TRIYONO – COMP594
FALL 2018

ANNOTATED BIBLIOGRAPHY

Related works

1. A. Karpathy, L. Fei-Fei. “Deep visual-semantic alignments for generating image descriptions.” IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 20157298932
The goal of this paper is learning to generate Natural Language descriptions by studying intermodal correspondence between an image and its description. The author proposed Multimodal Recurrent Neural Network architecture which is Bidirectional RNN for representing sentences and Region Convolutional Neural Network for image representation. Then it uses inferred latent alignment to generate a description of a particular area of an image. Sentences are treated as a weak label, where a continuous word written by people will be corresponded to a particular region of the related image. Here, a single word will match to specific best region in the corresponding image. The authors utilize the Markov Random Field and treat the true alignment as latent variables for addressing goal in generating contiguous words instead of only single words. But, there are limitation about generating description only from one input and how RNN receives the image information.
Food image-recipe task may be handled as a similar task to the problem in the paper. By treating each ingredient as a sentence that correlates to a particular region in the food image. Model that tackles image representation in my project will employ a structure (fine-tuned pre-trained image recognition deep learning model) with a goal for extracting content ingredients that occupy on the image. The difference with applied in this paper is, instead of using Bi-RNN, my project will utilize LSTM that proven outperforms RNN in performance.
2. W. Min, S. Jiang, J. Sang, H. Wang, X. Liu, L. Herranz. “Being a Supercook: Joint Food Attributes and Multimodal Content Modeling for Recipe Retrieval and Exploration.” IEEE Transactions on Multimedia, 2016.
This paper presents recipe-oriented-image-ingredient correlation learning problems with the use of additional attributes compared to previous studies. With Yummly 28k consisting of 27,638 pairs of items of image, ingredients, course, and cuisine, the author divides ingredients into two types, i.e. visible and non-visible. The proposed method is based on Multimodal Restricted Boltzmann Machine (RBM) to handle the weak dependence between visual content and textual ingredient that is M3TDBN (MultiModal MultiTask Deep Belief Network). M3TDBN aims to combine multitask learning so that several attributes can be collaborated with one another. This paper applies the method to three areas: multimodal cuisine classification, attribute-augmented cross-modal recipe image retrieval, ingredient and attribute inference from food images. This paper presents promising performance when using more attributes as additional to ingredients in the recipe-oriented-image-ingredient correlation task. The idea is with additional attributes, the performance on image recognition task can be leveraged to another level. Our project will use a dataset that contains only three main attributes which are title, ingredients, and instructions. The baseline that was done by the following paper need to be tweaked. With several additional attributes, we hope the performance of my project’s model will be above the baseline. Instead of using a model based on RBM, we will apply deep learning model in the project.
3. A. Salvador, N. Hynes, Y. Aytar, J. Marin, F. Ofli, I. Weber, A. Torralba. “Learning Cross-Modal Embeddings for Cooking Recipes and Food Images.” IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017.
This paper introduces and releases Recipe1M which is a large scale image-recipes dataset to tackle data limitations that might affect performance in existing work. The goal is to find joint embedding

between image and recipes for image-recipe retrieval. For achieving the goal, it proposes a multi-modal neural network with jointly learn visual and textual context in embedding space extended with regularization via high-level classification task. The model consists of two tasks, one task manages image representation that utilizes state-of-art deep convolutional networks, such as VGG-16 and ResNet-50, and the other part is for recipes representation that include ingredients-encoder and instructions-encoder that jointly using a fully connected network. The outputs from both tasks then are embedded into recipe-image joint embedding space. The work employed semantic regularization by adding a classification task and the cosine loss in joint embedding space. Aligning two modalities may come from shared discriminative weights from them. The semantic regularization was proved in quality of embedding in im2recipe task. By the model structure, the recipe and image embedding are semantically aligned.

This paper manages the image-recipe retrieval problem using multi-task multimodal that tackles each modal with different tasks. This idea is actually a general idea for several existing cross-modal task for image-text retrieval. Likewise, this project will utilize the same idea as this paper, that is using multi-task model and learning similarities from the outputs of the tasks. Moreover, Recipe1M that was introduced in this paper will be used as the dataset for my project. Furthermore, we will use the results in the paper as comparison baseline. On the other hand, we will employ and experiment semantic hashing in my project for semantic-embedding space learning.

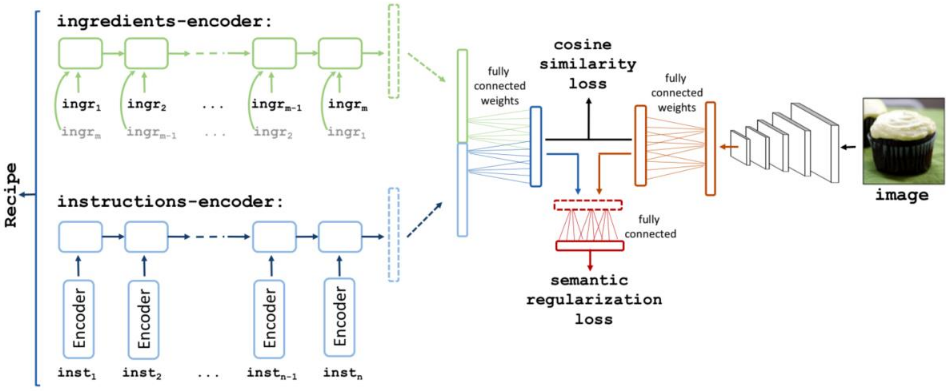

Figure 1 the joint neural embedding model that is proposed in [3]

4. J. Chen, C. Ngo. “Deep-based Ingredient Recognition for Cooking Recipe Retrieval.” ACM, 2016.
   Like the previous paper, this paper uses and releases image-ingredients dataset which is VIREO-172 that consists of Chinese food with split into 172 categories. The main goal is trying to analyze food with an unknown category called zero-shot retrieval by combining content based which is an ingredient and category to recognizing large variations in food appearance. The retrieval task is conducted in one direction which is image to ingredients retrieval. Given an image, the model provides a list of set ingredients based on the image query. Moreover, since a Chinese recipe that in nature, have diverse appearance and wild composition of ingredients that may be a problem with image recognition task, the paper proposed network by modifying DCNN, and build 4 types stacked architecture, placing different strategies for food categorization task and ingredients recognition. This model based on the idea about directly and indirectly influence between tasks. Directly means input of one task is connected to the output of another task. Further, finding co-occurrence correlation between ingredients based on conditional random field (CRF) and finding improvement for Arch-D model, then directly inferring ingredients using VGG16 are done and improved by Implementing graph

modeling. This paper concludes that image retrieval is affected by image query resolution, same ingredients in different categories, visibility of ingredients in image query, dish appearance diversity. One handy idea of this paper is my project should overcome limitation about dish's diverse appearance and some latent ingredients in food image-recipe recognition task. By applying and treating image representation task as an image-object recognition, hopefully it can improve the performance in recognizing ingredients on dish image, that implies learning similarities in the embedding space become easier. The other difference against this paper is, in my project we plan to conduct two directions of image-recipe retrieval instead of only image to recipe retrieval.

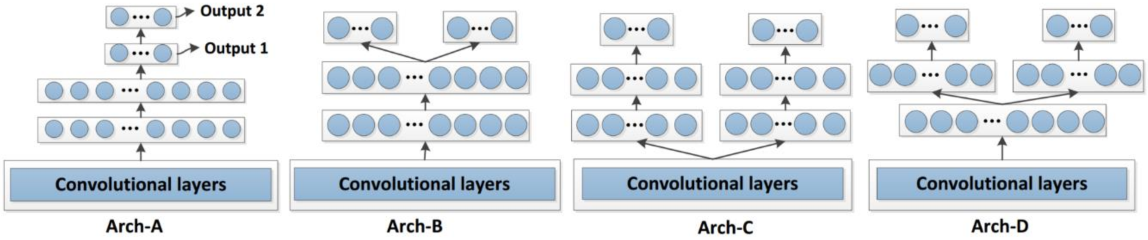

Figure 2 Four different deep architectures for multi-task learning that were proposed in [4]

5. L. Bossard, M. Guillaumin, Y. Van Gool. "Food-101-mining discriminative components with random forest.", European Conference on Computer Vision, pages 446-461. Springer, 2014
Bossard et.al uses random forest algorithm to mine discriminative parts of food image for classifying them. They also release new dataset 101-food, contains images and corresponding category. This paper focuses on supervised learning in classification task and only consider discriminative object that called as components that are aligned with super-pixels. The steps for mining discriminative components are as follows, generating candidate cluster by training weakly supervised random forest on super-pixels related image's class label that resulting candidates for discriminative components, mining components that increase diversity of the components and maintaining the strongest one, training the component models, and recognition from mined components by score all images super-pixels. By using this method, the paper shows average accuracy of 50.76% that only lower than classification using CNN.
Food image recognition requires handling image pixels precisely to get the expected results. The idea can come from this paper in handling the pixel array of dish image to only use the right part so that it can reduce the computational cost, but on the other hand it still maintains accuracy. Moreover, fundamentally, since learning similarities in a semantic embedding space can be considered as clustering or classification task, even we plan to use hashing method, increasing diversity between food category while maintaining similarities will help leveraging retrieval accuracy. Instead of doing classification task, our project will try to solve image-recipe recognition task in a different manner.
6. J. Chen, C. Ngo, T. Chua. "Cross-modal Recipe Retrieval with Rich Food Attributes.", ACM, 2017.
This paper uses not only ingredients as a textual component, but also additional attributes which are cutting and cooking method along as image for food recognition and recipe retrieval. Food appearances may be different even has similar ingredients, since having different cutting and cooking methods. Therefore, localization of ingredients and its cutting and cooking method is employed in region of an image, by using DCNN and a mechanism that predicts them in embedded features. Learning classification of the image is conducted in region-wise manner which is utilizing grids that correspond to a small region of the image. The model was able to recognize the ingredients in different dishes with some limitation for mixed ingredients. It employs an average of all of loss

functions for each the image attribute which using cross-entropy as its loss function. This paper found that joining cutting and cooking method improves image-recipe retrieval significantly, with the cutting method outperforms cooking method attributes. The multi-task proposed model was also compared with single-task model that show the second model underperform due to lack of training samples.

This paper together with paper #2 show that adding appropriate attribute for food image-recipe recognition is useful. Moreover, this paper given an example how to implement additional components in the deep learning model structure that suit with my project. Further, we should still aware with some ingredients that may be non-visible on a dish presentation, because of its nature, hidden, or mixed with other ingredients that change its shape and color. On the other hand, my project is different with this paper, since this paper only does retrieval in one direction, i.e. image to recipe retrieval, instead of two directions.

7. C. Liu, Y. Cao, Y. Luo, G. Chen, V. Vokkarane, Y. Ma. "DeepFood: Deep Learning-Based Food Image Recognition for Computer-Aided Dietary Assessment.", ACM, 2016.

   The goal is for improving dietary assessment accuracy by analyzing captured image using mobile devices. The authors proposed method that is based on deep learning for food image recognition using Convolutional Neural Network. The proposed model is based on pre-trained model such as LeNet-5, AlexNet, and GoogleNet with additional called inception module to increase neural network representation power. Besides improvement accuracy using the model compared with preceding paper, the author also proved that employed bounding box in the image and used the information in training, boost the classification accuracy.

   It is proven that by using bounding box in an image, can improve the classification performance. This is one of the ideas in object detection problem that similar to ingredients recognition from corresponding image in my project. The idea is similar with paper #1, but has been applied in cases of food image-recognition that match with my project. Instead of focusing on mobile device application our project will have no boundary in the implementation and not only limited to classification task.