Research on Application of Deep Learning in Text Generation and Image Captioning

Junhong Ma

Xi'an International University of Shaanxi province China, 710077

maxiaofei913@163.com

Abstract. Deep learning has attracted much attention in various fields because of its excellent learning ability. This paper studies the application of deep learning in text generation and image recognition, and analyzes the flow of text generation and the structure of deep convolutional neural network. Finally, the future development trend based on deep learning research is prospected, and the network design idea of integrating different data information for learning is proposed.

Keywords: Deep Learning; Image Captioning; Image Features; Text Generatio

1. Introduction

Deep learning is a new research direction in the field of machine learning, which is introduced into machine learning to make it closer to the original goal -- artificial intelligence. Deep learning is the internal law and presentation layer of learning sample data, and the information obtained through learning is very helpful for data interpretation, such as text, image and sound. The goal is to make machines infinitely close to human analytical learning. At present, deep learning has made remarkable achievements in text and image recognition.

2. Research Background

In the Internet era, the generation, dissemination and accumulation of large-scale data promote the development of data processing technology. With the popularization of intelligent terminal devices and the explosive growth of multimedia applications, the generation and accumulation of corresponding data are also increasing day by day. How to better use and process these data has become a common concern. Images and text are the most common forms of data in daily life, and also the main components of Internet data. Related researches also tend to focus on images and text data, giving birth to a series of effective data processing methods and intelligent application systems.

In the application of large-scale data, relying on human analysis can no longer meet the demand, which needs to be processed by computer and artificial intelligence algorithm. The research and application of artificial intelligence in the popularity of high-performance computers and GPU ushered in a new upsurge. In the context of big data, data diversity puts forward higher requirements for the generalization ability of algorithms, and traditional methods based on artificial rules are often too complex to be applied. On the other hand, the dramatic increase in computing and storage capacity makes large-scale data computing possible. Supported by data volume, deep learning based on statistics gradually replaces the traditional methods and becomes a popular direction in the current research

field.The research on artificial intelligence has been placing human beings' curiosity and dream of intelligent life in the form of science fiction. People hope to give computers the ability to understand the visual world around them and communicate with human beings in language. However, the information processing and transformation that is natural to human beings is not so easy for computers, so the study of image (vision) and text (language) data is an important proposition in the study of artificial intelligence.

3. The Application of Deep Learning

Nowadays, deep learning is widely concerned and has become a new development direction in machine learning, and has achieved good results in many practical applications. The main reason for the success of deep learning lies in its excellent feature learning ability. In the field of text classification, the classification effect largely depends on the characteristics of the data set, but the traditional statistical methods such as information gain, mutual information and principal component analysis are often used in feature selection. These methods rely on manual feature extraction, which may lead to inaccurate feature extraction when the dimension of data set is too large, and feature extraction depends on luck component. Deep learning enables the original features to be expressed through continuous mapping combinations to form abstract high-level features. In the field of image recognition, Google Brain USES deep learning model to identify cats in a large number of video images. In the field of speech recognition, the deep learning model is used in the simultaneous interpretation system developed by Microsoft.

4. Text Generation and Image Captioning

Deep learning has a deep connection with traditional neural network, which is also a kind of neural network. Both are a multi-layer network structure, including input layer, multiple hidden layer, output layer, this layered structure is similar to the structure of the human brain. There are some differences between deep learning and traditional neural network training mechanism. The traditional neural network USES BP algorithm to train the whole network. During the training process, the initial weight of the whole network is generated randomly. Because of the gradient diffusion, the error is not well adjusted. In the training process of deep learning, after each layer of network training, the advanced features generated upward by the features of this layer will be consistent with another feature generated downward by the advanced features.

5. The Related Tasks of Text Generation

Parts-of-speech (POS): one of many tasks of text generation, which is defined as the process of assigning specific part of Speech tags to each word in a sentence. Part of speech markers can identify whether a word is a noun, verb, adjective, etc. Part of speech tagging can be applied to a variety of problems, such as information retrieval, machine translation, NER, language analysis, etc.

Parsing (also known as syntactic parsing): It is defined as the process of checking whether a sequence of characters written in a natural language conforms to the rules defined in formal grammar. It is the process of breaking a sentence into a sequence of words or phrases and giving it a specific category of components.

Semantic analysis: It is defined as the process of determining the meaning of a character or sequence of words and can be used to perform semantic disambiguation tasks. When analyzing a given sentence, if the syntactic structure of the sentence has been constructed, then the semantic analysis of the sentence is completed.

Emotion analysis: It is defined as the process of determining the emotional information behind a character sequence. Affective analysis can be used to determine whether the speaker or person expressing the textual idea is in a happy or sad mood, or represents only a neutral expression. Chinese emotion classification is mainly based on the concept of convolution control block. The method is to regard the sentence as an individual unit, based on the model of convolution control block, and compare the dependence of various periods of context for affective classification. The segmentation of a single

sentence is put into five layers of convolution control block for experiment, and the accuracy rate is 92.58%.

6. The Flow of Text Generation

Generally, text generation goes through such processes as data acquisition, preprocessing, attention mechanism and text evaluation. The process is as follows:

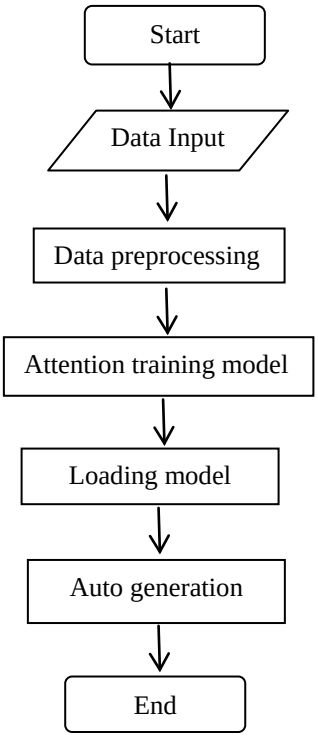

Figure 1. Text generation flowchart

7. The Tasks of Image Captioning

Image description is the combination of computer vision and natural language processing. The goal is to make the computer recognize the image content and automatically generate natural language text to describe the image content, which can be regarded as the translation process from image to text. The task of image description is to make full use of the image information to obtain a more accurate and complete description of the image content. The process includes not only the recognition and extraction of image content and feature representation, but also the generation of description text based on image features.

Most of the existing image description networks are based on codec networks. Deep convolutional networks are used to extract image features. Combined with external information, RNN(Recurrent Neural Network) is used as the decoder network to generate description text.There are two forms of image features extracted from the convolution network: one is the global feature extracted from the whole image, which is mostly directly used to initialize the text generation network; The other is to extract the local features of the features from the image, or to make weighted enhancement to the local areas in the image features, or to combine with the target detection method to extract the targeted local features, or to directly use deep network learning to obtain the target areas, and then extract the local features. The extracted local features are often used in combination with attention mechanism to guide the generation of description text.

8. The Process of Image Captioning

The utilization of image information can be considered from two aspects: one is to introduce richer image information at the image coder level; the other is to decode image features as completely as possible in the description text generation.

In terms of introducing image features, there are two research ideas: one is to introduce local features of images and combine them with attention mechanism to select features in the text generation process; the other is to optimize in the image encoder to obtain image features with stronger expression ability and better matching with the description generation network. Since the global features are characterized by simple network structure and low resource consumption, and the processing and application mode of global features can be easily extended to local features, image global features are generally adopted.

In the deep convolutional network used for image recognition, the convolutional layer usually exists in the form of grouping, which can be regarded as the convolution operation of different stages to extract features of feature graphs of different sizes, and the image features of the next stage can be obtained. According to the number of network layers in the input image of the depth or distance of the convolutional layer, different convolutional layers can be grouped and named as conv-n according to the serial number n, and the output features of convolutional layers of different depths can be named as f-n. Thus, the schematic diagram of the deep convolutional network structure is as follows:

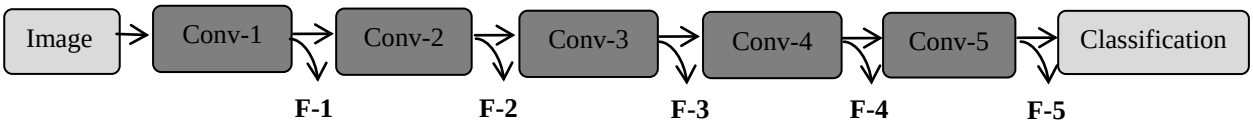

Figure 2. Deep convolutional neural networks structural representation

9. Conclusion

In the process of computer processing, both image and text data are presented in the form of discrete Numbers with certain structure rules. The difference is that the image data reflects the continuous visual data in reality, while the text data is naturally discrete. According to its different characteristics, the processing ideas and algorithm structures for image and text data are often different. References. Network design based on fusion learning of different data information has gradually become a hot issue in the field of deep learning. The integration and utilization of different input data information will also greatly enhance the computer's ability to deal with data. How to extract semantic information from text data and how to recognize image content and feature expression will continue to be important topics in natural language processing and computer vision. If image information and text information are integrated and utilized, description text is generated by image information, corresponding images are generated by text data, and finally the combination of image information and text information is realized, it will be an important field for the development of deep learning.

References

[1] K. Wang, X. Wan. Automatic generation of sentimental texts via mixture adversarial networks. Artificial Intelligence, (2019).

[2] Anderson, P., He, X. D., Buehler, C. Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering, (2018).

[3] Zhou, C. Research on Text Classification Based on Hybrid Model of Deep Learning. Lanzhou University, 2016.

[4] Huang, J. J., Li, P. W., Peng, M., Xie, Q. Q., Xu, C. Review of Deep Learning-based Topic Model. Chinese Journal of Computers, 2019.