Towards Effective Image Annotation by Exploiting Multimodal Data

徐, 行

https://doi.org/10.15017/1543997

出版情報：九州大学, 2015, 博士（学術）, 課程博士
バージョン：
権利関係：全文ファイル公表済

（別紙様式 2）

氏　　名　：徐　行

論 文 名　：Towards Effective Image Annotation by Exploiting Multimodal Data
（マルチモーダルデータ分析に基づく効果的な画像アノテーションに関する研究）

区　　分　：甲

論　文　内　容　の　要　旨

Automatic image annotation is an active and challenging problem in computer vision research. Given an unseen image with no caption, the goal of image annotation is to predict relevant textual words from a word vocabulary to describe visual content of that image.

In this thesis, firstly the author focuses on the traditional image annotation problem, and particular interests are focused on developing effective scheme to capture the topic correlations in the latent Dirichlet allocation (LDA) based topic models to improve the annotation performance. To directly model correlation between topics when generating the topic proportions for a given image, the author uses logistic normal distribution proposed in the correlated topic model (CTM) instead of the Dirichlet assumption to capture the topic correlations. Then, the author extends the idea of modeling topic correlations from popular LDA based models to CTM based models. In the proposed CTM based models, topics are now correlated with each other to mix the topic proportions for one image. The author presents a comprehensive comparison between CTM based and LDA based models on three benchmark datasets, illustrating the superior annotation performance of proposed CTM based models on image annotation task, by means of propagating topic correlation among image features and annotation words. Extensive experiments on three benchmark datasets demonstrate the superior annotation performance of proposed CTM based models on image annotation task

Secondly, the author analyzes the statistical properties of user-provided ground truth labels (also called “tags”) of training images in the image annotation problem, and investigates how the quality of the ground truth labels affects the annotation results. The author first addresses a special case of image annotation with incomplete labeling, where the multiple objects in each training image are not fully labeled. The author leverages the structured output learning method to boost the performance of conventional one-versus-all support vector machine (OVA-SVM) classifiers, and proposes an image specific structured loss function which is more appropriate to explore the dependency of predicted multiple labels. Experimental evaluation demonstrates that the proposed method is efficient to handle the issue of incomplete labeling, and it performs superior than several existing methods on image annotation task.

After discussing the special case above, the author considers a more general case of tag completion with defective tag assignments, where ground truth tags cover both incomplete and noisy situations. The author proposes an image-tag re-weighting scheme and formulates a

re-weighted empirical loss term, which is more efficient for optimization in tag completion problem under defective tag assignments. Experimental evaluations verify that using proposed re-weighting scheme achieves superior tag completion results on various degree of incomplete and noisy settings.

Lastly, the author extends the image annotation problem to a more general issue of cross-modal retrieval considering mutual associations of images and texts. The cross-modal retrieval tasks such as text-to-image search and image-to-text search are cast to modeling images and associated text through latent subspace learning. To handle the diversities of different modalities of data, the author develops a novel framework that consists of two procedures: (1) a coupled dictionary learning method is first developed to generate homogeneous sparse representations for different modalities by associating and jointly updating their dictionaries; (2) a coupled feature mapping scheme is then used to project the derived sparse representations from different modalities into a common subspace in which cross-modal retrieval can be performed. Experiments on a variety of cross-modal retrieval tasks demonstrate that the proposed framework outperforms the state-of-the-art approaches.