Saliency Guided Activity Recognition

Varun Gupta, Yasin Hasekioglu, Anand Rajaraman

December 2017

1 Introduction

Activity recognition is a very important and challenging problem that tries to track and understand the behavior of agents through videos taken by various cameras. There exist a number of methods such as optical flow, Kalman filtering, Hidden Markov models, etc., under different modalities such as single camera, stereo, and infrared. Vision-based activity recognition has found many applications such as human-computer interaction, user interface design, robot learning, and surveillance.

Through this paper, we propose an alternate approach to the problem of activity recognition, namely through information from saliency maps from first person viewpoint. Saliency maps help identify regions or objects in an image which attract the most attention and could essentially represent most of the features relevant for activity recognition. Based on this hypothesis, we present a systematic approach to obtaining saliency maps and predicting activity based on both saliency maps and RGB images. The proposed architectures in the paper have been derived from published papers which are cited. We present several experiments that were tried and certain interesting observations that were made in the results section.

2 Saliency Prediction Model

Visual saliency prediction is one of the most challenging and interesting problems in computer vision. The idea is to identify key salient points in 2D images and extract meaning from images as to where the attention could be. An extended application involving these salient points in a scene can be used for gaze tracking[2] from first person view. Identifying objects in a scene that drive instantaneous focus to the human eye, is still something machines need to learn to do well if they are to be used to make decisions for real world applications.

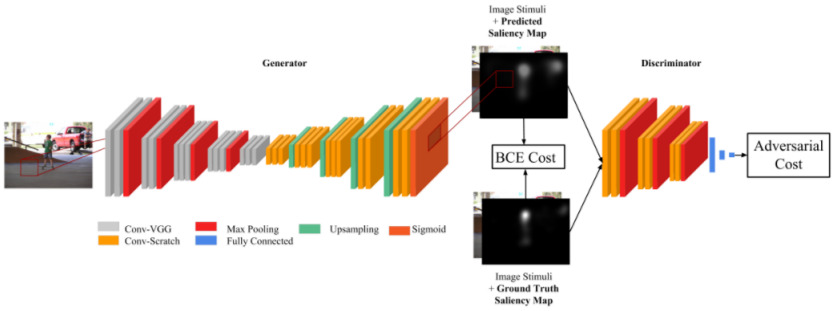


Figure 1: SalGAN Architecture[1]

2.1 SalGAN

We use a GAN[1] based deep network architecture called SalGAN for visual saliency prediction trained with adversarial examples. The first stage of the network consists of a generator model whose weights are learned by back-propagation computed from an appropriate loss functions over downsampled versions of the saliency maps. The generator model can be any of the standard architectures likes VGG-16, ResNet etc. The resulting prediction is processed by a discriminator network trained to solve a binary classification task between the saliency maps generated by the generative stage and the ground truth ones.

3 Activity Prediction model

The activity recognition model consists of a feature extractor and a classifier. A CNN feature extractor is used for extracting spatial information and an LSTM is used for temporal information gathering for activity classification using a linear classifier on the output feature maps. The entire architecture is described in the sections below and illustrated in Figure 2. Note that the input is not necessarily the RGB image. In our case, it is either RGBS (RGB + Saliency) or saliency overlapped RGB image.

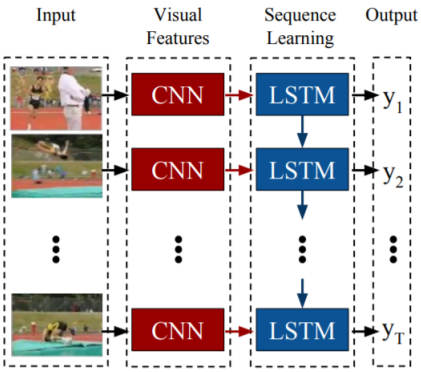


Figure 2: CNN + LSTM Architecture for activity recognition

3.1 CNN Feature Extraction

The CNN feature extractor was designed to incorporate 4 input channels and produce a 512-d feature vector output. In the first layer, the 4 channels were mapped to a 3 channel output followed by a partial pretrained VGG-16 feature extraction network that outputs a 16x12x512 feature map. Two different approaches that were tried to convert to a single 512-d vector are:

1. Global Average Pooling: Average pooling over the entire 16x12 feature map to output a single scalar. The idea is to incorporate global feature information for better activity prediction using fully connected LSTM. This idea would work better in experiments where saliency overlapped RGB images are used for activity prediction because of the sparsity in the input layer.

2. Multiple strided convolutions: Multiple average pooling layers to convert 16x12 feature maps to a single scalar. The idea is to incorporate partial global context which would work better when the input layer is not so sparse.

3.2 LSTM

The LSTM architecture basically consisted of 32 LSTM units that each carried a 512-d feature vector representation for each image. The output vector accumulated for each image after overlapping with temporal information from LSTM was used to classify into action labels using Fully Connected Layers. The exact architecture will be discussed in the experiments section once the training data is introduced.

4 Experiments

In this section, we present experiments on both saliency prediction as well as efforts taken towards activity recognition.

4.1 Model Setup and Data

The dataset we use for this saliency prediction is GeorgiaTech Egocentric Activites (GTEA Gaze+) dataset. This dataset consists of first person cooking videos of different recipes with multiple action labels. We are also provided with the gaze information of where the person is looking at any instant. The architecture mentioned above is trained on Salicon dataset with third person data. The image features extracted by the network from these images does not scale up to first person data as the feature definition is different in this setup as can be seen in Figure 4. The scale at which the saliency is learned by the pretrained model will not correlate to the feature space of our setup. Also we noticed that the third person network applied to our data generated multiple hotspots which will not aid our activity prediction. So we have to retrain the network with frames from our video so that the network learns this scale.

But the problem with retraining SalGAN on first person data is that we don't have true saliency map which is required for our generator training. To address this issue we use the ground truth gaze points to fit a gaussian filter around it to guide as ground truth saliency. The issue with gaze points is that it is incoherent with actual saliency as a person's gaze might sometimes not address the salient point in the image. We hope that our network doesn't get biased by this gaze and predicts the actual saliency.

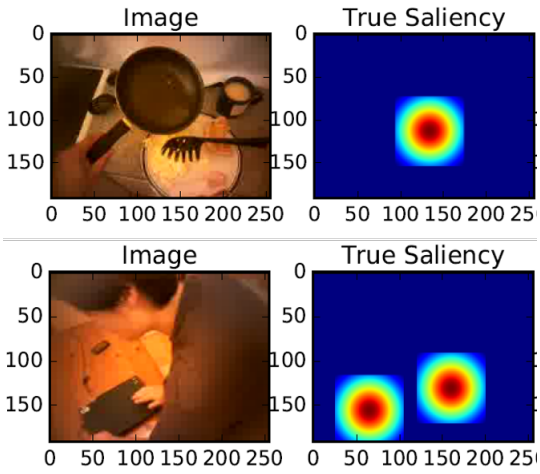


Figure 3: Ground truth saliency maps generated using gaze points

The results from our network as shown later indicates that the network

is actually able to learn better saliency maps compared to the synthetically generated ground truth saliency maps from gaze.

4.2 Retrain SalGAN

The SalGAN was retrained using frames from just one recipe which is a 10 minute long sequence. The training process followed in this GAN setup is first bootstrapped for 30 epochs by pure generator training. The encoder-decoder type generator model predicts a saliency map given an input image which is compared with the ground truth saliency by a Binary Cross Entropy loss. The adversarial training begins after 30 epochs. The appended image with predicted saliency is fed to the discriminator network to calculate a discriminator loss which is the score of fooling the network into predicting the predicted saliency as real. To aid the generation of larger spatial saliencies, the training of generator and discriminator is alternated every iteration during adversarial training.

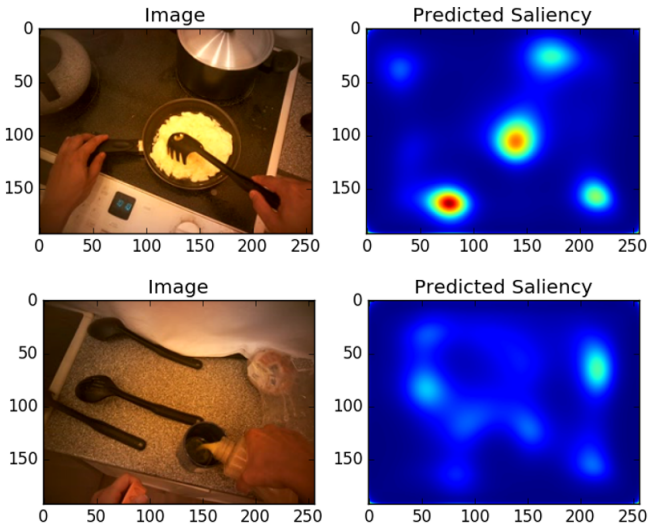


Figure 4: Saliency Prediction using pretrained weights

4.3 CNN + LSTM

The CNN and LSTM architectures designed for the purpose of activity recognition have been described in Section 3. Here we will describe how the architecture is modified for the purpose of classification.

The CNN takes as input a 4 channel RGBS image and transforms it into a 512-d feature vector. 32 such feature vectors are extracted corresponding to 32 consecutive frames to be fed into an LSTM with 32 units. In our experiments, we have tried different number of LSTM layers but the results have been shown for 1 layersince it performed reasonably well.

5 Results

In this section, we show results and observations made for the 2 major experiments (saliency prediction and activity recognition) and also shown certain variations experimented with. For saliency prediction, we will show the results of the retrained SalGAN network on first person videos. We will also compare the results obtained using only the generator model as well as the adversarially trained model. For activity prediction, we will show quantitative results of the performance of the network.

5.1 Saliency Prediction

We trained the SalGAN network using generator loss for 30 epochs having started from the pretrained SalGAN weights followed by 60 epochs of both generator and discriminator training. We observed good trends in the loss function as shown in Figure 5 and 6.

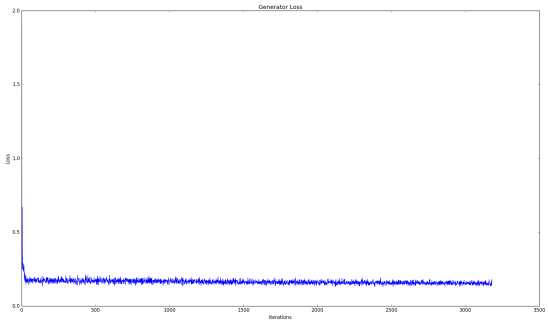


Figure 5: BCE loss during training. Note that the generator loss continues to reduce albeit marginally with discriminator training

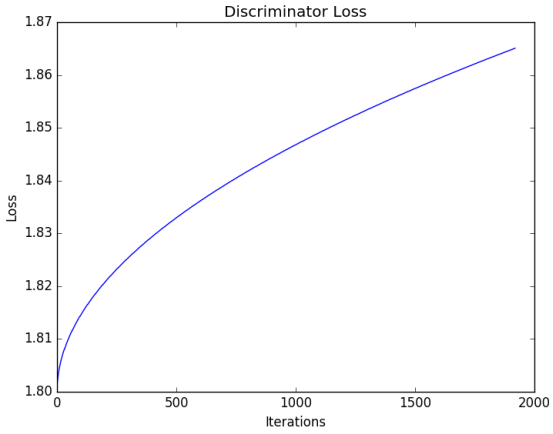


Figure 6: Variation of discriminator loss during training. Shows that the generated saliency map is able to fool the discriminator network.

After various experiments and hyperparameter optimization, we obtained reasonable good results which can be articulated in the form of the following observations:

1. Unbiased by gaze points - The network was able to predict salient regions in images where the gaze point was either away from important regions or even outside the image frame. This is illustrated in Figure 7.

2. Adversarial training improves saliency prediction - The network was able to expand around salient regions when trained with the help of a discriminator network whereas it was limited to a small region around gaze point when trained with only the cross entropy loss. This is illustrated in Figure 8.

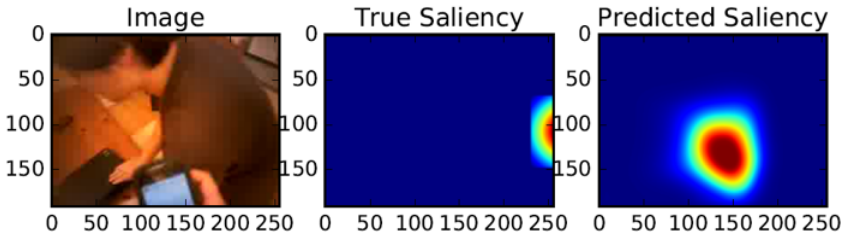
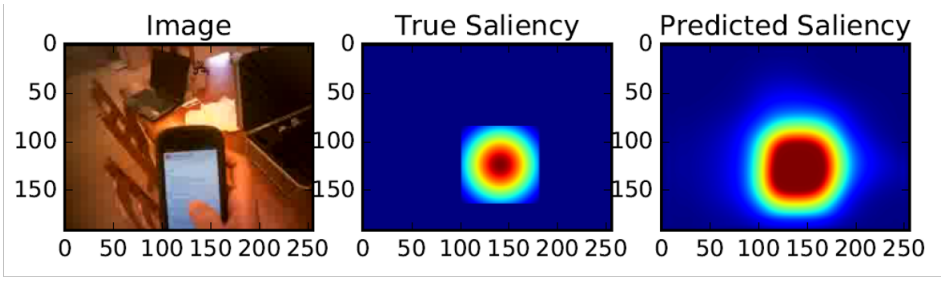


Figure 7: Saliency Prediction unbiased towards gaze points



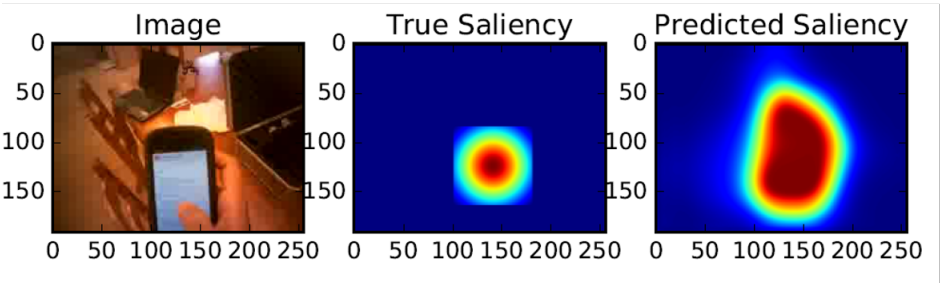


Figure 8: Saliency Prediction using generator alone shown above. Prediction using GAN shown below. This shows that the adversarial architecture drives the prediction towards salient regions rather than gaze regions.

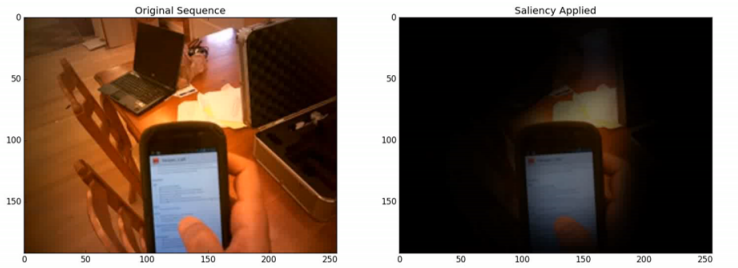


Figure 9: In-training saliency prediction overlapped over the RGB image indicating that the saliency map isolates regions useful for activity prediction

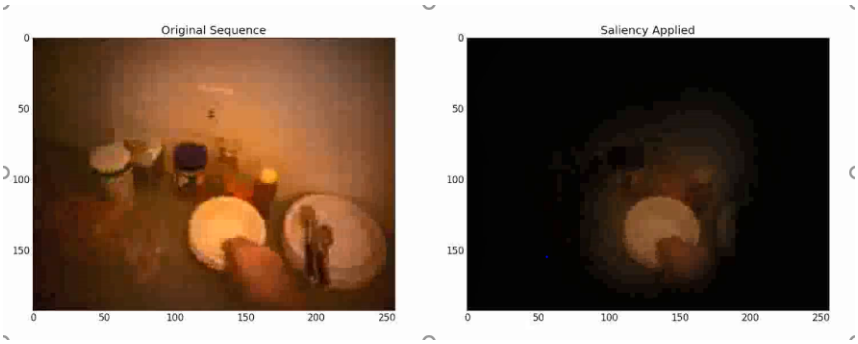


Figure 10: Test image saliency prediction overlapped over the RGB image indicating that the saliency map isolates regions useful for activity prediction

5.2 Regions to improve

1. Could be improved to handle saccade motions by learning temporal saliency relations.

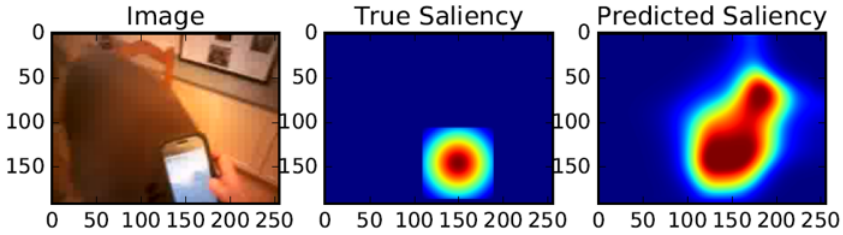


Figure 11: Saliency Prediction not so accurate when gaze is not fixated. Improvement in handling saccade gaze type would aid in activity recognition

5.3 Activity Prediction

In this experiment, we fed 32 consecutive RGBS images into the CNN + LSTM network consisting of 32 LSTM units with only 1 layer of LSTM and predicted the action amongst the 17 possible actions and object in context amongst 34 possible objects for each frame. The experiments were performed on both ground truth saliency maps and the SalGAN predicted saliency maps but the differences in the result were not so pronounced.

The training data was randomly sampled batches of 32 consecutive frames which were captured at 8 frames per second. We obtained a prediction accuracy of 35.9% for action recognition and 38.9% for object recognition. We experimented with the number of LSTM layers but the performance did not improve significantly compared to the cost of adding another layer.

From all our experiments, we inferred the following:

1. The sequence length we fed in was not sufficient for the LSTM to infer much much about the activity being performed. The predicted actions and object labels stagnates after a while indicating that the memory component of the LSTM is not significant.

2. Unclear on how to represent information that the LSTM can effectively utilize. Our guess is that VGG16 feature extraction does not provide good spatial context for activity recognition from saliency maps. We intend to use better architectures like ResNet or spatial meshgrid guided maps for feature extraction.

6 Future Directions

The performance of the activity prediction could be improved using some or all of the following methods:

1. Including spatial information in the form of meshgrid into CNN feature extractor

2. Using ResNet architecture for better feature extraction

3. Overlap saliency map onto RGB image in order to use regions of interest which are most relevant to the activity

4. Feed in larger batches of video sequences to aid LSTM

5. Using 3D convolutions instead of LSTM

There are other experiments that could be performed to improve the saliency prediction. These are:

1. Context based saliency maps (using a separate parallel network for global features)

2. Hourglass based architecture which uses low level context information in high level feature maps to possibly get better saliency maps

Experiments on other first person datasets need to be tried to validate the performance of the SalGAN. More intricate analysis of activity recognition could be performed using lesser number of actions that are predicted or by merging together multiple actions or objects (for example, merging together bacon and bacon container, fridge and freezer, etc.). Other application areas like abnormality detection or sports activity recognition for which the datasets are readily available, could be experimented with to generalize the importance of this model and the usefullness of the idea of using saliency maps for activity recognition.

7 References

[**1**] Junting Pan, Cristian Canton, Kevin McGuinness, Noel E. O'Connor, Jordi Torres, Elisa Sayrol and Xavier Giro-i-Nieto. "SalGAN: Visual Saliency Prediction with Generative Adversarial Networks." arXiv. 2017.
[**2**] K. Krafka, A. Khosla, P. Kellnhofer, H. Kannan, S. Bhandarkar,W. Matusik, and A. Torralba. Eye tracking for everyone in IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016.
[**3**] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. WardeFarley,S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets in Advances in Neural Information Processing Systems, pages 2672–2680, 2014.
[**4**] Donahue, J., Hendricks, L. A., Guadarrama, S., Rohrbach, M., Venugopalan, S., Saenko, K., Darrell, T. (2014). Long-term recurrent convolutional networks for visual recognition and description. arXiv preprint arXiv:1411.4389.Chicago