ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

IJARCCE

International Journal of Advanced Research in Computer and Communication Engineering

PCON-2019

National Conference and Seminar on Innovations in Engineering & Technology

Govt Polytechnic College, Palakkad

Vol. 8, Special Issue 1, January 2019

Survey on Visual Question Answering and Multimodal Compact Bilinear Pooling

Pradeep T[1], Rafeeque P C[2]

PG Student, Dept of Computer Science & Engineering, Government Engineering College, Palakkad, Kerala, India[1]

Head of the Department, Department of Computer Science and Engineering, Government Engineering College, Palakkad, Kerala, India[2]

Abstract: Visual Question Answering (VQA) is the process of extracting the answer of the question based on the given image. Here the input is an image along with a natural language question regarding the image. The system will analyze the question and image, then extracts the answer of the question from the image. So this process is the combination of both Computer Vision (CV) and Natural Language Processing (NLP). Computer vision is used to analyze the image and NLP is required when analyzing the question and generating the answer. In VQA the answer is obtained by the mutual interaction between the image and textual vectors. Among that outer product based method between the two vectors are superior to all other. But since outer product is infeasible due to its high dimension, Multimodal Compact Bilinear Pooling (MCB) is used to efficiently combine the different features. Multimodal Compact Bilinear Pooling is one of the recent technique to perform VQA. For VQA here uses MCB twice, one for predicting the spatial attention over the images and another for combining these attentions with the question features. When applied on Visual7W dataset, this model outperforms the baseline approaches and the VQA challenge.

Keywords: Natural Language Processing, Computer Vision, Deep learning, VQA, Count-sketch projection

I. INTRODUCTION

Combining the image and textual features in an effective way will produce a reasonable result in various NLP and CV problems. For obtaining the text features from a set of words we can use the Recurrent Neural Networks(RNN) such as LSTM or GRU. For obtaining the image features from an image we can use the Convolution Neural Networks(CNN) such as VGG, ResNet, AlexNet etc. For combining these two features there are lots of approaches using today such as concatenation and element wise sum etc. But all of these produce a joint representation, it might not be expressive enough to fully capture the deep correlation between the two different features. In this paper we use the Multimodal Compact Bilinear Pooling method for combining the Image and Textual features together. Bilinear pooling computes the outer product between the two vectors so that way it allows the multiplicative interaction between the features rather than the element wise product. Each element in one vector interacts with all other elements in the other modality vector. So t other VQA methods. However, given that it shows the better performance among the their high dimensionality ($n2$), bi-linear pooling has so far not been widely used. So in this paper here discussing how to efficiently compress the bilinear pooling for single modality.

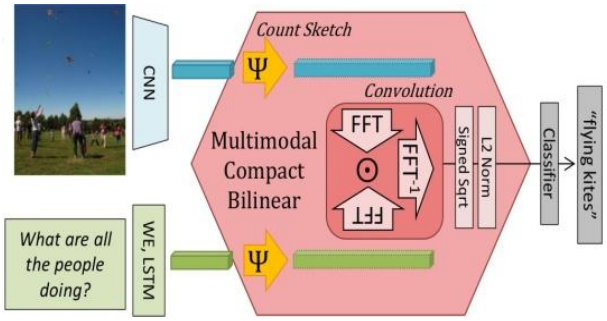

Fig. 1. Multimodal Compact Bilinear Pooling (MCB)

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

IJARCCE

International Journal of Advanced Research in Computer and Communication Engineering

PCON-2019

National Conference and Seminar on Innovations in Engineering & Technology

Govt Polytechnic College, Palakkad

Vol. 8, Special Issue 1, January 2019

Figure I shows the basic architecture of Multimodal Compact Bilinear Pooling. In this architecture MCB is approximated by randomly projecting the image and text representations to a higher dimensional space and then convolving them by Fast Fourier Transform (FFT). Here using the MCB in VQA to predict the answers of the given question from the image and select the locations for the visual grounding task. For open-ended question answering, we need MCB twice, one for predicting spatial attention and another time to predict the answer. For multiple-choice question answering it require a third MCB to relate the encoded optional answer to the question-image space. Here the count sketch production is using to reduce the higher dimensionality issues in MCB.

The rest of the paper is organized as follows: Section 2 describes about Multimodal Compact Bilinear Pooling for Visual and Textual Embeddings. Section 3 describes about Simple Baseline for Visual Question Answering. Section 4 describes about Bilinear CNN Models for Fine-grained Visual Recognition. Section 5 describes about Fisher Vectors Derived from Hybrid Gaussian-Laplacian Mixture Models for Image Annotation. Section 6 is the comparison of different methodologies using for VQA. Section 7 concludes this paper with future improvements. And finally the references are also added.

II. MULTIMODAL COMPACT BILINEAR POOLING FOR VISUAL & TEXTUAL EMBEDDINGS

In VQA we predict the most likely answer or location $\hat{a}$ for the given image **x** and question or phrase **q**. The equivalent formula is given by

$$\hat{a} = \underset{x \epsilon A}{argmax} p(a|x, q; \Theta)$$

- Θ = Parameter
- X = Image embedding
- q = Question embedding
- A = Set of answers or locations

CNN and question embedding LSTM, here are interested in getting a good joint representation by pooling both representations. With a multimodal pooling $\varphi(x, q)$ that encodes the relation- ship between x and q well, Here discussing about multimodal pooling φ for combining representations from different modalities into a single representation and then detail architectures for VQA and visual grounding , after that how we are predicting $\hat{a}$ with the given image vectors and text vectors.

A. *Multi modal Compact Bilinear Pooling (MCB)*

Consider two vectors x and q. Where $x \epsilon R^{n1}$ and $q \epsilon R^{n2}$ and which learns W such that $z = W[x \otimes q]$ where $\otimes$ denotes the outer product (xq^T) and [ ] denotes linearizing the matrix in a vector. Bilinear pooling is interesting because it allows all elements of both vectors to interact with each other in a multiplicative way. But it will enlarge the dimensional of the product. So it is difficult to handle such huge numbers. So we need some other mechanisms to overcome this. So here using the Count Sketch projection Ψ function for this purpose. So we needs to projects the outer product to a lower dimensional space and also avoids computing the outer product directly.

Count Sketch projection which projects a vector $v \epsilon R^n$ to $y \epsilon R^d$. Here initialize two vectors $s \in \{-1,1\}^n$ and $h \in \{1,...,d\}^n$. Both s and h are initialized randomly from a uniform distribution and remain constant for future invocations of count sketch. For every element v[i] its destination index j = h[i] is looked up using h, and s[i]•v[i] is added to y[j].

Project the outer product to a lower dimensional space, which reduces the number of parameters in W. Count sketch of the outer product of two vectors can be expressed as convolution of both count sketches.

$\psi(x \otimes q,h,s) = \psi(x,h,s) * \psi(q,h,s)$ ($*$is the convolution operator).

The convolution x0 * q0 can be rewritten as

$FFT^{-1}(FFT(x') \Theta FFT(q'))$,

Where Θ refers to element-wise product. Finally the system converges to a multi classification problem of 3000 classes. Then extracts the answer.

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

IJARCCE

International Journal of Advanced Research in Computer and Communication Engineering

PCON-2019

National Conference and Seminar on Innovations in Engineering & Technology

Govt Polytechnic College, Palakkad

Vol. 8, Special Issue 1, January 2019

III. SIMPLE BASELINE FOR VISUAL QUESTION ANSWERING

In this paper [1] describes the simple Bag Of Words approach for visual question answering. So for this purpose first extracts the feature vectors from the image and textual vectors from the question. So that it produce some reasonable accuracy in the VQA dataset. Combining natural language and image is the one of the recent trend in machine learning. In most of the works QA simplified as a Classification Task. That is finally the system will be a classifier to classify the answer to the question. The number of different answers in the training set will be the number of different classes in the system.

The general pipeline of those models is that the word feature extracted from the question sentence is concatenated with the image feature extracted from the visual element, then they are pass into a soft max layer to predict the answer class. For obtaining the image features we generally use the VGG or ResNet [2] pre-trained model. For obtaining the text features here generally use the LSTM or GRU technique.

In the iBOWIMG [3] model, simply use naive bag-of-words as the text feature, and use the deep features from GoogLeNet [4] as the visual features. The input question is first converted into a numeric embedding vector like one hot vector. Then this one hot vector transformed to a word embedding layer to convert to the word vector. After that it is concatinated to an image vector like CNN. The combined feature is sent to the soft max layer to predict the answer class, which essentially is a multi-class logistic regression model.

possible. Proposed method in this paper, social pseudo relevance feedback (sprf), combines user feed- back with query expansion using the contextual vectors presented earlier. The contextual vectors are derived from user-generated content (i.e., tweets) and can think of those n-grams as explicit terms selected by users as votes in aggregate.

A. Benchmark Performance

From several comparisons and observation it has been found that the model performs well on RNN [5] on the VQA dataset, and the behaviour of the model is easily interpreted manner. Essentially, the BOWIMG baseline model learns how to memorize the relationship between the answer class and the informative words in the question or phrase along with the image feature. So split the learned weights of soft max into two parts, one part for the textual feature and the other part for the visual feature.

IV. BILINEAR CNN MODELS FOR FINE-GRAINED VISUAL RECOGNITION

Here [6] propose a model which has two feature extractor modules in its structure. The outputs of each feature extractor module is multiplied using the outer product. After that the generated output is pooled to obtain the actual result.

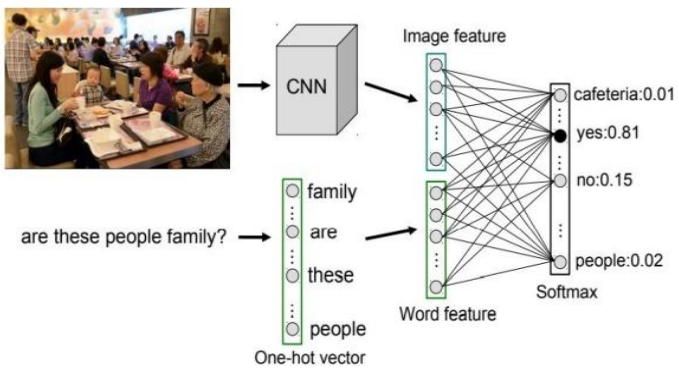

Fig. 2. Framework of the iBOWIMG

It is more helpful in fine grained classification. It can also handle the various textual descriptions such as VLAD and O2P. Here the Convolution Neural Networks (CNN) are used to obtain the features from the images. Fine-grained recognition classification is much more difficult to handle because the tasks such as identification of a special type of bird, or the categorization of the different types of Ants are complex tasks. Because the visual differences between the categories are small.

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

IJARCCE

International Journal of Advanced Research in Computer and Communication Engineering

PCON-2019

National Conference and Seminar on Innovations in Engineering & Technology

Govt Polytechnic College, Palakkad

Vol. 8, Special Issue 1, January 2019

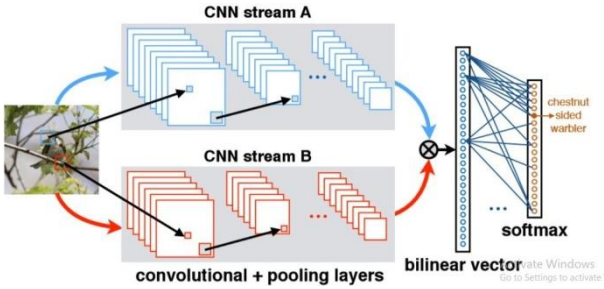

Fig. 3. Framework of the Fine-grained Visual Recognition

A. Methodology

Here first localize the various parts of the images and model these detected parts in the visual elements. A drawback of these approaches is that annotating parts is significantly more challenging than collecting image labels. Here it is taking various features of the images such as VLAD, SIFT, FOG. etc. For this purpose here using the pre-trained CNN models such as ImageNet and ResNet. The most important contribution is a recognition architecture that addresses several drawbacks of both part-based and texture models. The outer product between the two CNN outputs are capture the pair-wise correlations between the feature channels and can model part-feature interactions ,e.g., if one of the networks was a part detector and the other a local feature extractor.

B. Summarization Based Approach

A natural candidate for the feature function f is a CNN consisting of a hierarchy of convolutional and pooling layers. In this experiments here use CNNs [7] pre-trained on the Imnal layer including non-linearities as feature extractors. By pre-training we benefit from additional training data when domain specific data is scarce. This has been shown to be beneficial for a number of recognition tasks ranging from object detection, texture recognition, tone-grained classification.

V. FISHER VECTORS DERIVED FROM HYBRID GAUSSIAN-LAPLACIAN MIXTURE FOR IMAGE ANNOTATION

In general visual classification tasks the normal procedure is extracting the features of the images and then pooled to a classification layer. But latest trends shows that Fisher Vectors are more expressively handle such correlations than the normal feature vectors. The fisher vector [8] is taken as the gradients of the log-likelihood of descriptors, with respect to the parameters of a Gaussian Mixture Model(GMM) [8].

Here present two other Mixture Models and derived their Expectation Maximization and Fisher Vector expressions. The first one is Laplacian Mixture Model (LMM) [9] which is based on the Laplacian distribution. The second one is a Hybrid Gaussian-Laplacian Mixture Model (HGLMM) which is based on a weighted geometric mean of the Gaussian and Laplacian distribution.

The standard pipeline of Image recognition task is generally going through the main 3 stages.

- The first step is extracting the various image features such as SIFT,FOG ..etc from the image.
- The second step is transforming the extracted features to the BoW model or the numerical equivalence of the image vectors.
- The third step is vector representation is usually passed to a classier that is trained by a suitable machine learning algorithm, e.g., Neural Networks.

Here describes the suggested modification in the pooling step(Second step).

A. Methodology

In all of these contributions, the Fisher Vector of a set of local descriptors is obtained as a concatenation of gradients of the log-likelihood of the descriptors in the set with respect to the parameters of a Gaussian Mixture Model that was TTED on a training set in an unsupervised manner. Different methodologies are proposed in the Fisher vectors but all of them are in the context of the Gaussian Mixture Model.

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

IJARCCE
International Journal of Advanced Research in Computer and Communication Engineering
PCON-2019
National Conference and Seminar on Innovations in Engineering & Technology
Govt Polytechnic College, Palakkad
Vol. 8, Special Issue 1, January 2019

B. Discussion

The Fisher Vector is proven useful in variety of the visual recognition tasks. Most of them based on the Gaussian Mixture Model. So in this work here discussing how the Fisher Vector derived from other distributions, such as , LMM and HGLMM. So it shows a significant improvement in the computer vision tasks. The Hybrid Gaussian-Laplacian Mixture Model (HGLMM) that we presented, allowed us to gain benefits from both under- lying distributions by having the edibility that each dimension in each component would be modelled according to the most suitable distribution. Such geometric-mean mixtures could be generalized to any two distributions and t real world data of any distribution shape. It is also not limited to just two parametric distributions.

VI. COMPARISON OF DIFFERENT METHODOLOGIES

Table 1 shows the different methodologies and their features in visual question answering that we discussed in the above sections.

Table I : Comparison Table

Title	Method Used	Advantage	Limitation
Simple Baseline for Visual Question Answering	Multimodal pooling	1) Attention mechanism 2) Simple interpretation	Lack of memorizing the correlations to actual reasoning and understanding of the question and image
Bilinear CNN Models for Fine grained Visual Recognition	Bilinear pooling	Fine-grained visual recognition task	Using two CNN models for the purpose
Fisher Vectors Derived from Hybrid Gaussian- Laplacian Mixture Models for Image Annotation	Joint multimodal embedding	Capture different interactions	Elements interact on higher dimensional space
Multimodal Compact Bilinear Pooling for Visual Question Answering and Visual Grounding	Multimodal compact Bilinear pooling	Multiplicative interaction between all elements of both vectors	Space and time complexity is higher

VII. CONCLUSION AND FUTURE SCOPE

Visual Question Answering is one of the recent trending field in Machine learning and Natural Language Processing. This is the combination of both Natural language and computer vision techniques. Computer vision for analysing the image and NLP to analyse and generate the text. There are several different kinds of techniques we are using for the VQA. Some of them are:

- Simple Baseline for Visual Question Answering
- Bilinear CNN Models for Fine-grained Visual Recognition
- Fisher Vectors Derived from Hybrid Gaussian-Laplacian.
- Multimodal Compact Bilinear Pooling for Visual Question Answering and Visual Grounding Among the above methods Multimodal Compact Bilinear Pooling [10] technique shows the most feasible results.

Some of the suggested further improvements are given below:

- Answering questions which require world knowledge to find out the answer. The VQA systems should able to analyse the questions which require world knowledge to find out the answer.
  Eg :-Which team is playing the cricket in the image.

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

IJARCCE

International Journal of Advanced Research in Computer and Communication Engineering

PCON-2019

National Conference and Seminar on Innovations in Engineering & Technology

Govt Polytechnic College, Palakkad

Vol. 8, Special Issue 1, January 2019

- Extract the texts from the image. The VQA systems should able to extract the texts from the image. Eg :-What is written in the blackboard in the picture.
- Ability to analyse some loigc questions. The VQA systems should able to analyse some loigc questions. Eg :-How the two persons in the picture are relate to each other.

REFERENCES

[1]. S. Qu, "Visual question answering using various methods,"
[2]. C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning.," in AAAI, vol. 4, p. 12, 2017.
[3]. B. Zhou, Y. Tian, S. Sukhbaatar, A. Szlam, and R. Fergus, "Simple base- line for visual question answering," arXiv preprint arXiv:1512.02167, 2015.
[4]. Z. Zhong, L. Jin, and Z. Xie, "High performance offline handwritten chinese character recognition using googlenet and directional feature maps," in Document Analysis and Recognition (ICDAR), 2015 13th International Conference on, pp. 846–850, IEEE, 2015.
[5]. T. Mikolov, M. Karafia´t, L. Burget, J. C˘ ernocky`, and S. Khudanpur, "Recurrent neural network based language model," in Eleventh Annual Conference of the International Speech Communication Association, 2010.
[6]. T.-Y. Lin, A. RoyChowdhury, and S. Maji, "Bilinear cnn models for fine- grained visual recognition," in Proceedings of the IEEE International Conference on Computer Vision, pp. 1449–1457, 2015.
[7]. A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in Advances in neural information processing systems, pp. 1097–1105, 2012.
[8]. B. Klein, G. Lev, G. Sadeh, and L. Wolf, "Fisher vectors derived from hybrid gaussian-laplacian mixture models for image annotation," arXiv preprint arXiv:1411.7399, 2014.
[9]. N. Mitianoudis and T. Stathaki, "Overcomplete source separation using laplacian mixture models," IEEE Signal Processing Letters, vol. 12, no. 4, pp. 277–280, 2005.
[10]. A. Fukui, D. H. Park, D. Yang, A. Rohrbach, T. Darrell, and
[11]. M. Rohrbach, "Multimodal compact bilinear pooling for visual question answering and visual grounding," arXiv preprint arXiv:1606.01847, 2016.