A Joint Model for Text and Image Semantic Feature Extraction

Jiarun Cao
Digital Media Research Institute
Beijing Institute of Technology
Beijing, China
Jiaruncao.china@Gmail.com

Chongwen Wang†
Digital Media Research Institute
Beijing Institute of Technology
Beijing, China
wcwzzw@bit.edu.cn

Liming Gao
Digital Media Research Institute
Beijing Institute of Technology
Beijing, China
649386435@qq.com

ABSTRACT

Most of the current information retrieval are based on keyword information appearing in the text or statistical information according to the number of vocabulary words. It is also possible to add additional semantic information by using synonyms, polysemous words, etc. to increase the accuracy of similarity and screening. However, in the current network, in addition to generate a large number of new words every day, pictures, audio, video and other information will appear too. So the manual features are difficult to express on the this kind of newly appearing data, and the low-dimensional feature abstraction is very difficult to represent the overall semantics of text and images. In this paper, we propose a semantic feature extraction algorithm based on deep network, which applies the local attention mechanism to the feature generation model of pictures and texts. The retrieval of text and image information is converted into the similarity calculation of the vector, which improves the retrieval speed and ensures the semantic relevance of the result. Through the compilation of many years of news text and image data to complete the training and testing of text and image feature extraction models, the results show that the depth feature model has great advantages in semantic expression and feature extraction. On the other hand, add the similarity calculation to the training processing also improve the retrieval accuracy.

CCS CONCEPTS

• Computing methodologies→ Artificial intelligence→ Natural language processing →Information extraction

KEYWORDS

Natural language processing, Information retrieval, Similarity Calculation

1 Introduction

Nowadays, with the development of network, computing technology, and multimedia data such as text, image and audio, different search technologies have been appeared, for instance, full-text search for text content itself, retrieval based on language model and depth feature-based search methods, etc. As the difference of search object, the IR techniques also can be classified into: retrieval of sequences such as text and audio, and retrieval of image files such as picture videos.

In the text retrieval, the expressed text or vocabulary is converted into a vector, and the obtained information is filtered by means of vector comparison[1,2]. In this way, the textual representation information is converted into a discrete vector representation, thus the semantic similarity can be represented by the distance in the vector space. And the introduction of semantic information can solve the problem of insufficient representation of manually-designed features. The combination of word vector mapping and multi-grammar model is used in semantic generation to free the semantic relationship from the original manual feature design and the preset knowledge collection, so that it can rely on the corpus resource training generation[3]. Multi-level convolution operations on semantic representations can generate models more efficiently and accurately, and using the newly-entered data to fine-tune on the basis of the original model can quickly adapt the model to new vocabulary or syntax. However, since the convolution size cannot be changed, therefore, only a fixed-length syntax rule can be implemented. It makes the influence of the preceding vocabulary when the length of the statement exceeds the convolution size cannot be passed to the subsequent statement feature generation, so its representation ability cannot cope with the text which is too long. In the RNN (Recurrent Neural Network), the input vector is cyclically updated according to the same network element, and the subsequent iterative update of the unit is used to predict the subsequent possible vocabulary[4]. In this update mode, the weight information of the intermediate network layer can be affected by the whole text, so that expression of the syntax information is more accurate.

In the search for image content, one of these methods is searching for related images by means of graph search [5], the other is using a descriptive statement to filter the images. In the picture tag generation field, there are also some kinds of competition such as using ImageNet as the dataset to classify pictures, and the classification or recognition of unknown data is predicted by providing a large amount of annotation data [6]. With the advantages of a large annotated data set and a multi-layer network model, the accuracy rate is continuously increased, and even exceeding the human accuracy rate [7]. Therefore, the

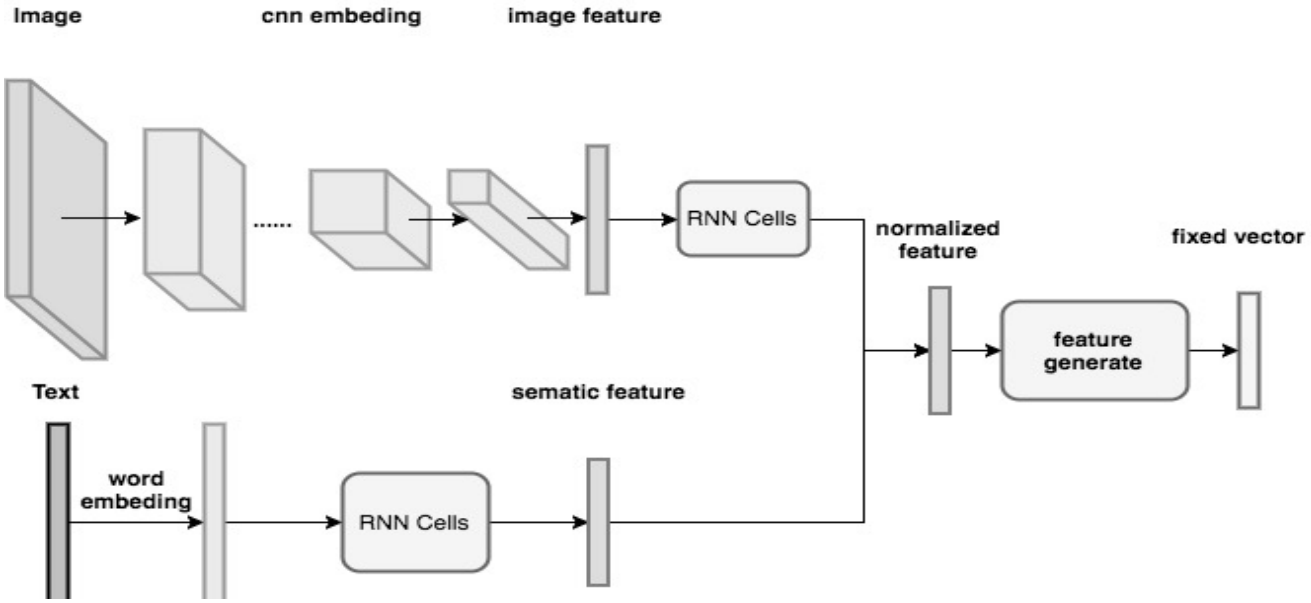

Figure 1 Model Structure Overview

image tag generation can quickly and accurately establish the index item of the image. However, this method is limited by the dataset and cannot generate an accurate label for the content beyond the coverage of the dataset. To break the range of intrinsic label expression, transforming the probability regression problem into the sequence feature classification is one of the current main solution. Specifically, the attention mechanism of making full use of the state transfer between the upper and lower layers is the dominant way to solve the problem of sequence generation. By combining the high response area in the picture with the picture information, the decoder can acquire the target information and the relationship between them, thereby generating natural language description information [8, 9].

In the information search, not only the target files need to be filtered according to the search conditions, but also sorted their relevance to get the final result . The model directly uses the twin network to calculate the similarity between two targets. This kind of structure can put the feature extraction and similarity calculation into the same model, and the selection of the feature merge layer position can retain the original to some extent[10]. This approach is suitable for similarity calculation problems that require a certain degree of positional relationship. Another way is to use the distance of the vector as a loss function in training, and the distance calculation method of the traditional vector space is the basis of the similarity measure [11].

This paper studies the feature extraction and semantic representation of texts and pictures. With the current research and development trends, this paper proposes a method of file retrieval and sorting, which combines the similarity calculation and the semantic feature of the image data. We utilize a independent feature extraction structure for different information representations of text and images. Moreover, to facilitate the measure of similarity, we convert features generated by different different information into a unified form and train together with the similarity measure function. According to the experiments in part three, our method gets a effective result only using a small amount of annotated data.

2 Model Structure

2.1 Overview

The feature extraction structure of text mainly contains three parts. As shown in Figure 1: image semantic extraction, text semantic extraction, and normalized feature generation. Since the difference between information representation of text and images is so large, it cannot process directly using the same structure, thus we use different feature extraction methods for both images and texts. This structure can be used to extract the semantic content of the documents and generate tags for the classification search. In addition, the same network structure uses different loss functions is able to perform different functions. By referring to similarity calculations and small sample generalization, the network has the ability to distinguish between different samples.

2.2 Text Feature Extraction

In order to accurately locate the target file when searching the text, in addition to using keywords, to extract the feature, RNN is also a effective way, which is able to increase the length of dependence between textual information during feature extraction, and make the representation of semantic features more accurate. At the same time, set a fixed value mapping before input for each vocabulary, and then generate the word vector by the mapping matrix in the network ,it helps the state of different network nodes in the model to represent different inputs, and the vector transformation including the semantic features. According to this

way, the word vector contains the semantic relationship between words.

The attention mechanism can automatically filter the key information to compress the output text .It helps to generate the fixed-length vector while maintaining the original semantics to the greatest extent. The same network element is used in feature extraction to process the input of sequence information, and the resulting state information is retained to combine with the next set of inputs to generate new state information. the final structure is a multi-layer network in a chronological order by expanding the processing of a set of data.

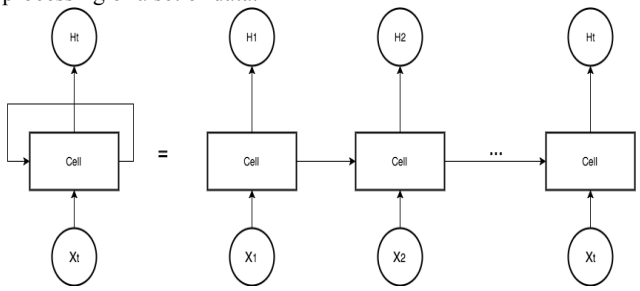

Figure 1 RNN Structure

After the word vector generation module, we add LSTM to extract the semantic information model structure, which is formed by the sequential combination of vocabulary. As shown in Figure 2, LSTM has the ability of feature abstraction, which helps to extract the semantic feature of whole sentence and finish the transformation from text to vector. The overall structure of the model is shown in Figure 3.

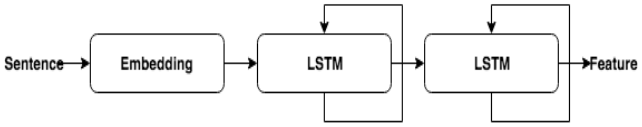

Figure 2 Semantic and Word2Vec Future Training

2.3 Image Feature Transformation

In order to implement retrieval of images, it is not only necessary to extract abstract features that can represent information in the image, but also to extract information with certain natural semantics. Image classification is one of the basic methods for extracting information contained in a picture. By using the model for training of picture classification problems, the model can have the ability to extract picture information. Based on the image classification model, the location and quantity information of image can be obtained by means of target detection, only in this way can we acquire more abundant information.

Directly using the classification model for feature extraction can achieve the image2image search method, but it is impossible to complete the semantic condition search because of the huge gap between its representation and natural language. We extract he image information of different kinds and scales to form a set of discrete vocabulary information, and then utilize the language model to convert the discrete vocabulary into a concrete expression statement, thereby we achieve the conversion of the image content to the text content.

Referring to the function of word bag in character feature extraction, we use the convolution layer as the encoding layer of the image, use the hidden layer feature of the image classification model as input of the RNN. Since the information of the category, the state of the object and the relationship between the original positions have been retained , it provide a richer information for the final generation of the text description. Moreover, we use the convolutional layer to extract the high-dimensional abstract features of the image, and regard the result as a input of the decoding layer composed of multiple LSTMs. The network structure for extracting the semantic information is as shown in Figure 4 .

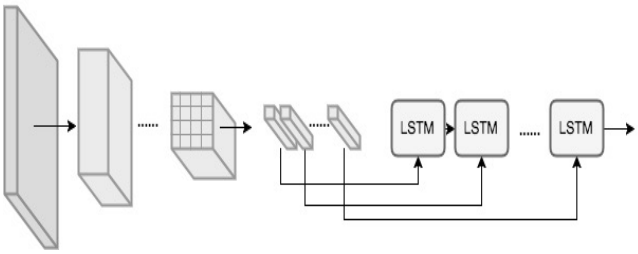

Figure 3 Image Semantic Feature Extraction

2.4 Semantic Similarity Measurement

In the field of information retrieval, in addition to filter the target files to meet the search criteria, it is also necessary to sort the results based on the similarity. When calculating the similarity, the target to be compared is often normalized to the same size by feature extraction. When evaluating the similarity between two vectors, a similarity algorithm with multiple different granularities is available, in this paper, we adopt the cosine similarity. The cosine similarity (Equation 1) considers the combination of values, and eliminates the difference in size between calculations. It makes the whole result more focused on the features produced by the combination of values between vectors.

$$d = \frac{\sum_{k=1}^{n} x_{1k} x_{2k}}{\sqrt{\sum_{k=1}^{n} x_{1k}^2 x_{2k}^2}} \quad (1)$$

After the feature extraction model pre-training, the information contained in the hidden layer can be used to generate a short semantic description, thereby ensuring that the hidden layer contains the overall semantic information, on the other hand, it keeps the value in the same range by adding the normalization parameter.

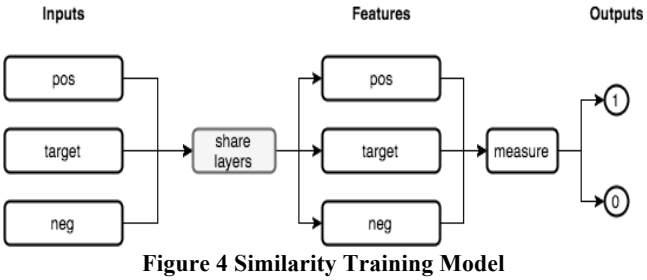

Figure 4 Similarity Training Model

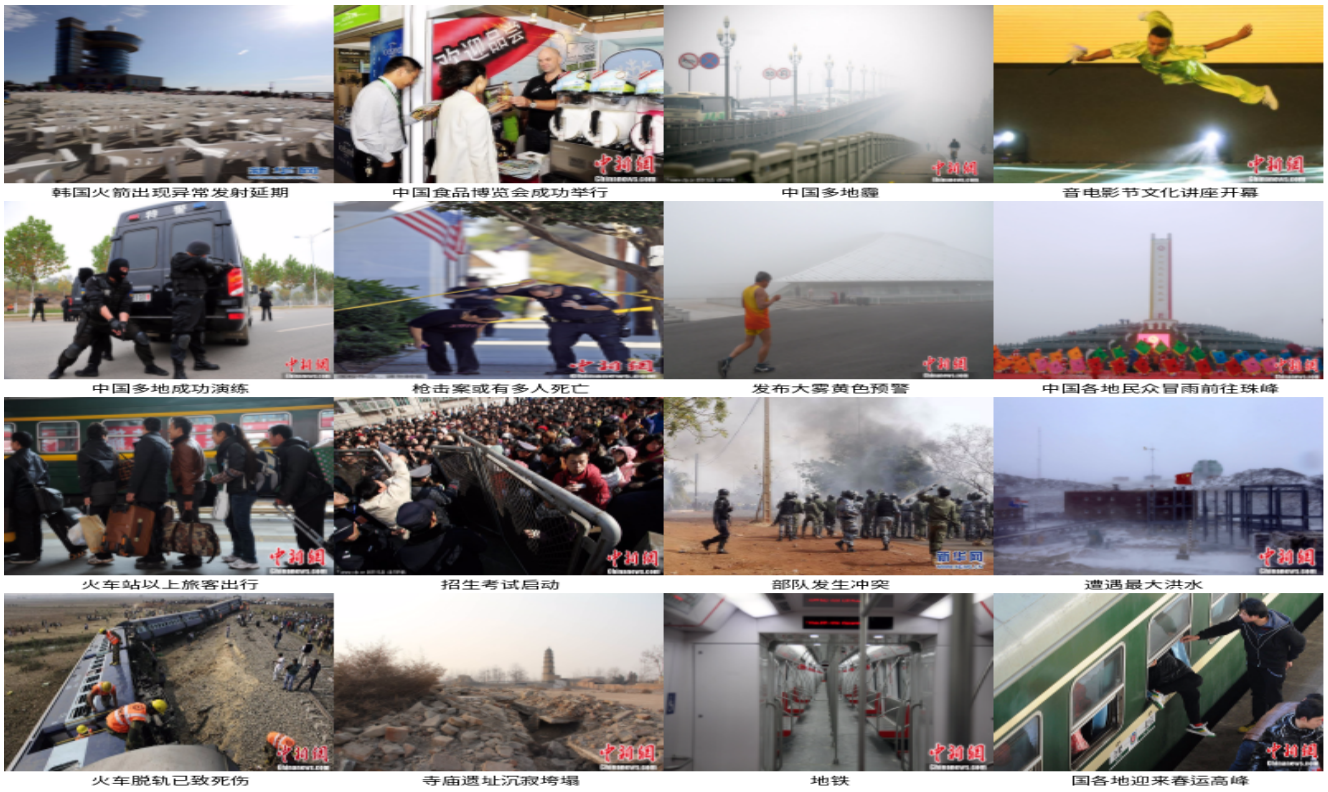

Figure 8 Transform Image to Text

As the network structure is shown in Fig.5, at each time, the inputs are a reference value and a positive and negative example selected according to the similarity, and the share layer is used to extract the feature vector, then the similarity is evaluated using the distance metric.

In this way, although the final specific distance metric is not constrained, the model still learned the ability of distinguishing semantic features after the training process.

3 Experiment

2017-03-16 #@# 八旬老太"开挂" 雨夜撬铁丝网翻墙逃出养老院 #@# （原标题：思乡心切
2017-03-20 #@# 济南街头上演“公交婚礼” 新郎自驾公交去迎亲 #@# 19日，山东济南春光
2017-03-16 #@# 外媒：法国一所学校发生枪击事件 造成多人受伤 #@# 中新网3月16日
2017-03-21 #@# 蔡英文主持潜艇自造签约 突然怪风将合同吹走 #@# [摘要]蔡英文在致辞
2017-03-20 #@# 日本大学生长春听讲“南京大屠杀” 被真相震惊 #@# 中新社长春3月20日
2017-03-18 #@# 女子花万元包下全城900多辆出租车 向男友求婚 #@# 这几天，一组照
2017-03-20 #@# 公务员丈夫长期家暴记者妻子将其打死 被判处死刑 #@# 法制晚报快讯
2017-03-16 #@# 法国一高中发生枪击事件，造成多人受伤 #@# 据法国媒体报道，法国南
2017-03-21 #@# 乐天免税店销售额锐减25% 韩国免税店开拓东南亚市场 #@# 资料图：
2017-03-16 #@# 南京一女子地铁遭遇“咸猪手” 一路追到男厕堵门抓人 #@# 现代快报
2017-03-18 #@# 岳阳一男子杀人逃17年变身千万富豪 已被刑拘 #@# 17年前，他因赌博抢
2017-03-21 #@# 甘肃：今起我省大部有一次降水降温大风天气过程 #@# 今起我省大部有
2017-03-20 #@# 港媒关注中国火箭回收技术：与美不同走伞降路线 #@# 资料图：Space
2017-03-18 #@# 【人生回眸】一片冰心在玉壶 #@# 吴文藻与冰心的爱情，可谓因错结缘
2017-03-20 #@# 香港男子走私仿真枪案重审开庭 一审曾判7年 #@# 香港人陈智勇走私武
2017-03-18 #@# 工程改造请和尚做法事？ 涉事国企：这个锅不背 #@# 近日有媒体报道
2017-03-20 #@# 菲总统称欢迎我海警船停靠：美军都行中国为啥不行？ #@# 资料图：
2017-03-16 #@# 耐克中国售卖缺少气垫产品涉嫌虚假宣传 已被立案调查 #@# 针对央视
2017-03-18 #@# 忆吴文藻先生：此去不缘名利去 万里归心对月明 #@# 他认为社会学要
2017-03-18 #@# 韩国会第一大党呼吁暂停萨德部署：应取得中俄同意 #@# 资料图：在韩国
2017-03-16 #@# 国际货币基金组织巴黎办公室发生邮件爆炸事件 #@# 新华社快讯：据
2017-03-20 #@# 票贩伙同售票人员倒卖门票?颐和园:涉事人被停职 #@# 新京报快讯(记
2017-03-21 #@# 日媒：安倍与奥朗德谈深化海洋安保合作牵制中国 #@# 资料图：日本军

Figure 6 Exhibition of Partial Dataset

In the experiment of extracting text, we select the news content as training data, the news headline as label. We train the news content to fit the summary information of different contents. We have compiled almost 200,000 news data as our dataset, and some part of dataset are shown in Figure 6.

In the semantic feature extraction, we use the sequence-to-sequence model to obtain the extraction of text abstract semantics. In order to supervise the extraction of semantic features in the hidden layer, we add a new encoding layer to fit the output of the preliminary encoding layer.

During the experiment, the gradient was slowly descended the training process, so we use no gradient decay. We use a fixed learning rate in the first 20,000 iterations of the training. We start to add gradient decay after the learning rate is severely fluctuated. The final loss curve are shown in Figure7.

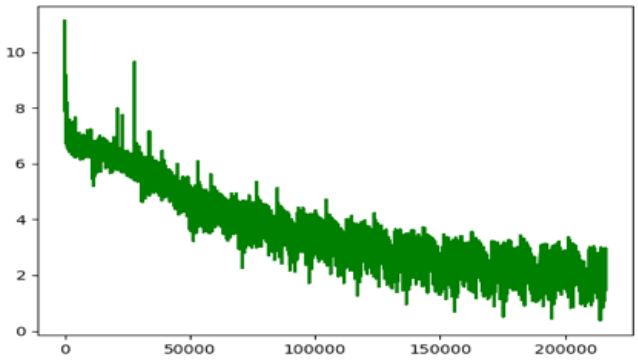

Figure 7 Loss Curve

Figure 10 SEARCH IMAGE BY IMAGE: the first input image and its corresponding title information is listed on the left, and the similar image and its similarity are shown on the right.

发生爆炸	发生5.9级地震 0.70	境附近发生3.4级左右地震 0.69	震地震 震源深度10公里 0.68	县附近发生4.7级左右地震 0.66
FBI首次承认：正调查特朗普与俄罗斯是否"勾结"	特雷莎谈与特朗普手拉手："只是他的绅士行为" 0.80	菲律宾总统再怼奥巴马 盛赞特朗普为"务实派" 0.80	白宫回应"特朗普通俄"：指控没有任何证据 0.80	美联邦调查局调查俄干预美大选 特朗普发文反击 0.80
广州深夜出楼市新政：单身限购一套 连缴社保3年变5年	统计局：2月一二线城市新建商品住宅价格涨幅回落 0.82	北京发布二套房新政 500万总价款可贷金额减50万 0.83	国家统计局：一线城市新房价格同比涨幅连降5个月 0.81	北京楼市调控再加码 专家：未来将有6-9个月萧条期 0.80
一代传奇！美国亿万富豪洛克菲勒去世 享年101岁	79岁亿万富豪罗斯成美国最年长商务部长 0.85	生前住10平方米房子 95岁中科院院士徐祖耀逝世 0.74	105岁老人全家共96口人 年龄相加达3850岁 0.72	美国亿万富豪洛克菲勒去世 享年101岁 0.94
跨16省助考舞弊产业链：试卷印刷点人员偷拍泄密	云南下发通知：重大突发事件最迟需5小时内发布信息 0.76	湖南通报祖保煤矿10死2伤爆炸：违规组织生产是元凶 0.77	福建一施工桥梁掉落致7人受伤：5名工人坠落 0.70	广西：9级大风吹翻两艘渔船 11名渔民全部获救 0.71
南苏丹小型客机坠毁 时天气较差 至少14人受伤	一艘载有11人渔船在舟山海域沉没 目前5人获救 0.83	兰州城关区超载货车连撞3车 致10人受伤 0.83	一艘载有索马里人船只遭直升机攻击 造成31人死亡 0.78	福州一公交车行驶中撞上行道树 导致16人受伤 0.73
郭金龙：坚决打好蓝天保卫战碧水攻坚战突出问题歼灭战	李克强设专项资金：集中攻关打一场雾霾"歼灭战" 0.72	人民日报：城镇化不能只重面子 要防止过度城镇化 0.74	教育部：短期内难消除校园欺凌 0.74	人民日报：城镇化不能只重面子 要防止过度城镇化 0.74
男子山顶摆拍坠崖被批作死 当事人回应	武汉一公交车司机劝阻抽烟被捅伤 嫌疑男子被刑拘 0.89	三男子网上相约银川自杀 一人念及父母悔悟报警 0.80	司机驾车冲撞交警被制服后死亡 成都警方回应 0.86	兰州多名行人闯红灯遭劝阻 殴打辱骂交警被拘留 0.82
北京市教委："过道学区房"不能作为入学资格条件	政协委员：购买"孝亲房"赡养老人应该减免税 0.83	人大代表：房地产税不是为了把楼市价格"砸"下去 0.80	政协委员谈"房住不炒"：房价永远涨的神话要打破 0.80	人大代表高西庆：监管不是确保股市"只涨不跌" 0.79

Figure 11 SEARCH TEXT BY TEXT: the first category on the left is the input text title, and the right is the corresponding similar news title and its similarity

原始标题	生成标题
索马里首都摩加迪沙发生爆炸	称首都机场发生交火 目击者称爆炸原因尚不清楚
国产卫星高警一号传回新图：西安钟楼清晰可见	_UNK 卫星进行地面空间地面空间可满足地面
新加坡真的危机四伏吗？总统李显龙罕见发声	专家称新加坡在中美间感觉是美国 应警惕

Table 1 Sample of Outputs

Through training on a large number of data sets, the model has wide applicability to different expressions of sentences, and also prevents over-fitting problems caused by excessive model parameters or insufficient parameters. Using the hidden layer in the model, we can re-predict the headline of the original text. The test output is shown in Table 1. Although the new headline is not exactly the same as the original one, it remains unchanged in general semantics.

As the dataset of the image, we use the title and description information about the image in the news, and the information can make the model have the ability to convert the image content into a short text description. In the text, in order to facilitate training, we introduce additional noise data and add more training data as a supplement. In order to ensure the integrity of the image semantic extraction, we adopt the fixed dimension for similarity calculation.

In order to reduce needs of data for image training and the time consumption, the existing VGG model and weights are used in the extraction of picture features. The same hidden layer feature as the text feature dimension is generated by adding an additional full join operation on the hidden layer. The results of text generation are shown in Figure 8.

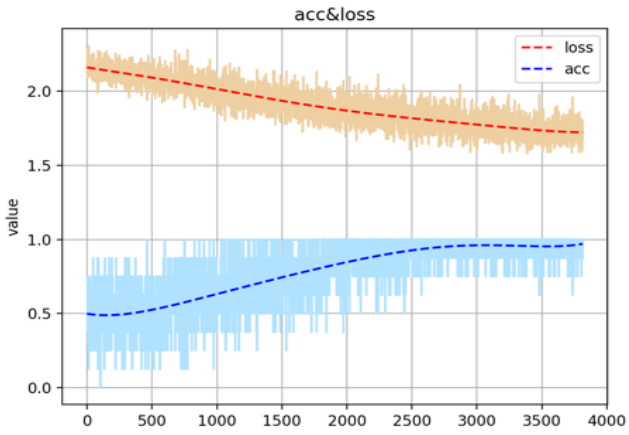

Figure 9 Loss & Accuracy Curve

Using the semantic feature extraction model of trained pictures and texts can ensure that the current output intermediate layer feature extraction results have been able to well characterize the original file. Based on this, the loss calculation method is used to optimize, so that the feature can be directly used to calculate the similarity between the two targets, and then use different conditions for file retrieval. As the number of iterations increases during the training, the accuracy of the output increases continuously, but the overall loss decreases significantly. The variation curve of the loss value and accuracy during the training is shown in Figure 9.

In the test, we use the title data and the picture data in the news respectively to search the relative content, as shown in Figure 10 and Figure 11, using a image or headline, we can obtain similar content respectively. And as the similarity increases, the content of its expression becomes more and more similar. When using a picture for semantic search, the headline results obtained are significantly different from the text. According to the analysis of the headline data, the main reason for this situation is that the semantic features of the text are directly obtained from the

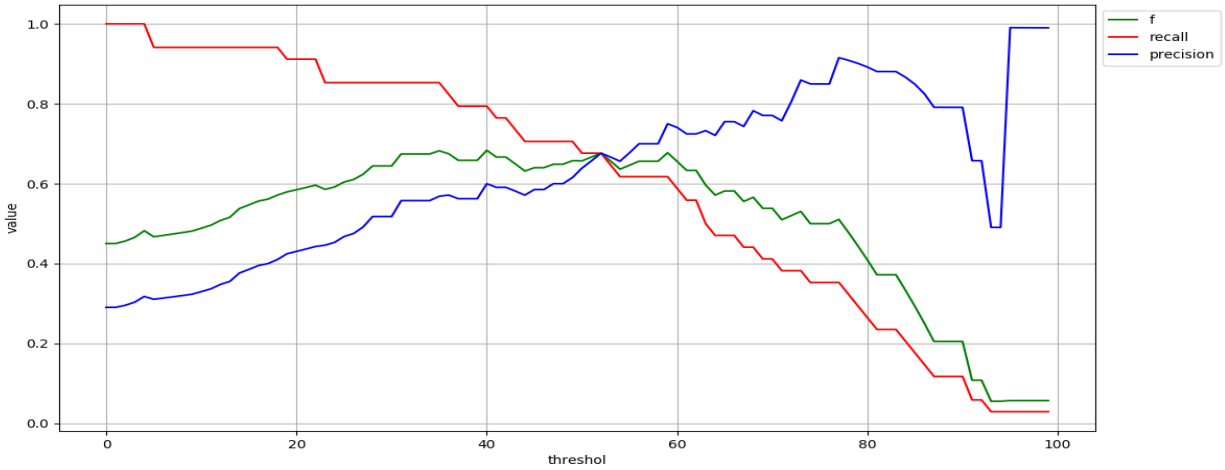

Figure 12 Comprehensive Performance Evaluation

original text, while image semantic features is extracted by image itself rather than its headline text.

4 Performance Metrics

We designed an indicator to evaluate the search results. The accuracy is the proportion of the correct results in all current detection results. On the other hand, we use recall ratio to measure the proportion of correct result in all candidate options. At the same time, the two kinds of measurement results were comprehensively evaluated by F-score, and the test results are shown in Figure 12.

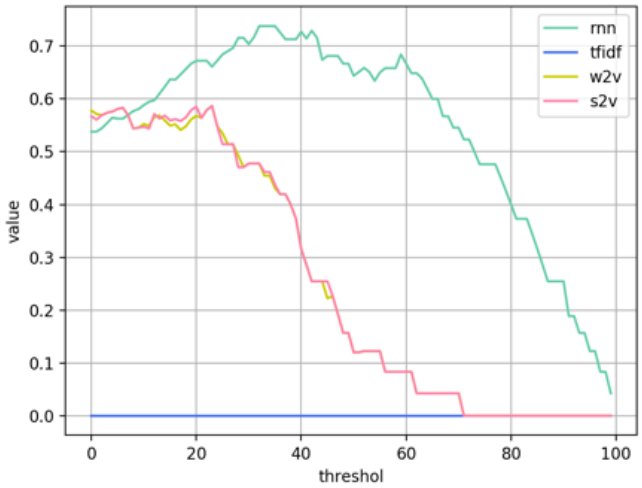

Figure 13 Accuracy compared with other feature extraction algorithms

In addition, in the comprehensive performance evaluation, we use the semantic similarity calculation model based on word vector and sentence vector and TF-IDF model to do comparison. The cosine distance is uniformly used as a measure of similarity .And we use F-score to measure the performance. The final results are shown in Figure 13.

The experimental results show that in the process of information search, we can add semantic features to the final feature representation, it helps get the higher similarity. And the feature extracted by the whole information is much better than local feature extraction. It also proves that the training of semantic similarity model can be completed by using the combination of probability loss and cosine distance.

5 Summary

Similarity calculation is one of the important technical components of information retrieval. Through the measurement of similarity, not only the sort results but also complete screening of information. As for model design, in order to realize the processing of different format information, we need to consider different underlying structure to do feature extraction. Unfortunately, although this method can accurately extract features, it still need to use other datasets to complete the pre-training. Meanwhile, it takes a lot of GPU resources because of the complexity of model structure. In the future work, we hope to design a model which is directly included all the way of the feature extraction, and add the training of the underlying model to the similarity training process.

REFERENCES

[1] Jing L P, Huang H K, Shi H B. Improved feature selection approach TFIDF in text mining[C]// International Conference on Machine Learning and Cybernetics, 2002. Proceedings. IEEE, 2003:944-946 vol.2.

[2] Wong S K M, Ziarko W, Wong P C N. Generalized vector spaces model in information retrieval[C]// International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM, 1985:18-25.

[3] AlexRudnicky. "Can Artificial Neural Networks Learn Language Models?." The Proceedings of the 2000:202-205.

[4] Vaswani, Ashish, et al. "Attention Is All You Need." (2017)[5] Lin K, Yang H F, Hsiao J H, et al. Deep learning of binary hash codes for fast image retrieval[C]// Computer Vision and Pattern Recognition Workshops. IEEE, 2015:27-35.

[6] Russakovsky O, Deng J, Su H, et al. ImageNet Large Scale Visual Recognition Challenge[J]. International Journal of Computer Vision, 2015, 115(3):211-252.

[7] Krizhevsky A, Sutskever I, Hinton G E. ImageNet classification with deep convolutional neural networks[J]. Communications of the ACM, 2017, 60(2):2012.

[8] Vinyals O, Toshev A, Bengio S, et al. Show and tell: A neural image caption generator[C]// Computer Vision and Pattern Recognition. IEEE, 2015:3156-3164.

[9] Xu K, Ba J, Kiros R, et al. Show, Attend and Tell: Neural Image Caption Generation with Visual Attention[J]. Computer Science, 2015:2048-2057.

[10] Ahmed, Ejaz, M. Jones, and T. K. Marks. "An improved deep learning architecture for person re-identification." Computer Vision and Pattern Recognition IEEE, 2015:3908-3916.

[11] Omkar M. Parkhi, Andrea Vedaldi, Andrew Zisserman. "VGG Face Descriptor ."