|  |  |
| --- | --- |
| **Project title** | **Learning Semantic Feature Map for Visual Content Recognition** |
| **Team number** | **3** |
| **Group member** | **梅建林SA19225286** |
| **张洲SA19225493** |
| **contact** | **1291247841@qq.com**  **1944996484@qq.com** |

**Abstract**

In this project, we propose to learn Semantic Feature Map (SFM) by deep neural networks to model the spatial object contexts for better understanding of image and video contents. First, we will study the fundamental of this proposal, including what, why and how about the approach through this paper, reference and material online. Then, We will learn how to achieve this argument, including extracting high-level semantic object features, organizing them to the designed SFM and employing either Fully Convolutional Networks (FCN) or Long-Short Term Memory (LSTM) on top of SFM for final recognition. To complete this subject, we need to develop a proposal, where the plan, the understanding of the paper,our arrangement of the work, a schedule and the expectation will be recorded, to guarantee sequential tasks every period methodically.

**Introduction**

Due to larger data, progressive computing power and more efficient algorithm models, such as CNN[1], visual recognition obtain great development and application in many significant fields, and will remain to be progressive in the near future. In this process. Convolutional Neural Networks (CNN) models plays the most significant role. A standard CNN model consists of several convolutional layers and a few fully connected layers for classification, operating on normalized images with single object. Its ability to decrease parameters and progressively learn discriminative features from different level mainly result in its success in visual recognition. However, abundant underlying contextual information among objects is not utilized when we recognize classes of images or videos with other objects via CNN, which can be considered to improve the effect of visual recognition..

Nowadays, deep learning frameworks usually couple CNN models with semantic region proposals[2], which generates separate features that can be aggregate. Although pooling operation and a certain encoder can complish the integration, they still discard the shape, position information. To address the above limitations, in this project, we propose to learn a generic semantic representation in a multi-task framework for general visual recognition tasks. So that is what this proposal aims to do and our project need to process.

**Literature Review**

Image and video classification problems have been extensively studied with much efforts paid to the design of high quality image feature representations. We focus the review on recent works related to this approach.

Deep CNN networks are the most popular and successful models for various visual recognition tasks in recent years. Recently, combining advanced CNN models with semantic regions is widely used in visual recognition systems since image or video classes are strongly related to the objects within it. Many researchers and scholars put forward many complex and effective models that originate from CNN and this combination, such as a two stream CNN model[3] for action recognition, R-CNN[4], Faster R-CNN[5], YOLO[6], SSDs[7] for more precise and fast object detection. These methods are mainly designed for image object detection while we aim to utilize the rich object clues for holistic image recognition. Generally, various extracted object features are required to be combined together into a unified image-level representation so that it could be fed into the classifiers, including Hypotheses-CNN-Pooling (HCP)[8], RNN coupled with CNN[9], multi-scale SPP[10] and so on.

**Project schedule**

To complete this subject, we develop a schedule to guarantee sequential tasks every period methodically.

9.9-10.7: **Understanding**. just like an opening report, we would finish the technology and background research related to our proposal, which can make us clearly understand what our proposal solve or improve. And a report will be generated for later procedure.

10.7-10.31: **Reference**. Other papers or literature mentioned in this proposal need to be checked and comprehended so that we can comprehend why this proposal is a state-of-the-art method and how it process.

11.1-11.31: **Implement**. After we know what, why and how about the proposal, we can implement the method actually by python language and tensorflow or keras frame. Then, to test whether the proposal is as good as the theory, we utilize data sets form Pascal VOC, MSCOCO and CCV, including images and videos to evaluate. In addition, we also compare this proposal with other method in the same metrics.

12.1-end: **Optimization and Document**. On the basis of this proposal, we can make somewhere better or subjoin other artifice to further improve the recognition speed or accuracy, such as pre-trained nn, multi-level features and so on. Before the project is completed, a summary report, as the ending, is supposed to be writed and handed on.

**Project Assignment**

|  |  |
| --- | --- |
| Zhou zhang | 1. grasp the knowledge and usage about deep learning and Computer vision such as CNN model, Semantic segmentation, Full convolutional network, LSTM, openCV and so on. 2. Practice one killed frame of deep machine learning at least, such as tensorflow and keras. 3. extract separate features map through a CNN model from an input image. And unite them to a SFM 4. Utilize FCN and LSTM to classify the images 5. Record the data and edit a report. |
| Jianlin mei | 1. grasp the knowledge and usage about deep learning and Computer vision such as CNN model, Semantic segmentation, Full convolutional network, LSTM, openCV and so on. 2. Practice one killed frame of deep machine learning at least, such as tensorflow and keras. 3. Find and download image and video dataset to train 4. Transform the model to classify videos 5. Record the data and edit a report. |

**Methods**

Accordingly to the paper which we have chose, we could learn a lot of visual recognition algorithms. Now we have initially established what methods to be used.

1. When an image is input, first of all need a region proposal algorithm to process the image so that we can obtain several thousand bounding box, which corresponds to all the patches in the image that are most likely to be objects. Then we process every ROI to produce fix sized feature maps through a CNN model.

2. Next is the significant and difficult. we reorganize these object-level semantic features into a contextual representation named Semantic Feature Map (SFM), which preserves the spatial layout and interaction information among multiple objects.

3. And this step is very important, too. We utilize both FCN and LSTM modules to model the SFM representation for recognizing targeted classes. For the LSTM module, we consider the SFM grids to be a sequence and each grid as a time step, upon which the LSTM is applied to explicitly model spatial context information.

4. Multi-task learning will use in this project, to adjust the size of label by multi-task loss function, which can obtain better ability of object recognition.

5. Finally, we get the data such as MPa, ROC, and display them in the form of an image.

**Dataset**

The theory can’t prove the actual advantages and progressive effects, so we must conduct full experiments based on our proposal.

For image classification, we can use dataset form network such as PASCAL VOC or MS coco to evaluate our method.A lot of image in the dataset we can use.

For video classification, we can use CCV dataset to evaluate our method. The dataset contains 9,317 YouTube videos and 20 classes.

After all data had been processed by the approach. We calculate the mean average precision to evaluate it. Also we can make a comparison with state-of-the-art methods methods on PASCAL VOC, and then we can intuitively see whether our method is more efficient or not.

**Expectation**

In this project, the most purpose is learn a semantic feature map for visual content recognition, which can model spatial contextual information among objects. When we do this job,large of visual processing knowledge is necessary. Multi-task loss function also used combined with FCN and LSTM in this project, which can adjust the location and size so that make a better classification results. At last, transfer the semantic feature map model of image data to video classification task is useful and necessary.

When we begin to achieve this project, we can learn more and more state-of-the-art to model to have a try to improve the semantic feature map so that can get a better results of classification task.

**Reference**

1. Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. ImageNet Classification with Deep Convolutional Neural Networks. In Advances in Neural Information Processing Systems. 1097-1105.
2. Yunchao Gong, Liwei Wang, Ruiqi Guo, and Svetlana Lazebnik. 2014. Multiscale Orderless Pooling of Deep Convolutional Activation Features. In European Conference on Computer Vision. 392–407.
3. Karen Simonyan and Andrew Zisserman. 2014. Two-Stream Convolutional Networks for Action Recognition in Videos. In Advances in Neural Information Processing Systems. Montréal, Canada, 568–576.
4. Ross B Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. 2014. Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation. In IEEE Conference on Computer Vision and Pattern Recognition. 580–587.
5. Ross B Girshick. 2015. Fast R-CNN. In IEEE International Conference on Computer Vision.
6. Joseph Redmon, Santosh Divvala, Ross B Girshick, and Ali Farhadi. 2016. You Only Look Once: Unified, Real-Time Object Detection. In IEEE Conference on Computer Vision and Pattern Recognition. IEEE, Las Vegas, NV, USA, 779–788.
7. Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott E Reed, Cheng-Yang Fu, and Alexander C Berg. 2016. SSD: Single Shot MultiBox Detector. In European Conference on Computer Vision. Springer International Publishing, Amsterdam, The Netherlands, 21–37.
8. Yunchao Wei, Wei Xia, Min Lin, Junshi Huang, Bingbing Ni, Jian Dong, Yao Zhao and Shuicheng Yan. 2015. HCP: A Flexible CNN Framework for Multi-label Image Classification. IEEE Transactions on Pattern Analysis and Machine Intelligence (2015), 1–8.
9. Jiang Wang, Yi Yang, Junhua Mao, Zhiheng Huang, Chang Huang, and Wei Xu. 2016. CNN-RNN: A Unified Framework for Multi-label Image Classification. In IEEE Conference on Computer Vision and Pattern Recognition. IEEE, Las Vegas, NV, USA, 2285–2294.
10. Li-Jia Li, Hao Su, Yongwhan Lim, and Fei-Fei Li. 2014. Object Bank: An ObjectLevel Image Representation for High-Level Visual Recognition. International Journal of Computer Vision 107, 1 (2014), 20–39.