Md. Musfiqur Rahman

Id: 1620277042

Paper title: Automatic Image Captioning

Author name: Jia-Yu Pant (Computer Science Department, Carnegie Mellon University, Pittsburgh, U.S.A.)

Abstract: In this paper, they examine the problem of automatic image captioning. Given a training set of captioned images, we want to discover correlations between image features and keywords, so that they can automatically find good keywords for a new image. They experiment thoroughly with multiple design alternatives a large dataset of various content, and our proposed methods achieve up to a 4. 7% relative improvement on captioning accuracy over the state.

Critical analysis: They proposed 4 methods to estimate a W – by – B. (Unweighted data matrix) Given an annotated image set 1=(11, ..., IN] with a set of terms W and a set of blob-tokens B, the unweigbted data matrix D~l’DwolD~~] is a N-by-W+Bj matrix.

They weighted the counts in the data matrix D according to the “uniqueness” of each tend blob-token. If a term appears only once in the image set, say with image Ii. In definition 3 they mention T,= measures the association between a term and a blob-token by the co-occurrence counts. Another possible measure could be to see how similar the overall occurrence patten (over the training images) of a term and a blob-token is. Such occurrence pattens are in fact the columns of DW or De, and the similarity can be taken as the cosine value between pairs of column vectors.

Dataset: The experiments are performed on 10 Core1 image data sets. Each data set contains about 5200 training images and 1750 testing images. The sets cover a variety of themes ranging from urban scenes to natural scenes, and from artificial objects like jet/plane to animals. Each image has in average 3 captioning terms and 9 blobs.

Conclusion: they do thorough experiments on 10 large datasets of different image content styles, and examine all possible combinations of the proposed techniques for improving captioning accuracy. The proposed methods are less biased to the training set and more general in terms of retrieval precision and recall.