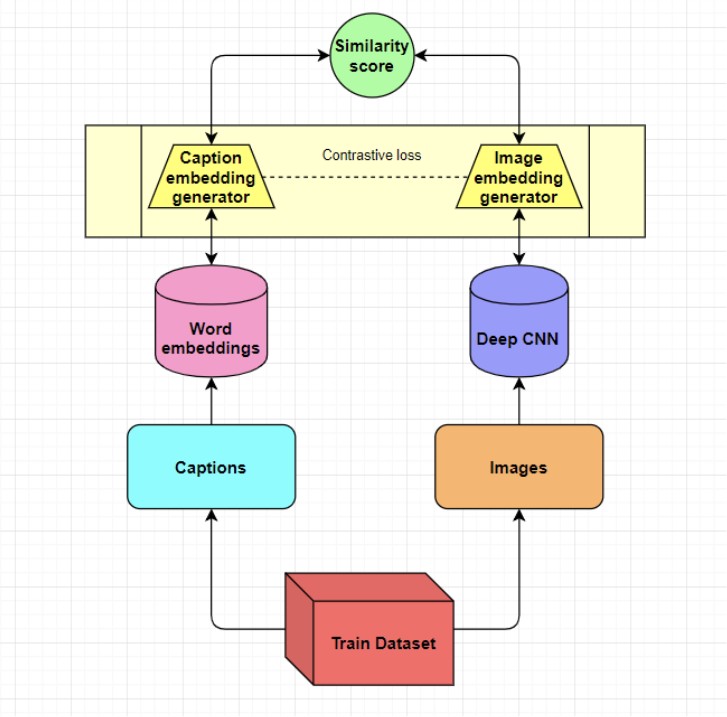


Fig. 1. High level architecure of the model

1. *The model Input*

The image dataset is picked by the MS-COCO dataset is used to train this model. It consists of 16 GB of nearly 400,000 images with every image associated with 5 captions each. The caption dataset consists of all the captions associated with the images picked for training. During training, each caption is split into tokens and each token is passed through a pre- trained neural word embedding model. Some of the popular word embedding neural models include word2vec, fasttext, elmo, bert. In this project we have extensively used FastText as our primary word embedding model after experimenting with Elmo, Word2Vec and Glove.

1. *Model Summary*

During training, each image is fed through a deep pre- trained convolutional stack for extracting the image’s feature vector for feeding into the multi head neural embedding. In this project we have extensively used Inception as it proved us with the best results, as the idea of the inception module combining the output of each module by concatenating over multiple kernel sizes accounts for all the features and captures every

detail precisely. The caption representation in the latent space is passed through a multi head self attention layer. Now that

*S*(*c, i*) = *− | f* (*i*) *− f* (*c*) *|*2 (1)

the results improved by improving the caption representation *i c*

in the latent space, we focused on improving the image feature

representation as well. We decided to add a layer of self- attention on the image side. What good is self-attention going to do? While convolutional filters are good at exploring spatial locality information, the receptive fields may not be large enough to cover larger structures. We can increase the filter size or the depth of the deep network but this will make the training complex. Alternatively, we can apply the attention concept. Now we pay attention to only particular feature maps, we have a larger receptive field and the context is more focused and more relevant. It has already been shown in the paper Attention is all you need [3]) that non-recurrent architectures (self-attention) have outperformed recurrent neural networks in neural machine translation. Self-attention networks can connect distant words via shorter network paths than recurrent neural networks, and it has been speculated that this improves their ability to model long-range dependencies.

The final step in the architecture is the second head of the multi-head model, where we take the image feature representa- tions from this CNN model and then pass it through a custom neural network to get the embedding of the image, which is used along with the caption embedding in order to arrive at the similarity score. When we train the two representations, we learn a mapping between how images and captions are correlated.

1. *Training procedure*

All the models were trained with 500,000 images picked up from the Microsoft COCO dataset, they were trained with stochastic gradient descent with the recently proposed Adam algorithm [7] .The captions were processed by building a vocabulary of all the words in the training corpus and mapping them to unique tokens using a dictionary.

The image side training consists of the first level training with the Inception model consisting of 27 layers, we have frozen the Inception architecture which outputs a vector of (batch size, 64, 2048) for every batch, which is passed on to the custom convolutional neural network stack which trains the model with respect to the description for the specific purpose we are using the image for. After this set of training, we obtain a (batch size, 1024) vector for the image representation.

The caption representation training consists of the processed caption of dimension(batch size, 52), the largest caption we are training the model for is of the length 52. The captions are embedded to make each token a 300 dimensional vector. The caption vector is position embedded and passed on to a global average pooling layer along one dimension to obtain a two dimensional vector map, which is then reduced to the same dimension as image by convolution filters. Now, we have both the image and caption vectors in the same shape(batch size, 1024). The following section explains the mechanism of calculating the similarity between the two modalities.

S(c,i) is defined as the similarity score between the caption “c” and the image “i”, Here, fi(i) and fc(c) represents the embedding functions for images and captions respectively. The negative sign indicates that we would like larger similarity value(S) to indicate more similarity. Here this function is basically called as L2 normalization. Our loss function can be defined as minimizing the following equation.

(*c,i*) [ *c max*[0*, α − S*(*c, i*) + *S*(*c, i*)] + *i max*[0*, α − S*(*c, i*) + *S*(*c, i*)]] (2)

This loss function is written in this manner for the purpose of contrasting correct and incorrect matches, we introduce negative examples that comes handy in the same training batch. Where (c, i) is the true caption-image pair, c’ and i’ refer to incorrect captions and images for the selected pair. Therefore, the cost function enforces positive (i.e. correct) examples to have zero penalty and negative (i.e. incorrect) examples to have penalty greater than a margin *α*. When training, the embedding functions fi(i) and fc(c) both return a 1024 dimensional vector, and to indicate to the model that a particular image is related to a particular caption we pass along a 1024 dimensional vector full of zeroes. After training the model using this objective function the weights of the embedding model of both the caption and that of the image will be aligned in such a way that for a corresponding caption- image pair the similarity score S(c,i) *≈* 0 and for any other non-matching caption-image pair the similarity is *≈* 1.

1. EXPERIMENTS CONDUCTED

We have tried out a multitude of experiments where each new model developed had some minor architectural differences which gave it a slight edge over its previous model. There are four major components involved in the architecture:

* 1. Image Representation Model
     1. Inception V3 [INC] - Inception is a deep neural network trained on ImageNet, which is a dataset of over 15 million labeled high-resolution images with around 22,000 categories. [8]
     2. ResNet [RES] - The core idea of ResNet is in- troducing a so called “identity shortcut connec- tion”(Residual blocks) that skips one or more lay- ers. [9]
  2. Sequence Model
     1. GRU [GRU] - Gated recurrent unit is an improved version of Recurrent neural networks with reset and update gates aimed at reducing he vanishing gradient problem.
     2. Bidirectional [LSTM] - Long short term memory networks is another flavour of Recurrent neural networks with an additional forget gate to retain only the relevant information in sequence.
  3. Caption representation model
     1. FastText [FSTX] - FastText is a text embedding introduced by Facebook research. It uses sub word information to capture the essence of each word vector. [10] [11]
  4. Self Attention Layer
     1. SelfAttention [SA] - Self-attention, also known as intra-attention, is an attention mechanism relating different positions of a single sequence in order to compute a representation of the same sequence. In our experiments this layer has been applied on both the image and text representations. [3]

Listed below are all the experiments we conducted so far:

1. Baseline models
   1. RES + GRU + FSTX
   2. INC + GRU + FSTX
2. Baseline with bidirectional sequence model for captions
   1. INC + LSTM + FSTX
3. Self attention models
   1. SA(INC) + LSTM + FSTX
   2. INC + SA(FSTX)
   3. SA(INC) + SA(FSTX)
4. *Baseline models*

Baseline models were the simple multihead architecture to extract the feature representations of image and caption and minimise the loss function. We used transfer learning to extract the image representations as well as pre-trained word embed- dings for the caption side. Therefore, for these two baseline models we decided to choose the “FastText” neural word embedding developed by facebook over other word embedding models such as word2vec, glove, elmo etc., because the way FastText is trained depends on n-gram models of the word, which implies that even for out of vocabulary words the model provides an excellent embedding representation because of the n-gram nature of the fasttext model. And on the image side we decided to go with two of the most famous deep convolutional neural network stacks which are Inception and ResNet as both models have achieved state of the art performance on the image net dataset.

1. *Baseline with bidirectional sequence model for captions*

Next we aimed at improving the features extracted for cap- tions. As a bidirectional sequence model will run the inputs in two ways, one from past to future and one from future to past and what differs in this approach from unidirectional is that, in the LSTM that runs backwards we preserve information from the future and using the two hidden states combined, we can at any point in time preserve information from both past and future and hence it understands the context better thereby providing a better feature representation of the caption.

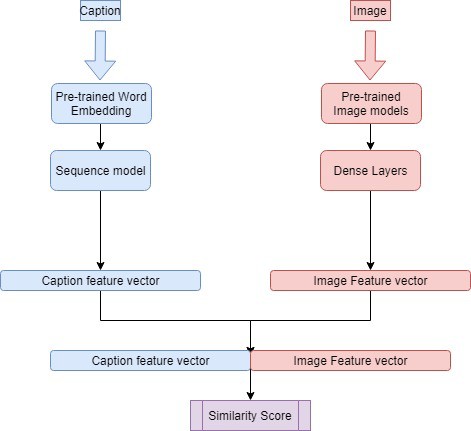


Fig. 2. Architecture: Baseline model

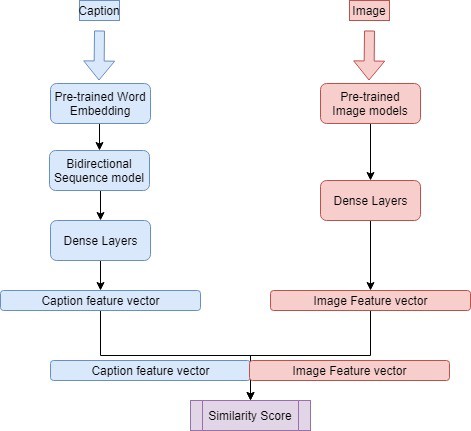


Fig. 3. Architecture: Bidirectional model

1. *Self Attention models*

Now that the results improved by improving the caption representation in the latent space, we focused on improving the image feature representation as well. We decided to add a layer of self-attention.

As a result, to the model configuration that had given the best results so far, self-attention was employed:

1.To the image representation alone 2.To caption representation alone

3.To both caption and image representations

1. INFERENCES

In our experiments we chose to implement the following two metrics to find out the performance of our model which are Recall@10 and mean rank respectively.

Our models performance can be thought of as directly proportional to the Recall@10 metric inversely proportional to the mean rank.

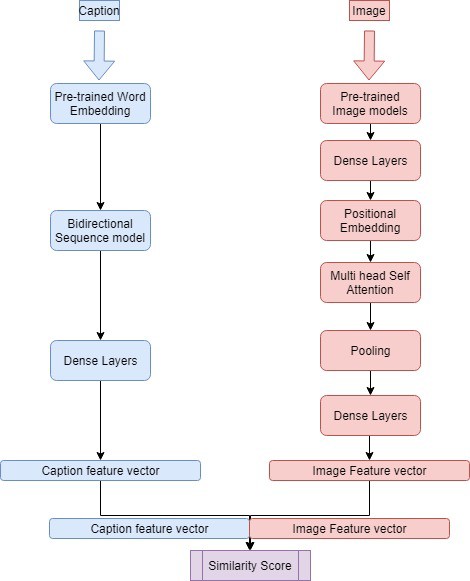
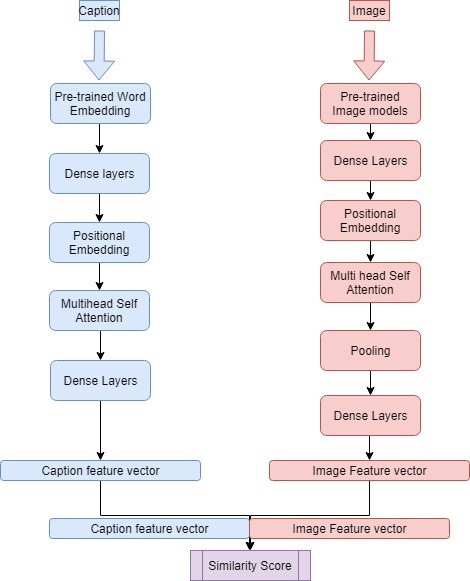
 

Fig. 4. Architecture: Image-self attention model

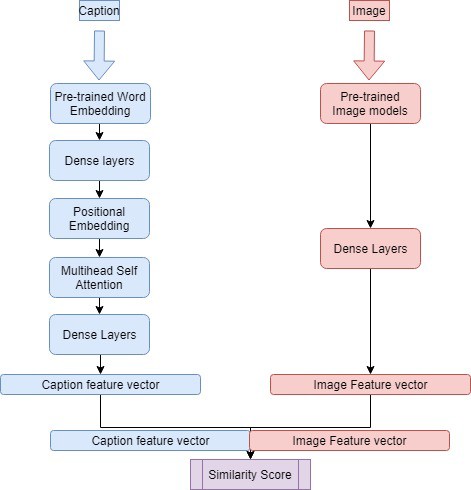


Fig. 5. Architecture: Caption Self Attention model

Going by this principle we can see that the bidirectional models have outperformed the base models and that the models which incorporate self-attention outperforms both the bidirectional and the base models.

The reason the bidirectional model outperforms the base model is primarily due to two factors: 1. Usage of LSTM over GRU provided the model with two hidden cell states as opposed to one. 2. Usage of the bidirectional layer gave the model over 4 hidden cell states, and generally models which have more context tend to learn finer abstractions of the captions and hence resulted in better performance. The

Fig. 6. Architecture: Image-Caption self attention

primary reasons the models with self-attention performs better than the model with the bidirectional model is because of the fact that multi head self attention trumps the recurrent neural networks as this provides more context for the model to learn from while not suffering from the vanishing gradient problem faced by recurrent neural networks.

1. CONCLUSION

With the above work, we have expanded the horizon for research in multi-modal machine learning. The fact that two varied modes, with structurally diverse dimensions can be combined and brought together on a common ground to process and extract the inherent relationship which is hard to perceive for human comprehension. The model is a proof of how the deep learning algorithm we are using tries to learn and create its internal representation of the inputs to correlate and retrieve the right features while it is being used for prediction. It essentially creates an internal map of image features in an effective way so it can retrieve the images later on based on text features. Our model, which provides a proof for the above mentioned concept was successful on most of the common scenarios that occur in reality. We saw that model brings together description and images in unique ways based on various levels of commonality. We observed the images being linked to a description based on its features, objects and also based on the background, colour, domain, action, complexity etc., this reassures the fact that attending parts of an image and caption based on the content can be applied on multi-modal machine learning problems and can be extended to any pair of modalities to learn the correlation between the two. The fact that the model with self-attention is performing consistently validates the idea and provides a wide scope for research along

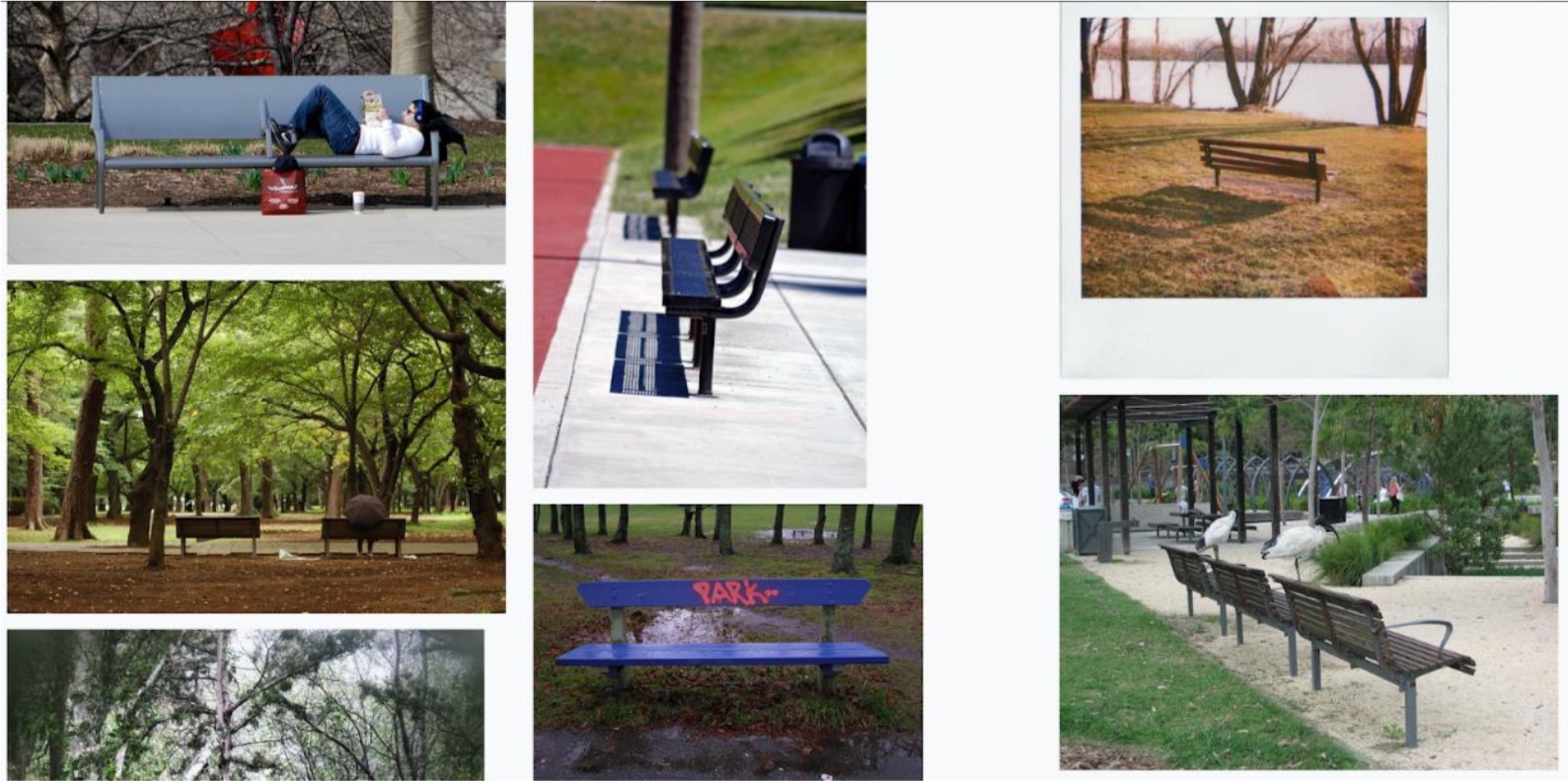


Fig. 7. Image Retrieval: A bench in the park.

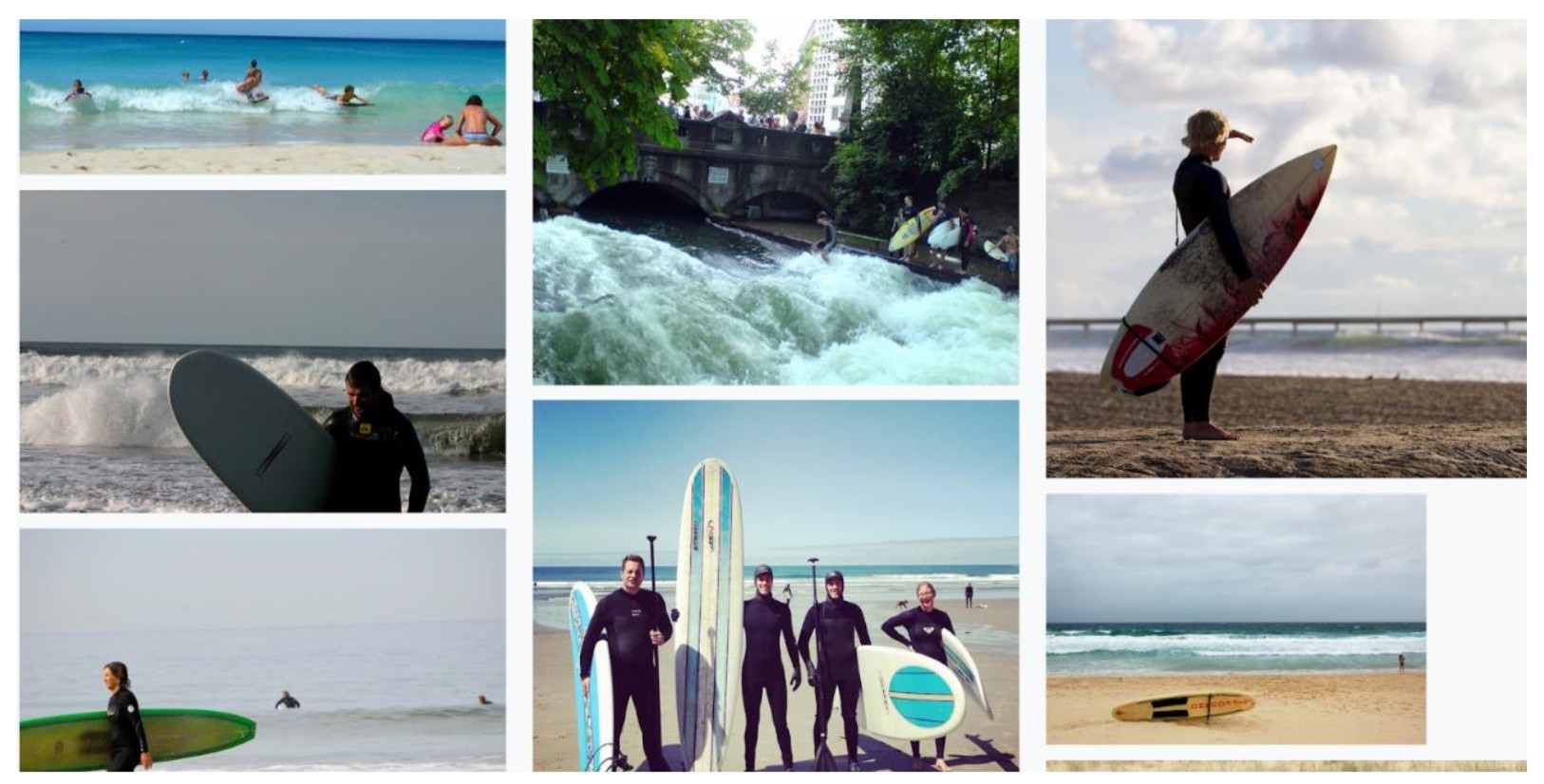


Fig. 8. Image Retrieval: Surfing on the beach.

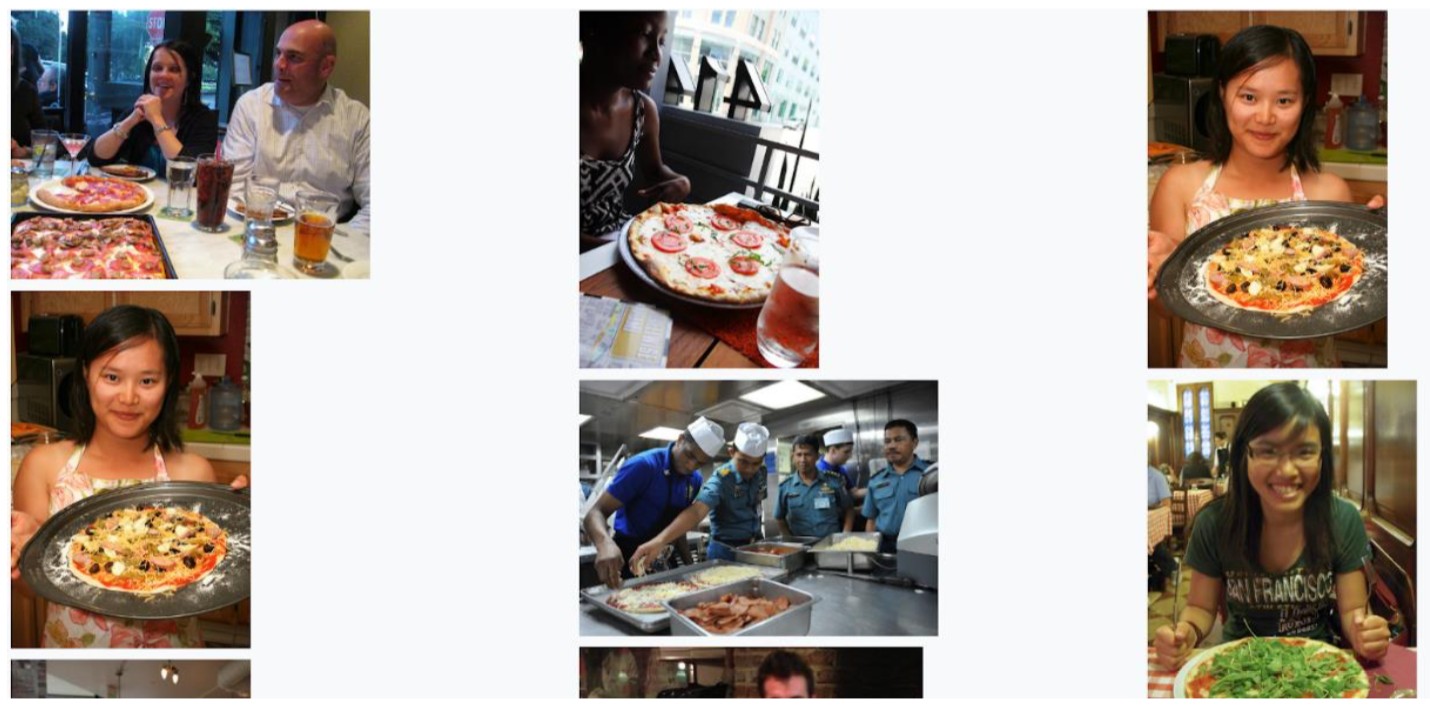


Fig. 9. Image Retrieval: People eating pizza.

the same lines. Applying cross attention between the word and image features and attending the most relevant features for the word level predictions will improve the model performance by multiple folds and clustering approach we followed to narrow down the search space was also successful and if incorporated with the model would prove to be a state-of-the-art solution for the domain.

1. FUTURE WORK

While retrieving the top images based on the user’s query, the more the pool of images to search from, the more the time required. We wish to reduce this retrieval time. This problem would be tackled using clustering. Since the images will be encoded in such a way that the concepts captured in the encoded form will relate to the caption, we can perform clustering on this encoded form and final search for relevant images from a specific cluster. This would speed up the retrieval process. We experimented by fixing an arbitrary number of clusters to see if images that go into each cluster capture some common theme. The results of this experiment can be observed in fig 7-10. Future enhancement would focus

Fig. 10. Cluster: Cluster capturing the concept of people.



Fig. 11. Clustering: A cluster of vehicles obtained from the samples

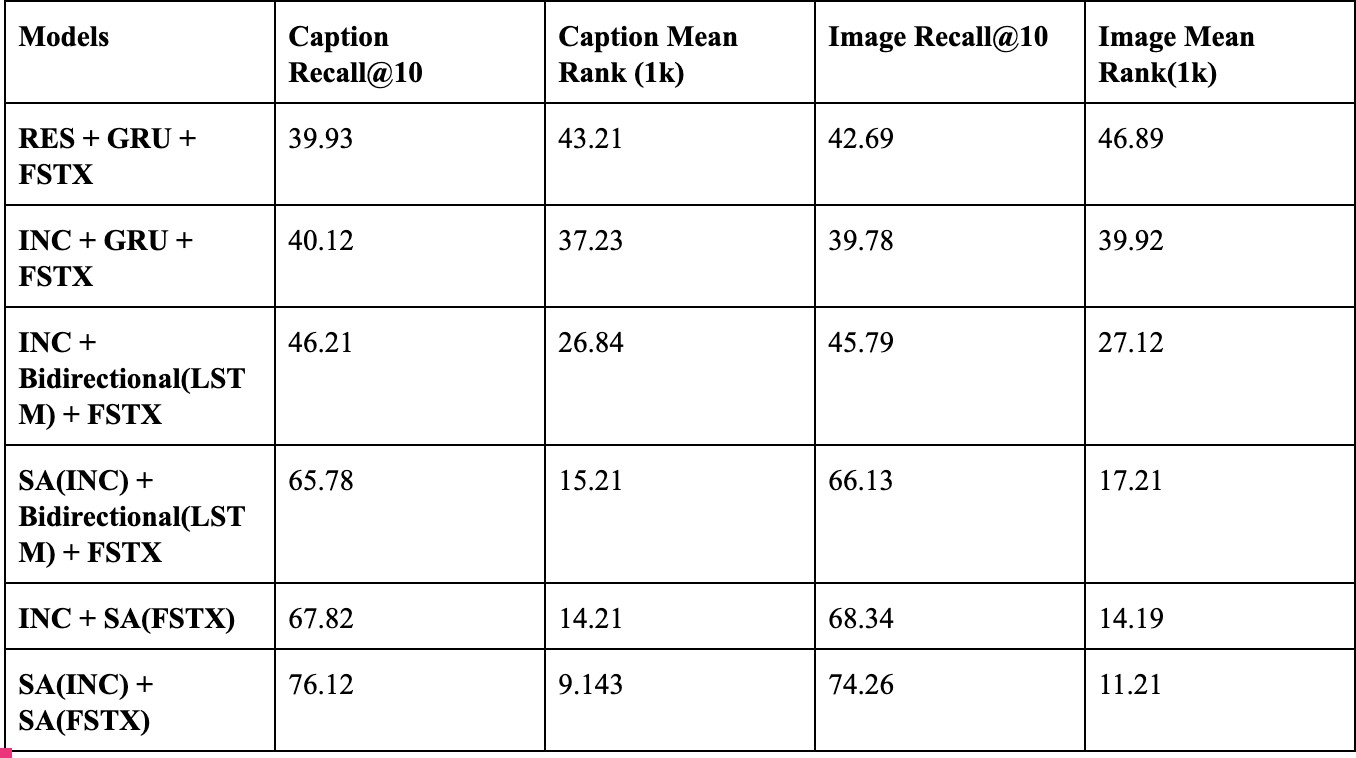


Fig. 12. Results - The metrics scores obtained for all the experiments tried. please refer to section IV(Experiments conducted) for detailed explanation of acronyms used)

on arriving at the cluster number dynamically so that we have as many clusters formed as the number of generalised concepts that can be captured [12].

Another approach worth trying out in the future is the idea of incorporating cross attention between the two modalities on top of our self attention framework, as we believe that only by having the right mixture of both, can the model learn what features to attend to, at the same time learning to self attend on key features by only looking at one modality alone. We believe that a ”transformer” [3] like architecture will be able to achieve this and give near state of the art results.

Another exciting approach to solve this problem was pro- posed by Jiuxiang Gu et. al in their work “Look, Imagine and match” [2]. An addition to their existing architecture which we believe would add better results, is the incorporation

of concepts like self-attention and cross-attention in their existing generative framework. This is because as humans we intuitively try to solve this problem using the “Look,imagine and match” philosophy but as humans, concepts like attention is wired in our brains naturally which is in fact missing in their work. So adding attention based models to their existing framework, should in theory give even better results.

Another key idea which we believe can shoot up accuracy in the “Look imagine and match” approach, is the inclusion of cycle loss [13] (i.e minimise the image-text-image conversion and text-image-text conversion) in their framework which will in theory improve their generative process to give constrained results which are much more aligned with the global informa- tion present in both the images and captions from the dataset.

REFERENCES

1. Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, Piotr Dollar ”Microsoft COCO: Common Objects in Context”
2. Jiuxiang Gu, Jianfei Cai, Shafiq Joty, Li Niu, Gang Wang ”Look, Imagine and Match: Improving Textual-Visual Cross-Modal Retrieval with Generative Models”
3. Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, Illia Polosukhin ”Attention Is All You Need”
4. Jason Weston, Samy Bengio and Nicolas Usunier ”Scaling up to large vocabulary image annotation”, 2013
5. Ivan Vendrov, Ryan Kiros, Sanja Fidler, Raquel Urtasun ”Order- Embeddings of Images and Language”
6. Oriol Vinyals, Alexander Toshev, Samy Bengio, Dumitru Erhan ”Show and Tell: A Neural Image Caption Generator”
7. Diederik P. Kingma, Jimmy Ba ”Adam: A Method for Stochastic Optimization”
8. Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, Andrew Rabinovich ”Going deeper with convolutions”
9. Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun ”Deep Residual Learning for Image Recognition”
10. Piotr Bojanowski, Edouard Grave, Armand Joulin, Tomas Mikolov ”Enriching Word Vectors with Subword Information”
11. Armand Joulin, Edouard Grave, Piotr Bojanowski, Tomas Mikolov ”Bag of Tricks for Efficient Text Classification”
12. Joris Guerin, Olivier Gibaru, Stephane Thiery, and Eric Nyir ”CNN features are also great at unsupervised classification”
13. Jun-Yan Zhu, Taesung Park, Phillip Isola, Alexei A. Efros Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Net- works