Name: Yuzhe He, NetID: 18yh46, Student#: 20143446

Hybrid Deep VGG-NET Convolutional Classifier for Video Smoke Detection

Fire brings danger and loss to people's lives and properties, but real-time wild smoke detections do not have proper accuracy. The author tries to use video smoke detection to identify the existence of smoke particles.

The data used in this experiment is using AdaBoost with staircase searching method. They first obtain arithmetical feature and Haar like features from the RGB images, and then applied dual threshold AdaBoost algorithm to image to categorize the smoke. The feature extraction of the smoke image includes having a pixel oriented high order directional derivative encrypting, high-order local ternary patterns. After noise eliminated, the SVM classifier is utilized.

The method used in this paper is a deep VGG-net CNN classifier. It first requires the frame pre-processing which spilt the video into 1162 frames and use 3 grey scale images to represent RGB. Then uses 2D-DCT to compress the image. After the data is processed, they used alpha channel to update background and threshold which are used for pixel calculation. The remaining process is to use Ostu method to identify the fire. Meanwhile, motion analysis is required since fire causes movement of air. The VGG-net can use minimum frame pixel as the training set and the convolutional layer has 16 output channels and 100 to 50 to 5 neurons fully connected. The training used f'(x) = 1/(1+e^(-x)) as activation function and training with back propagation. It used artificial bee colony as weight optimization.

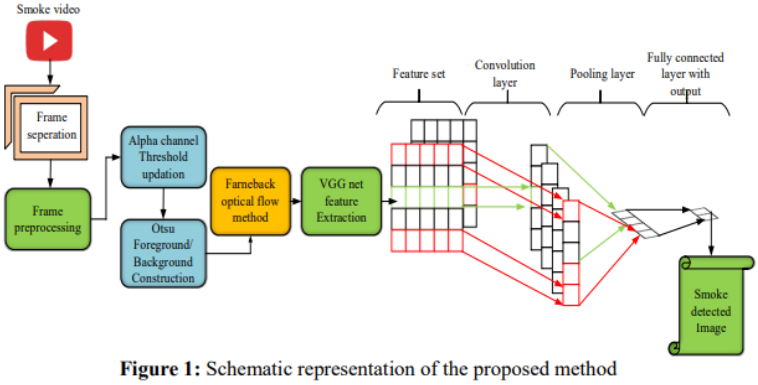


Figure 1: Schematic representation of the proposed method

The result is shown in picture, the suggested precision if 96% and VGG-Net CNN has attained minimum of 0.73 but when it is in use it nearly equals to 1.

Overall, this VGG-net CNN is very efficient as it requires less training set and time. But it is a bit hard to classify people's pose as this is used to classify grey scale of RGB (Princy Matlani, 2019).

Table 3: Performance comparison of proposed with existing algorithms

	Proposed VGG-Net CNN	GLNRGB	ALEXRGB
Specificity	0.8333	1.1973	1.1054
FPR	1.5000	0.5443	0.6848
FNR	0.9918	0.9831	1.0068
Precision	1.2833	0.3186	0.2813
Recall	1.6971	0.9626	0.6272
F1score	1.4854	0.2935	0.4671
Accuracy	96.3333	87.2222	92.2222
MCC	0.6402	0.4812	0.4617
TPR	0.9167	0.6136	0.7152
NPV	0.8333	1.1973	1.3900
FDR	0.9167	0.7014	0.5278
BER	0.7459	0.8676	0.9210
MSE	0.4171	0.7171	0.9171
NMSE	0.2652	0.7652	0.5652

Artistic Image Classification: An Analysis on the PRINTART Database

Artistic image understanding becomes popular in scientific researches as it is interdisciplinary. Usually, an artistic image contains global, local and pose annotations. In this paper, authors addressed an annotated database composed of artistic images and would let computer vision to understand these images.

The data used in this paper is composed of 988 images with global, local and pose annotations. All images are collected from Artstor digital image library and annotations are provided by art historians.

The method in this paper is using different approaches and get a best result.

Random. This lets random global annotation gets into visual classes, and using Hamming distance between query and test image annotations to give a rank. For local and pose annotation are achieved by selecting training image with smallest value.

Bag of Features. This is based on SVM(support vector machine). It requires to train each classifier for each label. The penalty factor of SVM for slack variables is calculate via cross-validation.

Label Propagation. This method encodes the similarity between pairs of images and then estimates the annotations of test image. This method is investigated the most by the authors.

Inverted Label Propagation. It inverts the problem and is able to produce global, local and pose annotations simultaneously. This method returns a vector to represent the possibility of landing in one of training images.

Matrix Completion. This method is aiming to find minimize rank for matrix represents annotation and features.

Structural Learning. This method is based on margin maximization quadratic problem.

The result of comparison is that inverted label propagation has the best result. The advantage of having inverted label propagation is its superior performance can explain the similar images from same theme which is suitable for our project as well. The distracted driver has many categories including drinking, using phones, looking back seats. If inverted label propagation can have good results telling same theme, it can generate good accuracy for our project (Costeira, 2012).

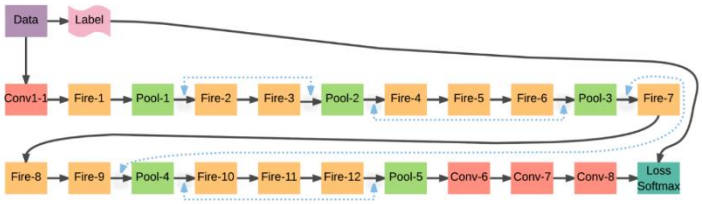
Name: Yanzhang Ma, NetID: 17ym24, Student#: 20090412

Compressed Residual-VGG16 CNN Model for Big Data Places Image Recognition

The issue addressed in this paper is a compressed the convolutional neural network Residual Squeeze VGG16. This model is based on the earlier VGG16 and by compressing VGG16 network, the improvements are as follows: 1. Small model size, 2. Faster speed, 3. The use of residual learning perform faster convergence, better generalization and solved the degradation issue, 4. Matches the accuracy of non-compressed model.

The image dataset is based on the large-scale MIT Places365-Standard image dataset created by MIT Computer Science and Artificial Intelligence Laboratory. There are 1,803,460 total training images in this dataset, and with 50 validations classed and 900 test classes and range from 3,068 to 5,000. The dataset is mainly consisted of images labeled with place or name.

The methodology used is the introduced Residual Squeeze VGG16 image shown below. This model contains 12 fire modules and 4 convolutional and 3 fully connected layers of VGG16. Compared to the original VGG16 model, a scale layer is attached to all fire modules. Then the second convolutional layer of VGG16 is replaced by one fire module. The fire module is proposed by Iandola et al, and its construction is composed of a squeeze convolution layer (1x1 filters) fed to an expanded layer (1x1 and 3x3 filters). In addition, residual connections are attached to four locations. In the image shown below, Pool-1 to Fire-3, Pool-2 to Fire-6, Pool-2 to Fire-9 and Pool-4 to Fire-12. The training is based on Berkeley Vision and Learning Center's open-source deep learning framework, Caffe. By pairing Caffe and an open-source deep learning GPU training system, NVIDIA DIGITS, allows users to build and examine artificial neural networks with real-time graphical representations.



The result is shown in the table below. It is smaller in size and faster in duration.

Network	Top-1 Validation %	Top-5 Validation %	Duration	Size	No. of Epoch
VGG16	54	84.3	3 Days 16 Hours	10.6 GB	20
Proposed Residual Squeeze VGG16	51.68	82.04	2 Days 19 Hours	1.23 GB	50

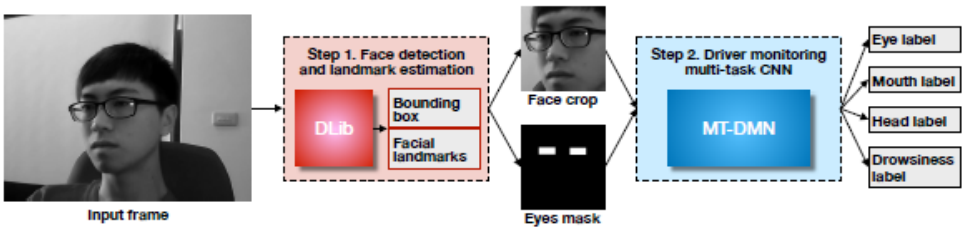
In conclusion, this Compressed Residual-VGG16 model combines three methods: VGG16, residual learning and Squeeze technique to perform a result that is very close to VGG16's accuracy also faster and smaller. And in our project, we intend to use VGG16 model and this model can be an improvement.

A Multi-Task CNN Framework for Driver Face Monitoring

The issue addressed in this paper is a vision-based Multi-Task Driver Monitoring Framework that simultaneously analyzes head pose, eyes and mouth status, also the drowsiness level of the driver. Which is similar to our selected topic, and we might get inspiration on this model.

The dataset is NTHU Drowsy Driver Detection video dataset, which is a database providing annotations for drowsiness, head, eyes, and mouth status. The database consists subjects of both genders and different ethnicities with various facial characteristics. The training set has 360 video clips of 18 subjects. The evaluation set consist of 20 video clips of 4 subjects. Frame level annotations are status like drowsiness status might be stillness or drowsy, head status could be stillness, nodding, or looking aside, eyes status are stillness or sleepy-eyes, mouth status is one among stillness, yawning and talking/laughing.

The methodology used is first use the hierarchical Temporal Deep Belief Network to separate drowsy video sequences from non-drowsy sequences. After that a combination of three CNNs has been proposed for predicting four class labels of various drowsiness status. Then evaluate it on the database. Multi-Task Cascaded Convolutional Neural Network (MTCNN) and Driver Drowsiness Detection Network (DDDN) are combined for the following steps. First step, MTCNN will detect the driver's face and locate 5 marks. The second step, DDDN will take the image of left eye and the mouth as input, separate it into three classes (normal, drowsy and yawning). A three hidden layers MLP classifier take marks coordinates as input to detect drowsiness.



The results are divided into different recognition tasks, including: eyes, head, mouth and drowsiness status. The accuracy of eye status is 75.91%, head status is 95.53%, mouth status is 89.93% and drowsiness status is 75.73%.

In conclusion, this model is an ensemble of several CNNs and divide tasks into different classes to get better accuracy. Our project of identify whether the driver is distracting can also use similar method.

Name: Jingyi Cheng NetID: 17jc66 Student Number: 20090394

Transfer learning using VGG-16 with Deep Convolutional Neural Network for Classifying Images

The main goal of the author is introducing the concept of transfer learning and applying Convolutional neural network (CNN) and VGG-16 model architecture for solving problems like image classification. The reason is traditionally, the widespread assumption in the field of machine learning of training data and test data must have identical feature spaces with the underlying distribution may not be tenable. Transfer learning reuses a pre-trained model on a new problem. Obviously that it can be more efficient by saving training time, better performance of neural networks (in most cases), and not needing a lot of data.

The dataset this paper takes into use are images with a small number of training samples per category. In this paper, the author uses 5000 data images of cats and dogs out of 25000 images for training and 2000 images for validation and builds an efficient model that can categorize images into bins of cats and dogs separately, which is a problem related to image classification.

The methodology used by the author can be summarized into firstly, pass images into CNN while go through several layers in the network and generates the CNN. In the convolution layer, convolution is processed, and feature map is generated by sliding kernel matrix on the input matrix as filter. Applying rectified linear unit activation function (ReLU) after in the nonlinearity layer. Then is the pooling layer for down sampling. Then the fully connected layer to compile the data from last layer to get the final output of CNN.

Then slightly adjust the model by using image augmentation technique.

Finally use VGG-16 model which had been pretrained in ImageNet Project to classify image and check accuracy for training data and validation data. This is the VGG-16 model architecture:

Result given by the author that after trained CNN the validation accuracy is 72.40%. Then achieve 79.20% after image argumentation, lastly, achieve accuracy of 95.40% which is quite impressive.

Overall, pre-trained networks for VGG are available freely and would be suitable for moving detection applying to the indication of distraction of drivers, although it maybe slow but shows good accuracy from the research.

Driver Distraction Detection Method Based on Continuous Head Pose Estimation

In this essay, authors take in the problem which is related to our project. After doing some research, I found that driver distraction can be detected by several features such as head position, body position, eye movement and so on. They want to solve the problem of Driver Distraction Detection Method through Continuous Head Pose Estimation. This can be the method we consider to be applied in our project.

It compares single regression and classification combined with regression in accuracy. And trained four classic networks with 300W-LP (the synthesized large-pose face images from 300W). and AFLW datasets (they are datasets employed as the training set for face attribute recognition and landmark localization.) HPE_Resnet50 with the best accuracy is selected as the head pose estimator and applied to test 20000 cases from ten-category distracted driving dataset SF3D. SF3D's ten categories of distracted driving based on a distracted driving identification competition on Kaggle which is the same as ours project, they choose 2000 images of each category of driving, that is the reason why I believe this essay is closely related to our project. Also, there is a collection of actual driving images (Driver_Imgs) from "BeiDou +" vehicle video surveillance platform in Jiangsu Province, China, for testing the effect of actual driving.

The method used by the authors can divide into two parts, first is the head pose estimation then distraction detection. In head pose estimation, they use CNN for mapping from image to pose space. Then is distraction detection based on videos and combined with SF3D.

And from their experiment part 1, they find that the average error of HPE_Resnet50 in AFLW2000 (test set) is 6.17°, while with the deepening of the layers, error deducts. and that there is an average difference of 12.4° to 54.9° in the Euler angle between safe driving and nine kinds of distracted driving on SF3D.

From part 2, they verify the application of head pose estimation in the actual driving image. Finally, after testing, the error is within an acceptable range, which can be used not only to analyze the images collected in the experimental environment but also pictures from the real world.

In this essay, they take the same indicators and categories of distracted driving as our project, also use Resnet which is a recommended model, and dataset can be reached in an easy way, so I believe there is something we can learn from this essay.

Name: Daniel Jang Student#: 20096632
Recurrent Convolutional Neural Network for Object Recognition

The goal of the Recurrent Convolutional Neural Network (RCNN) is to improve the performance of existing convolutional neural networks (CNNs) by introducing recurrent convolutional layers (RCL). CNNs have achieved better performance in the field of computer vision compared to other approaches, but the use of RCLs is relatively uncommon. Instead, deeper and more complex CNNs are built to improve performance. CNNs are primarily feed forward networks so to further emulate the biological neural network, recurrent connections can be introduced. The result is that object recognition becomes a dynamic process though the input remains static. Feed-forward models can only capture the context of an image in the higher layers, but this information cannot modulate the activities of lower layers. In pursuing a recurrent approach, RCNN seeks to improve performance over CNNs without increasing the number of parameters in the network. The RCNN model was evaluated using several benchmarks including CIFAR-10, CIFAR-100, MNIST and SVHN, and compared against a number of other CNN models.

The key component of RCNN is the RCLs which evolve over discrete time steps. The input for each RCL is a combination of both the current new inputs and the outputs of the previous time step, as well as both sets' respective weights. This allows for units in the network to be influenced by its neighbours. Convolutions are applied to the data to create feature maps which are used to identify and classify the features in the image. The outputs must be normalized to keep the values in the network from exploding in size. RCLs are stacked with optional max pooling layers interspersed which serve to down sample the output of the previous layer. The connections between RCLs are feed-forward. Finally, a Softmax layer is used to get the final prediction based on the features of the image. Training is performed by minimizing the cross-entropy loss function using the backpropagation through time (BPTT) algorithm. The paper includes several equations and diagrams which describe the structure and process of the network.

RCNN was compared to other CNN models such as Maxout, NIN, and DSN. When tested over the four evaluation datasets, RCNN consistently achieved lower testing error than the other models. In the case of NIN and DSN, RCNN had a comparable number of parameters. With Maxout, RCNN had substantially less parameters. With a recurrent approach, there are several advantages. Units within a layer are able to incorporate context information in an increasingly larger area, while in regular CNNs, this area is fixed, and only grows larger when going to the next layer. Through unfolding RCNN, it can be arbitrarily deep, while the number of parameters remains constant.

Improved Inception-Residual Convolutional Neural Network for Object Recognition

The goal of the Inception Recurrent Residual Convolutional Neural Network (IRRCNN) is to improve performance in the field of object recognition in machine vision using a recurrent approach over a more conventional deep convolutional neural network while improving accuracy with less complexity than other common or similar approaches. Conventional Deep Convolutional Neural Networks (DCNN) have produced very good results in the field of computer vision, but the use of recurrent convolution layers (RCL) is uncommon The IRRCNN model seeks to be an improved DCNN model based on inception, residual, and recurrent architectures, including the Recurrent Convolutional Neural Network (RCNN) architecture. The IRRCNN model was evaluated using several benchmarks including CIFAR-10, CIFAR-100, TinyImageNet-200, and CU3D-100.

The overall structure of the IRRCNN network involves and input image sent through several convolution layer, before being sent through a IRRCNN block and transition block an arbitrary number of times recurrently before a Softmax is applied at the output to get the prediction of the model. IRRCNN blocks are inception-residual units which includes RCLs. Through many discrete time steps, the inputs to these blocks go through RCLs which use the ReLU activation function. The outputs of these RCLs merged by are concatenation and are summed with the inputs before being used again as input for the next time step. Transition blocks are used differently depending on their position in the network. Operations that can be done include convolution, pooling, and dropout. Down-sampling is done here using the overlapping max-pooling operation for regularizing the network. Choices in the sizes of the convolution filters were done to reduce network parameters and to improve non-linearity. The output of these layers can either be sent to another IRRCNN-transition pair, or to the final output. Softmax is applied at the final output to get the prediction. The paper includes several formulas which describe these steps computationally.

Evaluation was done with four benchmarks, and IRRCNN's performance was compared to various other models including RCNN, EIN, EIRN, and showed improved performance over each. IRRCNN showed a 4.53% improvement in recognition accuracy over RCNN with the CIFAR-100 set. IRRCNN set out to improve accuracy over other DCNN approaches while having a similar number of network parameters or less. Compared to EIN and EIRN, the models share a similar number of network parameters, but has improved recognition accuracy.

IRRCNN draws inspiration from and competes against RCNN. In testing each network against the same datasets, IRRCNN consistently outperforms RCNN.

Summary table:

Paper ref	problem	data	method	result	Pros/cons
Hybrid Deep VGG-NET Convolutional Classifier for Video Smoke Detection	using cameras to find out existence of fire	1162 frames of a video containing fire and smoke	Use VGG-net CNN to train and Ostu to identify fire	Has a minimum precise of 0.73 and when in use nearly get 1.0	Can minimize the training set and time.
Artistic Image Classification: An Analysis on the PRINTART Database	Let computer understand what do artistic images mean (match the label given by artist)	988 images of artistic each with global, local and pose labels.	Use random, bag of feature, label propagation, inverted label propagation, matrix completion, structure learning to test accuracy	Inverter label propagation has highest accuracy	Can identify pose and theme of picture, has high accuracy
Compressed Residual-VGG16 CNN Model for Big Data Places Image Recognition	Compressed Residual-VGG16 CNN Model for Big Data Places Image Recognition	Improved VGG16 model namely Residual Squeeze VGG16	MIT Places365-Standard image dataset	Top 1 accuracy 51.68 Top 5 accuracy 82.04 Duration 2 Days 19 hrs. Size 1.23GB	Pros: faster in speed and smaller in size Cons: less accuracy
A Multi-Task CNN Framework for Driver Face Monitoring	A vision-based Multi-Task Driver Monitoring Framework	NTHU Drowsy Driver Detection video dataset	Ensemble of MT-DMF for face monitoring, MT-DMN predicts facial behavior; a three hidden layer MLP classifier for training	Accuracy eye status 75.91%; head status 95.53%; mouth status 89.93%; drowsiness status 75.73%	Pros: very well-developed facial monitoring model Cons: Accuracy could be low if not enough frame are provided
Transfer learning using VGG-16 with Deep Convolutional Neural Network for Classifying Images	Instead of traditional learning method, transfer learning reuses a pre-trained model on a new problem to increase efficiency for image classification	5000 images of cats and dogs out of 25000 images for training and 2000 images for validation	CNN (convolution layer, nonlinearity layer, pooling, fully connected layer) for training, image argumentation tools, VGG-16 model (pretrained in ImageNet Project)	Convolutional neural network gives validation accuracy of 72.40 %, after image achieved accuracy of 79.20 %, leverage VGG-16 trained on huge dataset of images and fine-tuned with image augmentation to	Increase in validation accuracy but may take relatively long time to process

				achieve accuracy of 95.40%	
Driver Distraction Detection Method Based on Continuous Head Pose Estimation	Analyze the difference in head pose between safe driving and distracted driving	300W-LP, AFLW are training set and AFLW2000 is testing set, SF3D for verify	Head pose estimation by CNN, HPE_Resnet50 was used to test SF3D for difference between safe driving and distraction driving	Average error of HPE_Resnet50 in AFLW2000 is 6.17° and that there is an average difference of 12.4° to 54.9° in the Euler angle between safe driving and nine kinds of distracted driving on SF3D	It uses data without preprocessing, the difference showed can be a part we can consider implementing in our project, also they choose the same categories for analyzing
Recurrent Convolutional Neural Network for Object Recognition	Improve the performance, and computational size of existing convolutional neural networks (CNNs) by introducing recurrent convolutional layers (RCL)	CIFAR-10 CIFAR-100 MNIST SVHN	Stack 5 RCL layers on top of input convolution layer with pooling layers between.	RCNN achieved the lowest percentage of test errors on the CIFAR-100 data set with 31.75% using 160 feature maps in each layer.	Has an equal or lower total number of network parameters with better performance over CNNs.
Improved Inception-Residual Convolutional Neural Network for Object Recognition	Improve accuracy and reduce network parameter complexity in object recognition tasks using aspects of recurrent convolutional layers (RCLs), inception networks, and residual networks.	CIFAR-10 CIFAR-100 TinyImageNet-200 CU3D-100.	Recurrently cycle through IRRCNN layers and transition blocks to apply. Merge output of many convolution operations at each cycle.	IRRCNN showed a 4.53% improvement in recognition accuracy over RCNN with the CIFAR-100 set	Improved performance over RCNN. Increased network complexity due to inception and residual components.

Reference:
[1] CosteiraCarneiroNuno Pinho da SilvaAlessio Del BueJoão PauloGustavo. (2012). Artistic Image Classification: An Analysis on the PRINTART Database. doi:https://doi.org/10.1007/978-3-642-33765-9_11

[2] Princy Matlani, M. S. (2019). Hybrid Deep VGG-NET Convolutional Classifier for Video. Retrieved from https://www.researchgate.net/profile/Manish-Shrivastava/publication/334542934_Hybrid_Deep_VGG-NET_Convolutional_Classifier_for_Video_Smoke_Detection/links/5d601a68a6fdccc32ccca596/Hybrid-Deep-VGG-NET-Convolutional-Classifier-for-Video-Smoke-Detection.pdf

[3] Qassim, H., Verma, A., & Feinzimer, D. (n.d.). Compressed residual-VGG16 CNN model for Big Data Places Image Recognition. Retrieved October 21, 2021, from https://ieeexplore.ieee.org/document/8301729

[4] Celona, L., Mammana, L., Bianco, S., & Schettini, R. (2018, December 17). A multi-task CNN framework for driver Face monitoring. Retrieved October 21, 2021, from https://ieeexplore.ieee.org/document/8576244

[5] Tammina, Srikanth. (2019). Transfer learning using VGG-16 with Deep Convolutional Neural Network for Classifying Images. International Journal of Scientific and Research Publications (IJSRP). 9. p9420. 10.29322/IJSRP.9.10.2019.p9420.

[6] Zuopeng Zhao, Sili Xia, Xinzheng Xu, Lan Zhang, Hualin Yan, Yi Xu, Zhongxin Zhang, "Driver Distraction Detection Method Based on Continuous Head Pose Estimation", Computational Intelligence and Neuroscience, vol. 2020, Article ID 9606908, 10 pages, 2020. https://doi.org/10.1155/2020/9606908

[7] Liang, M., & Hu, X. (2015). Recurrent convolutional neural network for object recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 3367-3375).

[8] Alom, M. Z., Hasan, M., Yakopcic, C., Taha, T. M., & Asari, V. K. (2020). Improved inception-residual convolutional neural network for object recognition. Neural Computing and Applications, 32(1), 279-293.

DataFlow Diagram