SSDNN-See and Say in Deep Neural Networks

In Deep learning area, Natural Language Processing and Computer Vision are popular to make amazing artificial intelligence applications. We combined both of them in our project called "See and Say in Deep Neural Networks" - (SSDNN) to automatically describe images using machine-generated sentences.

Our model made use of datasets of images and their sentence descriptions to learn about the relationship between language and visual data through the novel combination of Convolutional Neural Networks over image elements, as well as bidirectional Recurrent Neural Networks over sentences. Then we align the two modalities through a structured objective and multimodal embedding. We then utilize a Multimodal Recurrent Neural Network architecture that uses the inferred alignments to learn to generate novel descriptions of image regions.

We demonstrate that our alignment model produces state of the art results in retrieval experiments on popular datasets. We then show that the generated descriptions have good results on original images as well as new images.