ADL 2020 PROJECT-III REPORT

Mayank Gupta and Bazil Ahmed

{mayankg, bazilahmed}@iisc.ac.in

ABSTRACT

Generating explanations for the decision taken by deep models has recently gained significant attention by the research community. One of the work in this direction is Grad-CAM which offers visual explanations for the decisions taken by the large class of CNN based models. While Grad-CAM has given landmark results for image based tasks, it has not been explored much for video dataset. In this project, we have explored the use of Grad-CAM for video based classification task. We have explored 3 different models, first model uses 3D convolution over time and space, second uses LSTM over 2D CNN model without attention and the third model uses attention over the second model. In another trajectory we have tried to explore Grad-CAM for image based regression tasks. Given a set of images captured from the camera of a moving vehicle, we need to predict the ego-motion. In this task, we take gradient of the negative of the regression loss with respect to the activation maps of the last layer.

***Index Terms*—** Explainability, Grad-CAM, LSTM, CNN

1. INTRODUCTION

Deep learning has given significant advancements in various domains by beating state of the art. Despite of this success there has been very limited acceptability rate of deep learning solutions for the real world problems. This is due to the black-box nature of deep learning models which offers no explainability for the outputs it produces. There has been many attempts in this direction [[1], [2]] for offering visual explanations for the decision taken by deep CNN models for image based task. In this project, we are looking for visual explanation for the video based task. We also explore Grad-CAM for regression tasks. There are some regression tasks such as steering angle prediction[3], age-prediction[4] etc., which can be explained by converting the problem into a classification problem. But tasks such as visual odometry, trajectory prediction, etc. cannot be converted into classification tasks for explanation. So in this project we also try new ways to apply Grad-CAM to these regression tasks which cannot be converted into classification problem .

Thanks to Google for colab platform.

2. TECHNICAL DETAILS

We have worked with 3 models, the next subsections describe the same.

2.1. ResNet 3D (pre-trained)

We picked a pre-trained version of ResNet 3D which is a 18 layered 3D spatiotemporal convolution network as proposed in [5]. This network takes a sequence of frames and does 3D convolution over both the spatial and temporal space. We started with a input of 10 frames from a video sample and passed it through the ResNet 3D network and retrieved the activation maps from the last convolution layer before the average pooling layer for applying Grad-CAM. The dimension of the obtained activation maps is [batch size, 2, 128, Height, Width], here 2 represents the temporal dimension, 128 is the number of channels in each temporal dimension.

2.2. LSTM over convolution (trained)

We have used pre-trained ResNet 18 [6] for feature extraction of the frames extracted from a video sample. At each time stamp a frame is passed through ResNet 18, we extract the features obtained from the average pooling layer and then pass it to LSTM layer. The last hidden state of the LSTM is used for the classifying the HMDB51 dataset. In case of attention, we form a context vector out of all the hidden states of the LSTM which is then used for classification. We train this whole network (including ResNet 18) on HMDB dataset with and without attention. We applied the Grad-CAM on the activation map obtained from the last convolution layer. Also for taking gradient with respect activation maps corresponding to each frame we used 10 (number of input frames) instances of the trained CNN model.

2.3. Visual Odometry: a regression task

We are using pretrained SfMLearer[7] for this task. In this network, we have 2 main parts : *First*, we use a depth networks to predict the depthmap of the scene. *Secondly,* we use a pose network to predict the odometry parameters i.e. R(rotation) and t(translation). Both odometry parameters have 3 DOF. SfMLearer uses view-synthesis for loss function and therefore this method is completely unsupervised. In this

model, we calculate the loss between the predicted reference image and ground-truth reference image. For this task, we calculate the weights of the activation map using the gradient of the negative regression loss. We are using sequence 9 of the KITTI visual odometry dataset for our task.

2.4. Grad-CAM

We implemented Grad-CAM on the above models from scratch. We experimented with Grad-CAM on activation maps from different layers and observed the results. We observed that as we go towards the initial layers the gradients were uniform along the channel and hence did not yield any meaningful insights with respect to any class. In case of 3D convolution of the first model, as the temporal dimension reduced from 10 to 2 in the last layer, we could not find one to one map of heatmaps with the input frames. So we superimposed the first heatmap with the first frame of the input and second frame of the heatmap with the sixth frame of the input. The results were good considering the fact that focused part of the object did not move much across frames due to smaller duration of the video samples.

3. RESULTS

We observed consistent results for the video datasets as it was for image based datasets. The LSTM model with attention was able to perform better in terms of accuracy for the classification task also the focus of the Grad-CAM improved see figure 1. We also observed that attention weights were high for frames which contained the objects relevant for the classification task. While attention helps in selecting the relevant frames, Grad-CAM is able to highlight the region to focus in that particular frame. We conclude that for finding visual explanations for video based tasks, attention can be used for shortlisting the frames in order of relevance, and then Grad-CAM can be applied on those frames to understand the decisions of the deep model.

We observed the Grad-CAM with respect to different class than the ground truth and the results were meaningful, see figure 2. One interesting thing to note in this figure is the fact that model is able to recognize the robotic hand wave and that too at different resolution (see the bottom right frame of (a)).

We also observed the results for the ResNet 3D model and it followed the same pattern as the LSTM model results, see figure 3.

On the regression task, the domain knowledge of the task says that model should focus on non-moving objects which have distinctive features. In our case we see that model is not focusing on the moving vehicle as well as the road which does not give any textual context for any motion. At this point we can not draw any conclusion on regression task but we see some positive signs.

4. CONTRIBUTIONS

We have explored the use of Grad-CAM for video based classification task. We have explored 3 different models, first model uses 3D convolution over time and space, second uses LSTM over 2D CNN model without attention and the third model uses attention over the second model. We implemented Grad-CAM from scratch and are able to draw interesting insights for the video dataset. We implemented Grad-CAM for regression task, Odometry, by taking gradient with respect to the negative of the regression loss.

5. RESOURCES

`https://pytorch.org/vision/0.8/models.html`
`https://serre-lab.clps.brown.edu/resource/hmdb-a-large-human-motion-database/`
`https://deepmind.com/research/open-source/kinetics`

6. REFERENCES

[1] Ramprasaath R. Selvaraju, Abhishek Das, Ramakrishna Vedantam, Michael Cogswell, Devi Parikh, and Dhruv Batra, "Grad-cam: Why did you say that? visual explanations from deep networks via gradient-based localization," *CoRR*, vol. abs/1610.02391, 2016.

[2] Bolei Zhou, Aditya Khosla, Àgata Lapedriza, Aude Oliva, and Antonio Torralba, "Learning deep features for discriminative localization," *CoRR*, vol. abs/1512.04150, 2015.

[3] Mariusz Bojarski, Davide Del Testa, Daniel Dworakowski, Bernhard Firner, Beat Flepp, Prasoon Goyal, Lawrence D. Jackel, Mathew Monfort, Urs Muller, Jiakai Zhang, Xin Zhang, Jake Zhao, and Karol Zieba, "End to end learning for self-driving cars," *CoRR*, vol. abs/1604.07316, 2016.

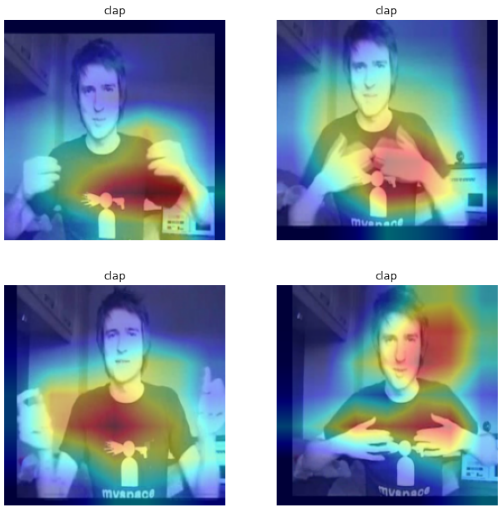
[4] Gil Levi and Tal Hassner, "Age and gender classification using convolutional neural networks," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR) workshops*, June 2015.

[5] Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri, "A closer look at spatiotemporal convolutions for action recognition," *CoRR*, vol. abs/1711.11248, 2017.

[6] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," *CoRR*, vol. abs/1512.03385, 2015.

[7] Tinghui Zhou, Matthew Brown, Noah Snavely, and David Lowe, "Unsupervised learning of depth and ego-motion from video," in *Computer Vision and Pattern Recognition*, 2017.

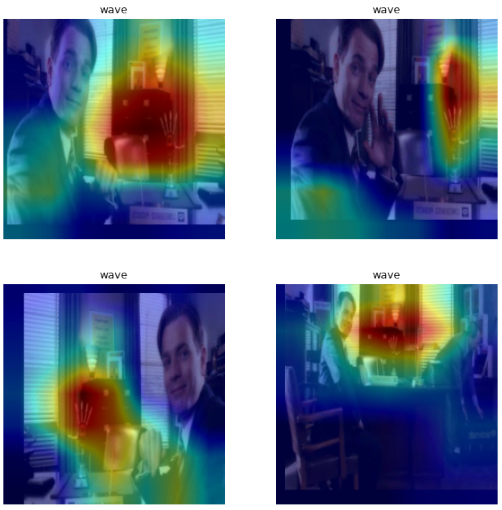
7. APPENDIX



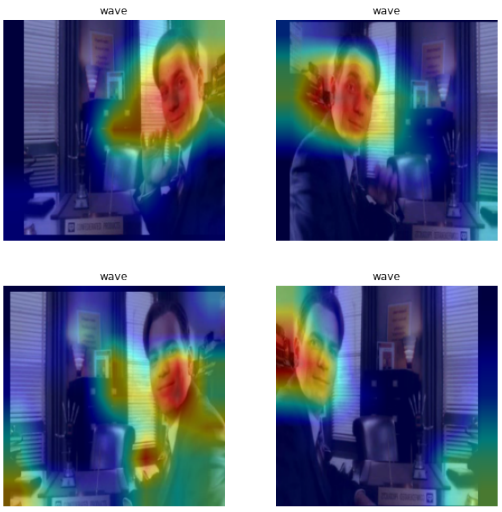
(a) Vanilla LSTM-CNN model

(b) LSTM-CNN model with attention

Fig. 1. Grad-CAM results of LSTM-CNN model with respect to the original class 'Clap' when the model predicted correctly



(a) Grad-CAM with respect to 'Wave'



(b) Grad-CAM with respect to 'Smile'

Fig. 2. Grad-CAM results of LSTM-CNN model with original class being 'Wave'

(a) Grad-CAM with respect to 'Arranging Flowers'

(b) Grad-CAM with respect to 'Brushing Hair'

Fig. 3. Grad-CAM results for ResNet 3D model with original class being 'Arranging Flowers'

(a) Grad-CAM with respect to negative of regression loss

(b) Original Image

Fig. 4. Grad-CAM results on Odometry, a regression task