Existing Work

Knowledge representation for the generation of quantified natural language descriptions of vehicle traffic in image sequences

The issue of generating natural language descriptions from visual data has long been studied in computer vision. This has led to development of complex systems composed of visual primitive recognizers combined with a structured formal language, e.g. And-Or Graphs or logic systems, which are further converted to natural language via rule-based systems. Such systems are heavily hand-designed, relatively brittle and have been demonstrated only on limited domains, e.g. traffic scenes or sports. Leveraging recent advances in recognition of objects, their attributes and locations, allows us to drive natural language generation systems, although these are limited in their expressivity.

Show and Tell: A Neural Image Caption Generator

A group of researchers at Google presented a generative model based on a deep recurrent architecture that combines recent advances in computer vision and machine translation and that can be used to generate natural sentences describing an image. Their model is trained to maximize the likelihood of the target description sentence given the training image. Experiments on several datasets show the accuracy of the model and the fluency of the language it learns solely from image descriptions. The model is one of the most accurate system developed for image captioning, which was verified both qualitatively and quantitatively.

A Comprehensive Survey of Deep Learning for Image Captioning

This paper reviews deep learning-based image captioning methods. It discusses different evaluation metrics and datasets with their strengths and weaknesses. A brief summary of experimental results is also given. It briefly outlines potential research directions in this area. Although deep learning-based image captioning methods have achieved a remarkable progress in recent years, a robust image captioning method that is able to generate high quality captions for nearly all images is yet to be achieved. With the advent of novel deep learning network architectures, automatic image captioning will remain an active research area for some time.

Research on Text Classification Based on CNN and LSTM

With the rapid development of deep learning technology, CNN and LSTM have become two of the most popular neural networks. This paper combines CNN and LSTM or its variant and makes a slight change. Unlike the typical CNN, which contains convolution operation and activation function, this paper constructs two text classification models called NA-CNN-LSTM and NA-CNN-COIF-LSTM by combining CNN without activation function and LSTM, and one of its variants COIF-LSTM. Through comparative experiments, it is proved that the combination of CNN without activation function and LSTM or its variant has better performance. The experimental results on the subjective and objective text categorization dataset show that the proposed model has better performance than the standard CNN or LSTM

Composing Simple Image Descriptions using Web-scale N-grams

Some existing systems use detections to infer a triplet of scene elements which is converted to text using templates. Similarly, Li et al. piece together a final description using phrases containing detected objects and relationships by starting off with detections. A more complex graph of detections beyond triplets is done but with template-based text generation. The mentioned approaches have been able to describe images “in the wild”, but they are heavily hand designed and not flexible when it comes to text generation.