Abstract:

We present a system for solving the holy grail of computer vision -- matching images and text and describing an image by an automatically generated text. Our system is based on combining deep learning tools for images and text, namely Convolutional Neural Networks, word2vec, and Recurrent Neural Networks, with a classical computer vision tool, the Fisher Vector. The Fisher Vector is modified to support hybrid distributions that are a much better fit for the text data. Our method proves to be extremely potent and we outperform by a significant margin all concurrent methods.