Automated image captioning

The research of comprehension of machine learning algorithms and procedures suitable for both image processing and natural language processing.

Technology has become an integrated part of our daily lives over the past decades. From offering readily-answers with the use of search engines, ease of world-wide cheap communication through the messaging applications that run on the internet to solving complex engineering problems.

Image caption generation is the problem of generating a descriptive sentence of an image. Also, automatically describing the content of an image is a fundamental problem in artificial intelligence that connects computer vision and natural language processing. A brief survey of some technical aspects and methods for the description-generation of images is presented in [7]. Paper [8] presents a new challenge of image captioning which is how to effectively inject sentiments into the generated captions without altering the semantic matching between the visual content and the generated descriptions.

One of the objectives is to research the vast field of machine learning to identify the best techniques to make use of in this specific context. Since there are a lot of algorithms and architectures available, it is easy to pick one that is suitable, but not the best technique possible, do a thorough research is implied whenever dealing with a problem that wants to be solved using deep learning or machine learning techniques. Luckily, there are a lot of papers and blog posts available, most of them referring to the theoretical aspects of the solutions proposed

The paper aimed to develop an application for automated description of images using deep learning, while also doing thorough research of the field. This prototype has to be further integrated with a text-to-speech engine to reach its goal of helping visually impaired people make some sense of pictures.

After researching the possible models and architecture, a composite model was selected, consisting of a convolutional neural network, as the encoder. That makes use of transfer learning, so reducing the training time and the computational complexity, and a recurrent neural network, as the decoder, which is the state of the art when it comes to dealing with textual data, especially predicting the next word in a sentence (sequential prediction). Since this classical architecture is not able to keep in memory long sentences, the attention mechanism was introduced as an interface between the two models, its output representing the features extracted from the encoder, concatenated with the attention weights.