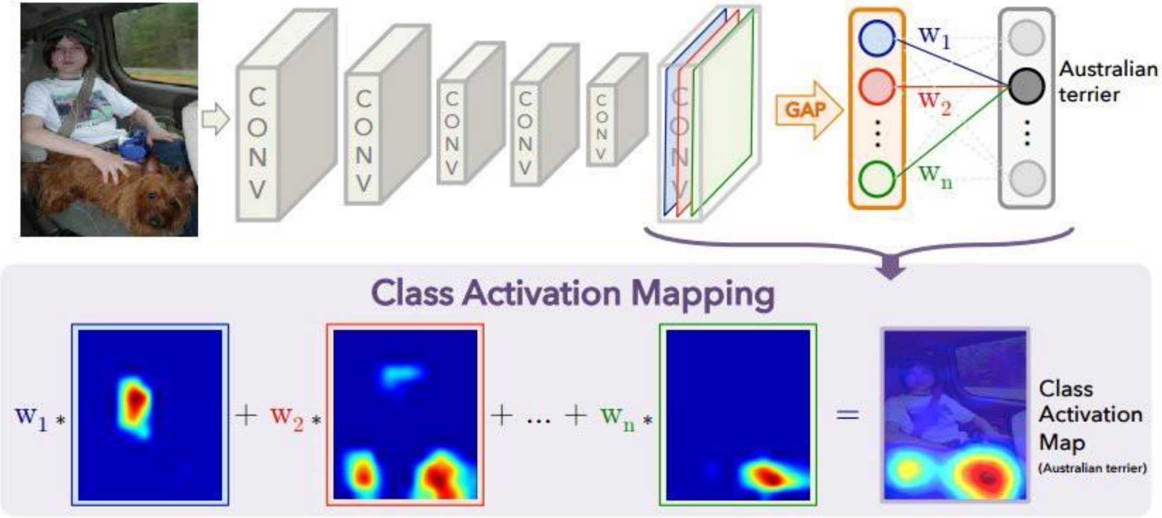
Grad-CAM for Video and Regression Tasks

Mayank Gupta and Bazil Ahmed

Background



Explaining the decisions of neural networks is an active are of research. This is an important topic as many applications e.g. healthcare require the explainability of neural networks

Grad-CAM is a very popular tool that is used to analyse the results of CNN for classification tasks for images.

Although Grad-CAM used a lot for images, there are not a lot of works that use Grad-Cam for videos and regression based tasks. So, in this project we explore some ways in which Grad-CAM can be applied to these tasks and show some results on these tasks.

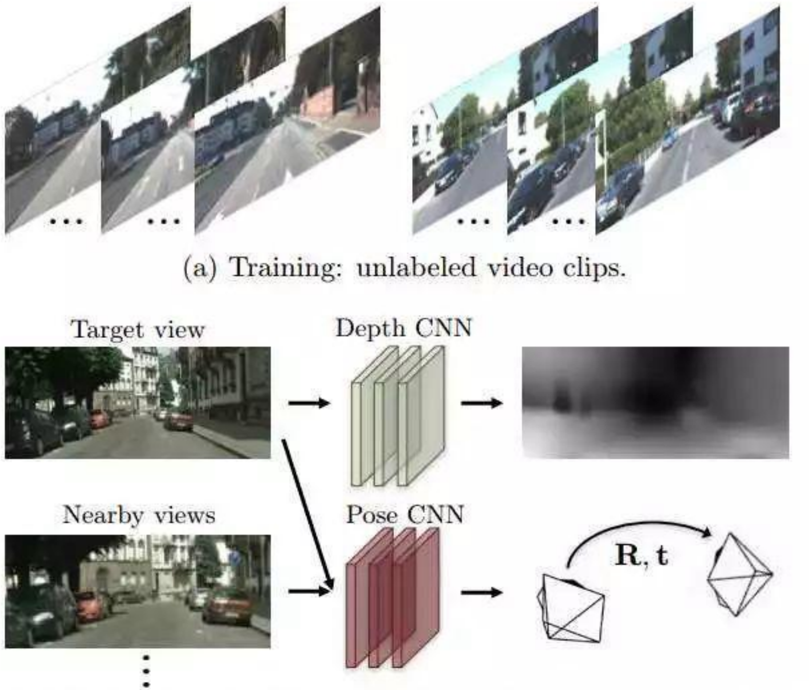
Proposed Approach

1. In this work, we apply Grad-CAM for videos using following approaches:

- We divide the video samples into equally spaced image frames which gets processed through different models
- First model is ResNet 3D which uses 3D convolution over space and time.
- Second model consists of vanilla LSTM over 2D CNN
- Third model uses LSTM over 2D CNN model with attention

2. We also propose a method for using Grad-CAM for regression. We show our results on Visual Odometry task which is an image based regression task.

Technical Details



(a) Training: unlabeled video clips.

(b) Testing: single-view depth and multi-view pose estimation.

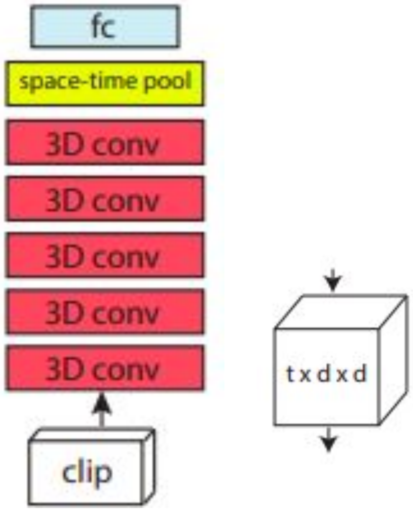
We are using SfMLearner for visual odometry task. It has 2 main components:

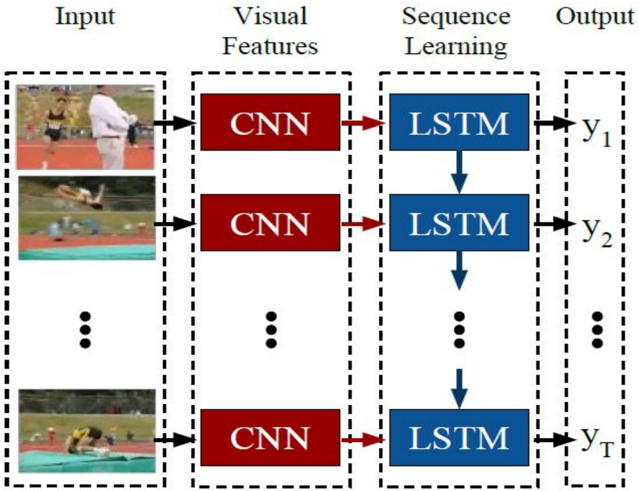
- DepthNet: predicts the depth map of the scene.
- PoseNet: predicts odometry parameter from 2 scenes.

Loss function used is based on view-synthesis.

We calculate the weights by taking the gradient of the negative loss function with respect to activation maps of the last layer for this task.

Technical Details





- ResNet does 3D convolution over both the spatial and temporal space hence one to one map from activation maps to original frame is lost
- Gradient with respect to initial layers is uniform for all pixels, hence taking gradient with respect to initial layers for one to one map is meaningless.

- Can easily map the contribution of each frame using Grad-CAM
- Multiple instances of trained CNN is needed for finding Grad-CAM for each frame, hence difficult to implement.
- Attention weights ranks the frames in order of relevance , also the Grad-CAM output on individual frames improves due to increase in accuracy.

Contributions (Novelty)

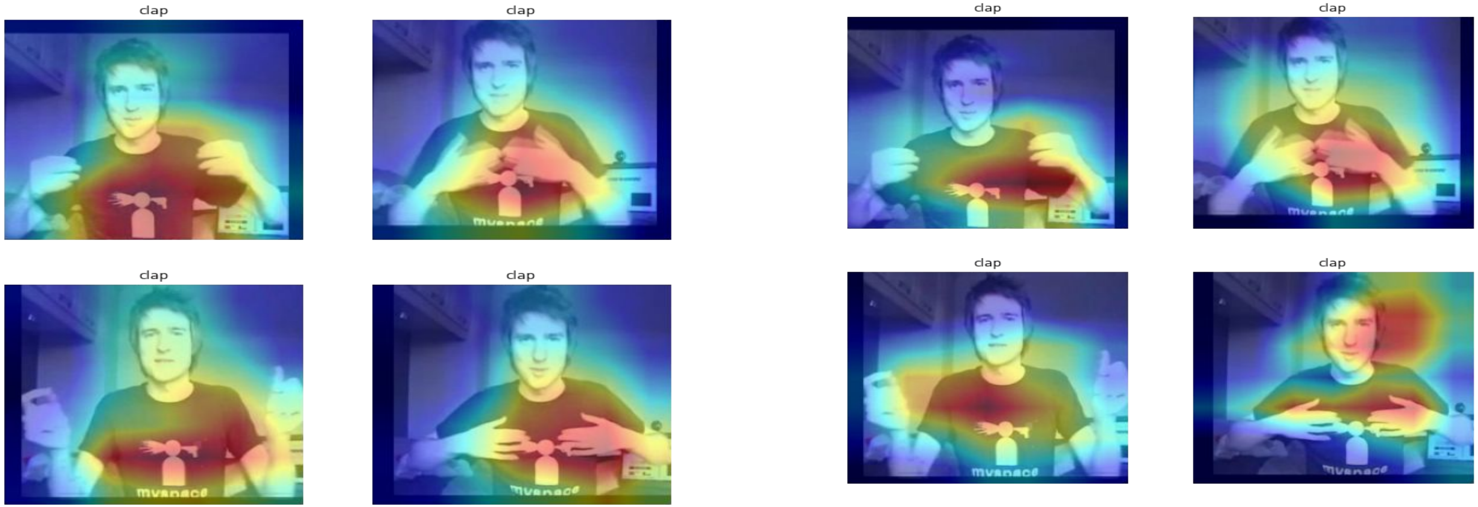
1) We have explored the use of Grad-CAM for video based classification task. We have explored 3 different models :
    - First model uses 3D convolution over time and space
    - Second uses vanilla LSTM over 2D CNN model
    - Third model uses LSTM over 2D CNN model with attention
2) We also explored Grad-CAM for regression task of monocular visual odometry by calculating the weights for activation maps using gradient of the negative of the loss function.
3) We implemented Grad-CAM from scratch for all above models and were able to draw interesting insights for both problems.

Results & Conclusion

- We observed consistent results for the video datasets as it has proven 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.

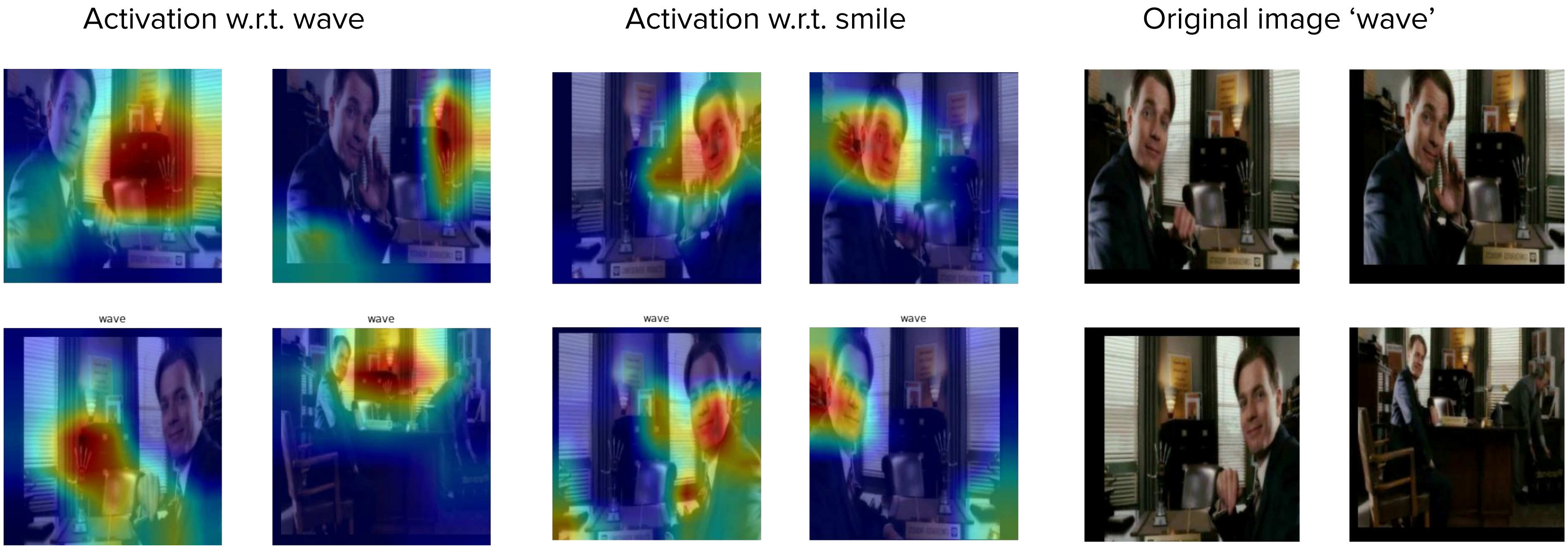
- We also observed that attention weights were high for frames which contained the objects relevant for the classification task.
- The Grad-CAM with respect to different class than the ground truth are meaningful and self explanatory.
- The results for the ResNet 3D model followed the same pattern as the LSTM model results.
- Results produced by Grad-CAM on visual odometry were mostly meaningful.

Results Visualized



LSTM-CNN model with(left) and without(right) attention with respect to original class ‘Clap’

Continued...



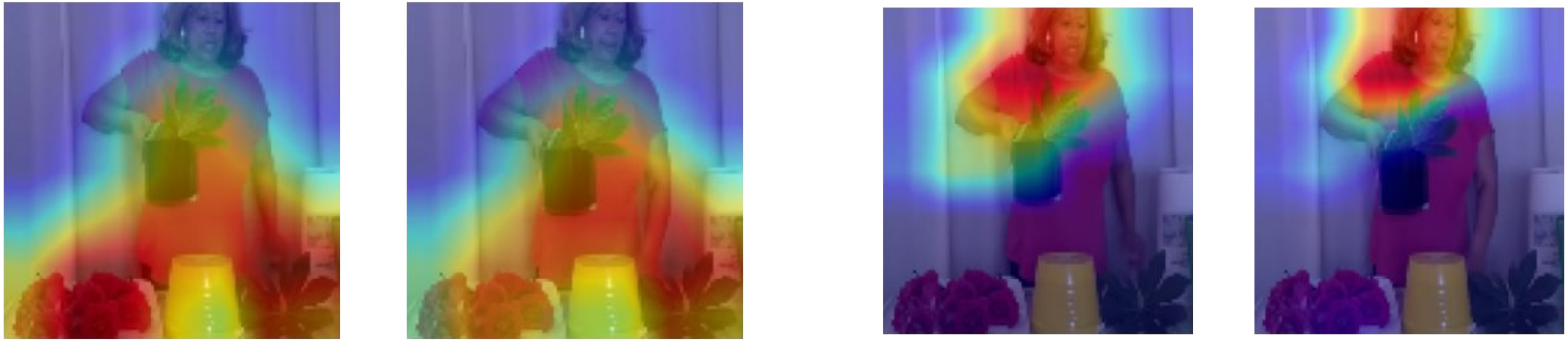
Activation w.r.t. wave

Activation w.r.t. smile

Original image ‘wave’

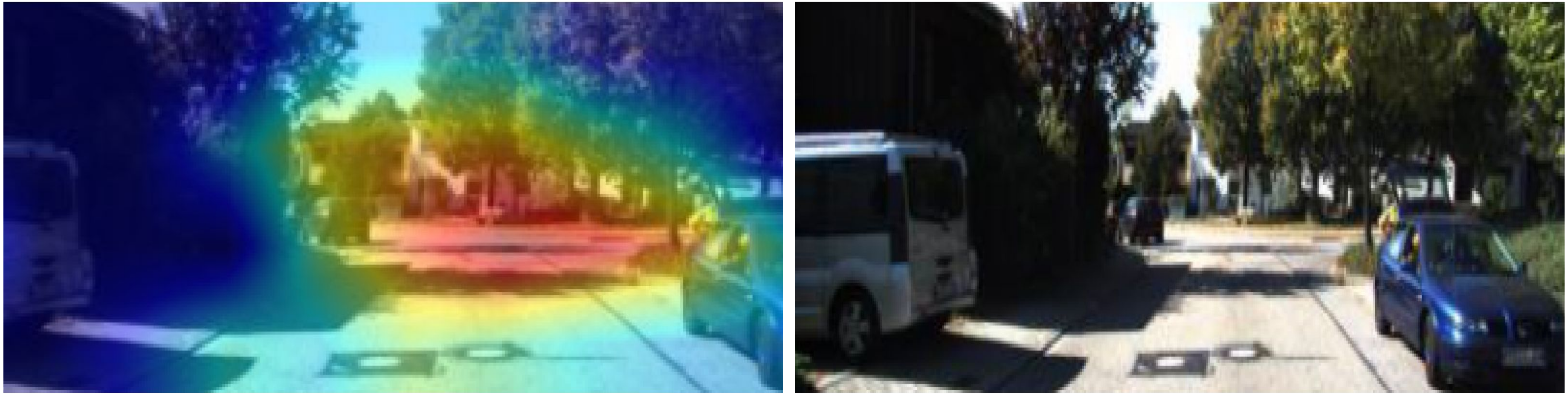
Grad-CAM with respect to original class ‘Wave’ and another class ‘Smile’

Continued...



Grad-CAM results for ResNet 3D, with respect to original class ‘Arranging flower’ (left) and ‘Brushing hair’

Continued...



Visual Odometry: An image based regression task