**Report of the VQA Model and Train**

Group Project Ⅲ

**Visual Question Answering**

Supervisor

郑锋

Team Members

11813226 张一弛

11812518 杨英特



catalogue

[Abstract 3](#_Toc95742095)

[Introduction 3](#_Toc95742096)

[VQA 3](#_Toc95742097)

[Model of VQA 4](#_Toc95742098)

[Process 4](#_Toc95742099)

[Word segmentation 4](#_Toc95742100)

[Word Embeddings 5](#_Toc95742101)

[Bert 6](#_Toc95742102)

[image feature extraction 7](#_Toc95742103)

[Model training 8](#_Toc95742104)

[Training and test 9](#_Toc95742105)

[Evaluation and analysis 9](#_Toc95742106)

[Reference 10](#_Toc95742107)

# Abstract

In the first reply, we completed the construction of the data set, and learned some problems of Chinese VQA through the construction of the data set. Then, we choose to use the classic VQA model of CNN and LSTM, and start to complete the extraction of semantic and image features after parsing the model. And input it into VQA model for training. This series of work is divided into four steps: word segmentation, word embedding, image feature extraction and model training.

# Introduction

## VQA

VQA is a subject that integrates computer vision and natural language processing through deep learning. This topic attempts to make machine learning answer questions related to pictures. At present, there are quite a lot of research in the field of VQA, and some teams have explored quite cutting-edge in the field of VQA. However, the current VQA focuses on the English semantic environment. We hope to build a feasible Chinese VQA model, so that understand the differences between Chinese and English VQA. And we plan to design a VQA application scenario in Chinese context to better apply VQA.

## Model of VQA

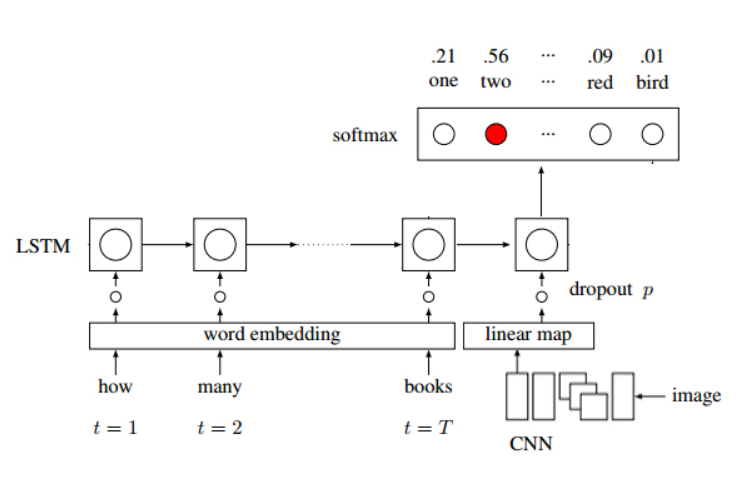


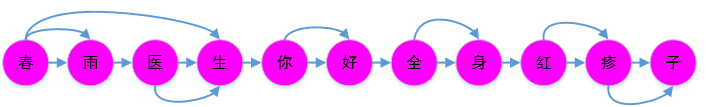
Figure 1: model of VQA

The model we use refers to paper *Exploring Models and Data for Image Question Answering(by Mengye Ren, Ryan Kiros & Richard Zemel)*. We input the question into an LSTM network to get the semantics of the question. At the same time, we use CNN to extract the image features, then splice the two and input them into the activation layer to get a definite answer.

# Process

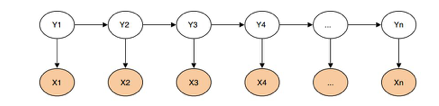
## Word segmentation

**Jieba**



Future 2.1: Jieba

It belongs to the word table participle in the participle. Jieba is mainly based on the search algorithm of directed acyclic graph. Through dynamic programming, the joint probability of word cutting combination is maximized from back to front.



Future 2.2: HMM

At the same time, Jieba will use HMM new word recognition. In word segmentation, the input character is regarded as an observation sequence, and the output sequence is represented by "BEMS". Where, B represents the starting word in the word, M represents the middle word in the word, e represents the ending word in the word, and s represents the formation of a single word. "Dr. Chunyu / hello / whole body / rash / son" can be expressed as bmme / be / be / s. therefore, HMM word segmentation is to predict the BEMS values of all characters.

**THULAC 2016**

THULAC (Thu lexical analyzer for Chinese) is a set of Chinese lexical analysis toolkit developed and launched by the laboratory of natural language processing and social humanities computing of Tsinghua University, which has the functions of Chinese word segmentation and part of speech tagging. Thulac has the following characteristics:

Strong ability. It is trained by using the largest artificial word segmentation and part of speech tagging Chinese corpus (about 58 million words) in the world, and the model has strong tagging ability.

High accuracy. On the standard dataset Chinese treebank (ctb5), the F1 value of word segmentation and part of speech tagging can reach 97.3% and 92.9%, which is equivalent to the best method on the dataset.

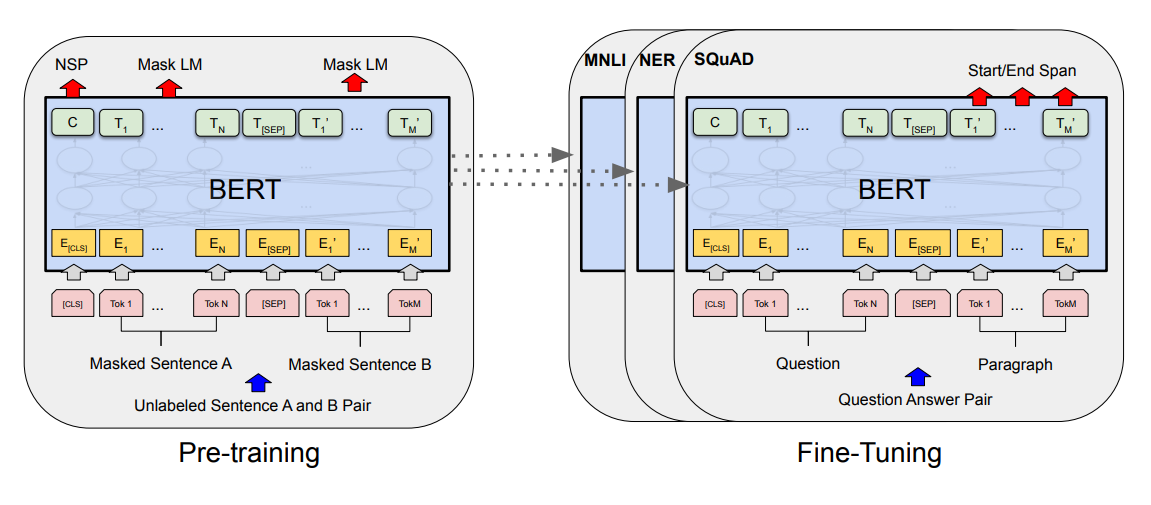
Faster. The speed of word segmentation and part of speech tagging is 300KB / s, and it can process about 150000 words per second. The speed of word segmentation can reach 1.3mb/s.

## Word Embeddings

We read the paper “Revisiting Correlations between Intrinsic and Extrinsic Evaluations of Word Embeddings”, this paper presents the first study on the correlation between results of intrinsic evaluation and extrinsic evaluation with Chinese word embeddings. The author use word similarity and word analogy as the intrinsic tasks, Named Entity Recognition and Sentiment Classification as the extrinsic tasks. A variety of Chinese word embeddings trained with different corpora and context features are used in the experiments.

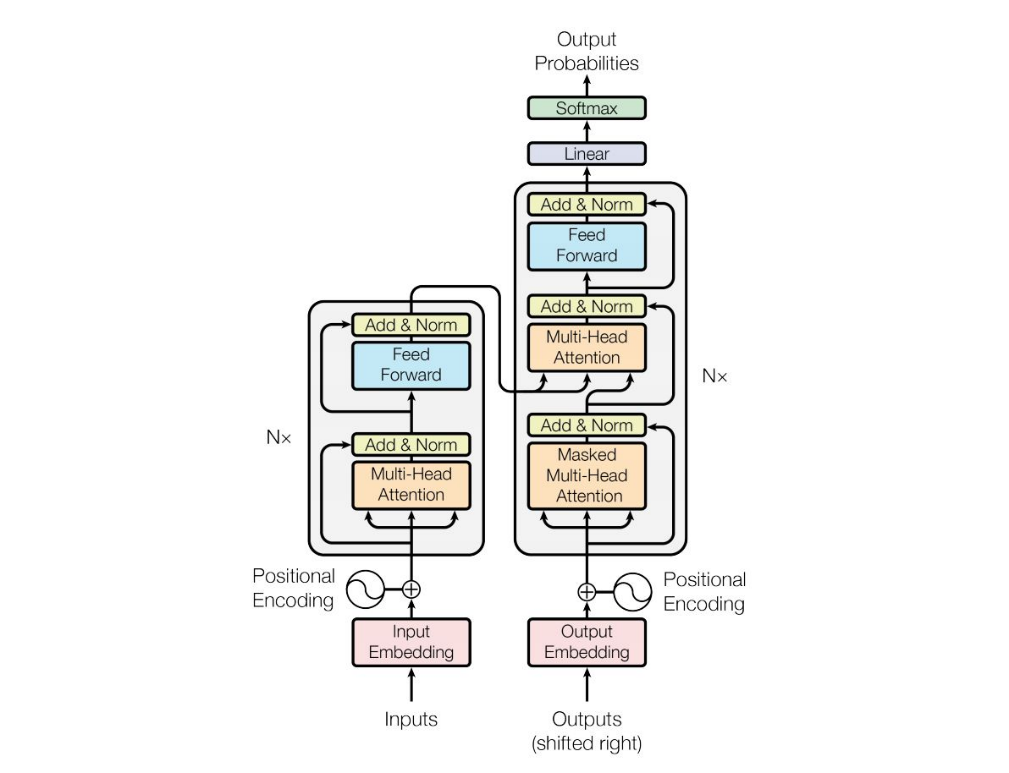
This project provides 100+ Chinese Word Vectors (embeddings) trained with different representations (dense and sparse), context features (word, ngram, character, and more), and corpora. One can easily obtain pre-trained vectors with different properties and use them for downstream tasks. And we used it in our VQA model.

## Bert

 Future 1.1: Bert

BERT, or **B**idirectional **E**ncoder **R**epresentations from **T**ransformers[3], is a new method of pre-training language representations which obtains state-of-the-art results on a wide array of Natural Language Processing (NLP) tasks.

The structure of the previous pre training model will be limited by the one-way language model (from left to right or from right to left), which also limits the representation ability of the model, so that it can only obtain one-way context information. Bert uses MLM for pre training and adopts deep bidirectional transformer components to construct the whole model, so as to finally generate a deep two-way language representation that can integrate left and right context information.

.

Future 1.2: Transformer

Transformer[2] is a model mentioned in the *Attention is all you need*. The architecture of the Transformer is some encoders and decoders. There are two basic conceptions: Sequence to Sequence (Seq2seq), and Encoder-Decoder. The Seq2seq is just a concept that emphasizes on purpose more: input a sequence then output a new sequence, which the lengths of the sequences are unknown. In NLP, sentences maybe have different length, use Seq2seq approaches can solve problems flexibly. And one of the Seq2seq approaches is Encoder-Decoder Model.

Finally, we use Bert's word embedding as a new QA extraction and embedding.

## image feature extraction

We use VGG16 model to extract image features.

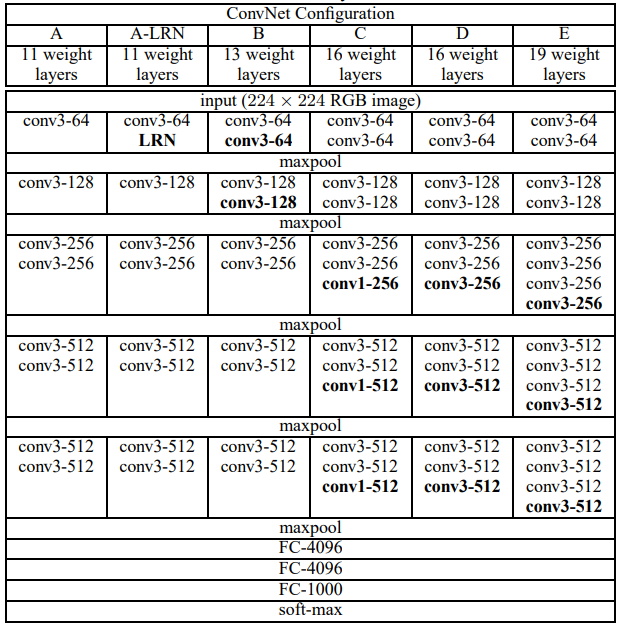


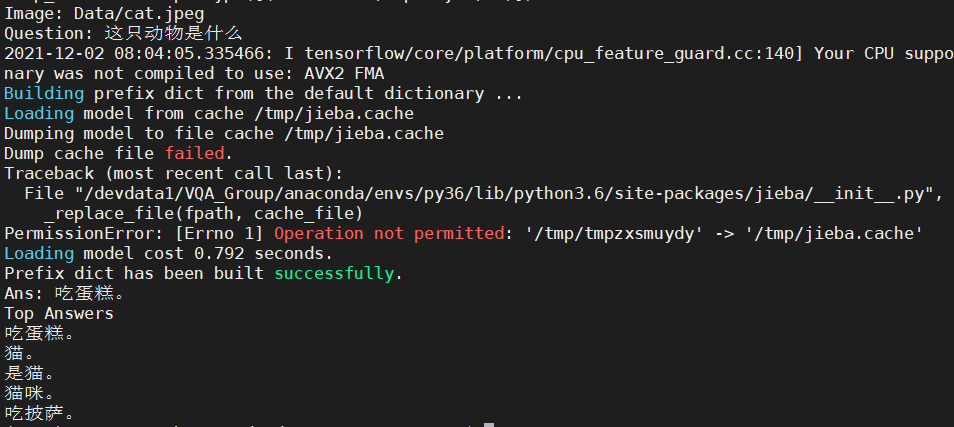
Figure 2.4: Model of VGG16

The default size of VGG net for the input image is 224 \* 224 \* 3 (as can be seen from the Figure 2.4). The VGG16 network structure has 16 layers with parameters, including 13 convolution layers, 5 pooling layers and 3 full connection layers, excluding the activation layer.

As this figure shown, VGG16 network structure can be divided into six module levels plus one input module. By repeatedly convoluting and pooling the pictures, the three channel pictures are finally transformed into one-dimensional vectors to output the characteristics of the pictures.

## Model training

Finally, we put the processed data into the designed VQA model for training.



Future 3: Result of VQA

As shown in the picture, three of the five most likely answers can accurately answer our questions. VQA in Chinese can meet the expectations. It can also be found that we need to improve the accuracy and build a verification set to measure our VQA model

## Training and test

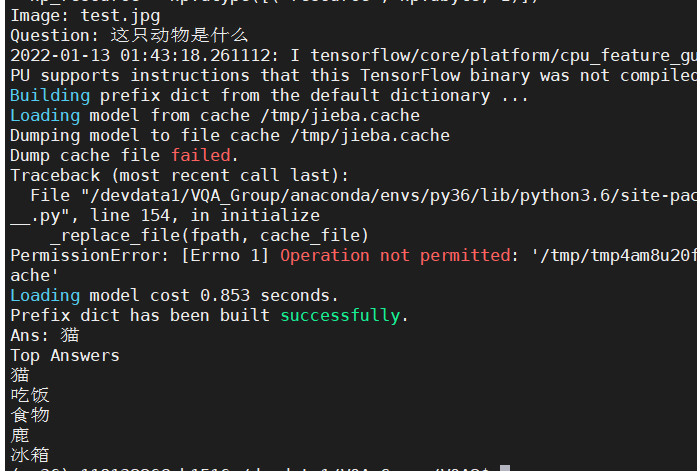
The data set is still the data set we make.

We still use VGG16 model to extract image features.

The words are embedded using Bert.

## Evaluation and analysis

Finally, we run the new designed VQA model.



Future 3: Result of same question in VQA2.0

(Other attempts are omitted in this article)

As can be seen from the figure, multiple words with the same meaning have been integrated into the same answer, and the improvement in the part of word embedding is successful.

At the same time we for many complex scene carried out quite a lot of tests, which, in the judgment on some of the more common and obvious things, more accurate, but in some need a certain combination, logic, a mistake in the two aspects of the VQA(e.g. *How many people are there*), eliminate the factors of word embedded can determine the images and deal with the problem of the model.

# Reference

**[1] Zhou, B. , Tian, Y. , Sukhbaatar, S. , Szlam, A. , & Fergus, R. . VQA: Visual Question Answering.**

**[2] Ren, M. , Kiros, R. , & Zemel, R. . (2015). Exploring models and data for image question answering. Litoral Revista De La Poesía Y El Pensamiento, 2953-2961.**

**[3] Simonyan, K. , & Zisserman, A. . (2014). Very deep convolutional networks for large-scale image recognition. Computer Science.**

**[4] YuanyuanQiu, HongzhengLi, ShenLi, YingdiJiang, RenfenHu, & LijiaoYang. (2018). Revisiting correlations between intrinsic and extrinsic evaluations of word embeddings.**

**[5] Zhongguo Li, Maosong Sun. Punctuation as Implicit Annotations for Chinese Word Segmentation. Computational Linguistics, vol. 35, no. 4, pp. 505-512, 2009.**

**[6] Vaswani, A. , Shazeer, N. , Parmar, N. , Uszkoreit, J. , Jones, L. , & Gomez, A. N. , et al. (2017). Attention is all you need. arXiv.**

**[7] Devlin, J. , Chang, M. W. , Lee, K. , & Toutanova, K. . (2018). Bert: pre-training of deep bidirectional transformers for language understanding.**

**[8] Zhang, Z. , Han, X. , Liu, Z. , Jiang, X. , Sun, M. , & Liu, Q. . (2019). ERNIE: Enhanced Language Representation with Informative Entities. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics.**

**[9] Sun, Y. , Wang, S. , Li, Y. , Feng, S. , & Wu, H. . (2019). Ernie: enhanced representation through knowledge integration.**