Attention-based Facial Behavior Analytics in Social Communication

Lezi Wang[1]
lw462@cs.rutgers.edu
Chongyang Bai[2]
cy@cs.dartmouth.edu
Maksim Bolonkin[2]
mbolonkin@cs.dartmouth.edu
Judee Burgoon[3]
judee@email.arizona.edu
Norah Dunbar[4]
ndunbar@comm.ucsb.edu
V. S. Subrahmanian[2]
vs@dartmouth.edu
Dimitris N. Metaxas[1]
dnm@cs.rutgers.edu

[1] Department of Computer Science,
Rutgers University,
NJ, USA

[2] Department of Computer Science,
Dartmouth College,
NH, USA

[3] Center for the Management of
Information,
University of Arizona,
AZ, USA

[4] Department of Communication,
University of California, Santa Barbara,
CA, USA

Abstract

In this study, we address a cross-domain problem of applying computer vision approaches to reason about human facial behaviour when people play *The Resistance* game. To capture the facial behaviours, we first collect several hours of video where the participants playing *The Resistance* game assume the roles of deceivers (spies) vs truth-tellers (villagers). We develop a novel attention-based neural network (NN) that advances the state of the art in understanding how a NN predicts the players' roles. This is accomplished by discovering through learning those pixels and related frames which are discriminative and contributed the most to the NN's inference. We demonstrate the effectiveness of our attention-based approach in discovering the frames and facial Action Units (AUs) that contributed to the NN's class decision. Our results are consistent with the current communication theory on deception.

1 Introduction

Research shows that when humans communicate, more social meaning comes from nonverbal than verbal cues, and among the nonverbal modalities, the face is the one upon which people typically rely [1]. Facial expressions convey one's identity, display emotions, show status, give context, open or shut down conversation, signal approval, and reveal strength of conviction, among other things. People rely on facial cues to glean both intentional and unintentional meaning. With so much communicated by the face, it is natural that facial expressions have been investigated for possible cues to deception for decades [4, 5, 8]. With advances in computer vision has come the possibility of detecting facial movement variations on a more granular scale than the human eye can perceive, and with it, the discovery of deception indicators not normally directly detectable by human perception [6, 20]. Although much deception research has focused on the emotional potential of the face, searching for micro-level "leaked" indicators that betray concealed true emotions [10], the face can reveal far more signals related to deception. It may reveal signs of cognitive effort and efforts to retrieve information from memory as a speaker attempts to formulate a believable verbal statement [3]. For example, blink patterns and lip presses may be associated with a speaker's thought processes. The face may signal not just internal emotional states such as fear or distress, but also affect directed toward another such as contempt or dislike. Nose flares and inauthentic smiles may signify these states. Communicators may also signal their attentiveness to others or their desire for a speaking turn [2]. Because people are aware that others' gaze is directed to the face, deceivers try to control their face and may, in the process, inadvertently overcontrol it, producing a pattern of rigidity [14, 21]. Head movement, facial animation and gaze patterns may all reflect this "freezing" of activity. However, if deceivers have opportunities to rehearse, plan or mentally edit what they say, any temporary missteps may be repaired [11]. Given the fluidity of facial expressions, temporal patterns can also be telling. For instance, blink patterns vary during versus after lying, and the onset and offset of smiles may differ by truth tellers versus liars.

In previous research, to automatically decipher the meaning in nonverbal human communication using computer vision methods, researchers first applied models inspired by communication theory. However, the underlying human-defined features for the computer vision based analysis where incomplete due to non-linearity and the multi-scale nature of the problem. The recent use of neural nets, has addressed the discovery of the features associated with the computer vision-based analysis of nonverbal communication and has improved significantly the recognition of desired events during nonverbal communication such as truth telling.

In this paper, we develop a novel attention-based neural network (NN) approach in order to advance the state of the art in understanding inference in deep neural nets. Our novel approach discovers the frames in a video sequence and their content through AUs that contributed the most in the final inference of the neural net. This is done by employing a novel learning approach, at the various layers of the NN, to discover those pixels and related frames which are discriminative for the NN's class inference.

We train and test our novel approach on facial video collected from a version of the board game *The Resistance*. In this game players were randomly and secretly assigned to play deceivers (called "Spies"), or truth-tellers (called "Villagers"). The video-based facial data of the players were collected in various countries. Using our method the goal was to recognize who is a spy and who is a villager and also discover which frames and which facial expressions (AUs) contributed to the NN's class decision. Our approach demonstrates that with over 280 videos (2hr length each), we are on par with human recognition of spies vs villagers. In addition, for the first time our NN can attend and discover the frames and associated facial action units (AUs) that contributed to the NN's class decision.

2 Related Work

Visualizing CNNs. A number of previous works have been proposed to visualize the internal representations offline in an attempt to better understand the model. In [16] [17] and [24], they compute the gradient of the prediction w.r.t the specific CNN unit, i.e. the input image, to highlight the important pixels. Specifically, Simonyan et al. [16] visualize partial derivatives of predicted class scores w.r.t. pixel intensities, while Guided Backpropagation [17] and Deconvolution [24] make modifications to 'raw' gradients that result in the better visualization. Despite producing fine-grained visualizations, these methods are not class-discriminative.

Erhan et al. [12] synthesize the images to maximally activate a network unit and Mahendran et al. [13], Dosovitskiy et al. [9] analyze the visual coding so as to invert latent representation. Although these can be high-resolution and class-discriminative, they visualize a model overall and not predictions for specific input images.

Our work is mainly inspired by recent works [7, 15, 25] addressing the class-discriminative attention maps. CAM [25] generates the class activation maps highlighting the task relevant region by replacing fully-connected layers with convolution and global average pooling. A drawback of CAM is the low flexibility, which requires retraining of the classifiers and feature maps to directly precede softmax layers. Hence it is unable to be applicable to any feature layers. Grad-CAM [15] is proposed to address this issue. Without retraining and changing network architecture, Grad-CAM generates the class activation maps by weighted combination of the feature maps in different channels. The weights are computed by the averaging of the gradient of the final prediction w.r.t the pixels in feature map. According to our observation, simple averaging is unable to measure the channel importance properly, which causes a large attention inconsistency among different feature layers. Grad-CAM++ [7] proposed a better class activation map by modifying the weight computation while its high computation cost of calculating the second and third derivatives makes it hard to be used to train the model.

Video Highlight Detection is highly related to our research topic since we intend to extract a brief synopsis containing segments of special interest from a video [23]. Many earlier approaches have primarily been focused on highlighting sports videos. A latent SVM model is employed to detect highlights by learning from pairs of raw and edited videos [18]. Success of deep learning also imparted improved performance in highlight detection [22]. However, most of these techniques may not generalize well to web videos since they are either based on heuristic rules or require huge amount of human-crafted labelling data which are difficult to collect in many cases. In our *The Resistance* games, we only have video-level annotations of players' roles (Spy/Villager) without knowledge about when and where, in the untrimmed videos, players show the notable facial movements for the roles. Finding those movements are important for understanding human behaviours during communication. To achieve this goal, we incorporate the interpretation in the learning to discover those pixels and related frames which are discriminative and contributed the most to the NN's prediction for the players roles.

3 Methodology

In this section, we describe the detail of how to extract the class-discriminative attention map for the videos. The procedure is illustrated by Figure 1. Motivated by the work of Grad-CAM [15] and Grad-CAM++ [7], we use the gradient to measure the importance of each feature map pixel to having the model classify the input image as class c. For the gradient of the class score Y^c is computed by taking derivative w.r.t feature map F^k in k-th channels, i.e. $(\partial Y^c)/\partial F^k)$. The pixel importance is denoted as $(\partial Y^c)/(\partial F^k_{ij})$. In [7, 15], the gradients

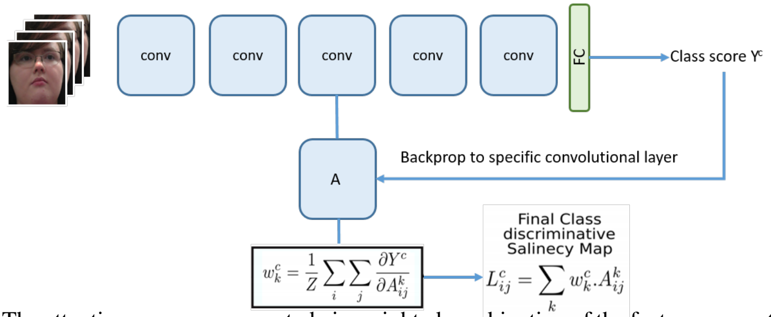
conv
conv
conv
conv
conv
FC
Class score Y^c
A
Backprop to specific convolutional layer
$w_k^c = \frac{1}{Z}\sum_i \sum_j \frac{\partial Y^c}{\partial A_{ij}^k}$
Final Class discriminative Salinecy Map
$L_{ij}^c = \sum_k w_k^c . A_{ij}^k$

Figure 1: The attention maps are generated via weighted combination of the feature maps at the specific layers. The weights measure the importance of the features, computed according to the gradients, where we take derivative of the class score w.r.t the feature maps.

are used to compute the channel-wise weights for combining the feature maps from different channels, generating the attention map of the last feature layer $\mathcal{A}_{Grad-CAM}$:

$$\mathcal{A}_{Grad-CAM} = ReLu(\sum_k \alpha_k^c F^k), \tag{1}$$

where α_k^c indicates the importance of the feature map F^k in the k-th channel. In [15], the weight α_k^c is a global average of pixel importance in the feature map:

$$\alpha_k^c = \frac{1}{Z}\sum_i \sum_j \frac{\partial Y^c}{\partial F_{ij}^k} \tag{2}$$

, where Z indicates the total number of pixels in feature map F^k In [7], higher order derivatives (second and third) involved to compute the channel weights increase the computational costs.

Besides only generating the attention map of the last feature layer as in [7, 15], we compute the category-oriented attention map for the intermediate layers. In terms of the interpretability, we propose two attention mechanisms for any feature layer with low computational cost, modeling the channel and pixel wise attention respectively. Then, we combine the model's channel and pixel wise attention to generate the final response map for the input video.

Channel-wise attention Different from Eq. 1 that the Grad-CAM uses the gradients of all the pixel to compute the channel weight, we only select the positive gradients and average them to obtain the channel-wise importance:

$$\alpha_k^c = \frac{1}{Z}\sum_i \sum_j ReLu(\frac{\partial Y^c}{\partial F_{ij}^k}) \tag{3}$$

The intuition is that the positive gradients model the pixels where the intensity increasing has positive impact on the final prediction score [7]. Substitute Eq. 3 to Eq. 1, we have the channel wise class-discriminative attention $\mathcal{A}_{ch}$.

$$\mathcal{A}_{ch} = \frac{1}{Z} ReLU(\sum_k (\sum_i \sum_j ReLU(\frac{\partial Y^c}{\partial F_{ij}^k}))F^k) \tag{4}$$

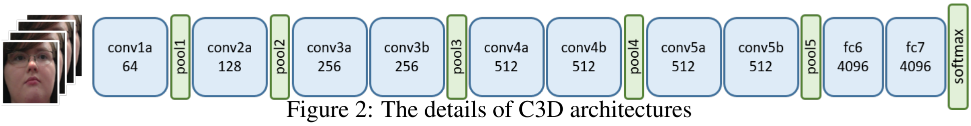

Figure 2: The details of C3D architectures

Pixel-wise attention Attention proposed by the previous works [7, 15, 25] are computed in channel-wise way, where the pixels within the same channel share the same weight for feature maps combination. Beside channel-wise attention, we also find that the pixel-wise attention demonstrates better guidance when training a model in low-quality images. Specifically, each channel acts as an expert to vote the pixel importance in the attention map. In the feature map F^k, the pixel intensity is scaled by its importance measured as $(\partial Y^c)/(\partial F^k_{ij})$ and the averaging is performed across channels to obtain the pixel-wise attention:

$$\mathcal{A}_{px} = ReLu(\frac{1}{K}\sum_k < \frac{\partial Y^c}{\partial F^k}, F^k >), \tag{5}$$

where the $< \frac{\partial Y^c}{\partial F^k}, F^k >)$ indicates the element-wise multiplication between the gradient and feature maps.

The harmonic attention According to our observation, the pixel wise attention captures more high-frequency items and the channel-wise attention maps are more smoother. Those two types of attention are complimentary where the $\mathcal{A}_{px}$ highlight the important pixels which are ignored by $\mathcal{A}_{ch}$ due the low value averaged channel weights. Hence, we propose to combine the $\mathcal{A}_{px}$ and $\mathcal{A}_{ch}$, generating the harmonic attention $\mathcal{A}$. Empirically, applying the pixel-wise weighting first and then computing the channel-wise attention as Eq.3 achieves better performance. The proposed harmonic attention is formulated as:

$$\mathcal{A} = \frac{1}{Z} ReLU(\sum_k \sum_i \sum_j ReLU(\frac{\partial Y^c}{\partial F^k_{ij}}) < \frac{\partial Y^c}{\partial F^k}, F^k >) \tag{6}$$

3.1 Training a 3D convolutional Neural Network for Spy Detection

We formulate the spy detection as a binary classification problem. Given a video sequence, we apply a 3D convolutional Neural Network (C3D) [19] to classify the player as spy or villager. Specifically, we crop the players' faces and the C3D takes a facial video clip as input, predicting the probability for the his/her role.The cropped face frames are normalized into the size of 112×112. In the C3D architecture, we design the model having 8 convolutions, 5 max-pooling, and 2 fully connected layers followed by a softmax layer. The 3D convolution kernels are $3 \times 3 \times 3$ with stride 1 in both spatial and temporal dimensions. The Number of filters are denoted in each box, as shown in Fig. 2. The 3D pooling layers are denoted from pool1 to pool5. All pooling kernels are $2 \times 2 \times 2$, except for pool1 is $1 \times 2 \times 2$. Each fully connected layer has 4096 output units. The output has 2 dimensions for binary classification. The model training and testing are conducted by using PyTorch and NVIDIA K80 XGPUs.

4 Experiments

4.1 Dataset

We tested our method on data from a real world game, where the goal is to examine deceivers' strategies and truth-tellers' deception detection abilities. Groups of participants were sought to play a board game adapted from Resistance, during which players in the roles of Villagers (truth-tellers) and Spies (deceivers) competed to win missions. To detect cultural differences in communication strategies and patterns, the games were played in eight different locales across the world. In the following we present details on dataset collection in terms of different locations, participants, procedure, game play and measurements.

Participants

There are 693 participants recruited via email, message boards and advertisements from public universities in the Southwestern US (9 games; n = 59), Western US (11 games; n = 67), Northeastern US (10 games; n = 74), Israel (10 games; n = 71), Singapore (12 games; n = 84), Fiji (14 games, n = 106), Hong Kong (15 games, n = 115), and Zambia (15 games, n = 117). The sample was 59% female, and was ethnically diverse, with the biggest groups being Asian (38%) or white, non Hispanic (18%). Nationalities represented 41 different countries. Participants were required to be proficient English speakers. Each game was approximately 2 hours long.

Procedure

Participants enrolled using an online scheduling system. Groups ranged from five to eight participants. Prior to arrival at the site, participants completed consent forms, cultural measures, and demographic questions. Upon arrival, participants were randomly assigned to one of eight computer equipped with a desk, a computer tablet with a built-in webcam, and a chair. Participants were informed that they would be filmed by the cameras.

Each group had a facilitator who explained the rules of the game. Interaction began with an ice-breaker activity, after which players rated each other on scales meant to capture baseline perceptions of dominance, composure, and trustworthiness. Participants took part in the game for an hour, during which they played between three and eight rounds. After the second, fourth, and sixth rounds, and at the end of the game, participants again completed ratings of one another and identified who they thought were the spies. Participants were paid for participating and received additional financial incentives for performing well in the game.

Game play

Similar to [26], we adapted a version of the Mafia game that closely resembles the board game The Resistance. We pilot tested several versions of the game to ensure the game best met the needs of the research questions. Players were randomly and secretly assigned to play deceivers (called "Spies"), or truth-tellers (called "Villagers"). In games of five or six players, two were assigned to be Spies, and in games of seven or eight players, three were assigned to be Spies. The goal of Villagers was to remove Spies from their community; the goal of Spies was to undermine the missions of the Villagers. Spies were aware of who the other Spies were, but Villagers did not. Villagers had to depend on shared information to deduce the other players' identities within the game.

Players completed a series of "missions" by forming teams of varying size. At the beginning of each round, players elected a leader, who then chose other players for these missions based on who they thought would help them win the game. All players voted to approve or reject the team leader, then voted on the leader's proposed team. Players voted secretly

#Validation/Training Games	Classification Accuracy
1/9	65.43(± 0.27)
2/8	62.28(± 0.30)

Table 1: The Spy/Villager prediction accuracy reported on the two different dataset splitting. The training data is randomly sampled without attention knowledge.

#Validation/Training Games	Classification Accuracy
1/9	67.85(± 0.25)
2/8	67.03(± 0.28)

Table 2: The Spy/Villager prediction accuracy reported on the two different dataset splitting. The model is trained with the video frames selected according to the model attention.

on their computer and publicly by a show of hands. Facilitators would announce if there was a discrepancy in public and private votes, thus informing participants when deception occurred. Players chosen by the leader to go on a mission team secretly voted for the mission to succeed or fail. Villagers won rounds by figuring out who the spies were and excluding them from the mission teams. Spies won rounds by causing mission failures. The ultimate winner of the game (Spies or Villagers) was determined by which team won the most rounds. Additionally, players won monetary rewards by being voted as leader or winning the game.

Measures

We design several measurements for monitoring the game play, including Game Outcome, Trust, Dominance and Previous Game Experience.

Game Outcome: In [26] Mafia study, they regard the deception detection success as the truth-tellers winning the game (i.e., if the truth-tellers win, they must have accurately detected deception). Similarly, in this study, game outcome was a dichotomous variable measuring whether or not Spies or Villagers won the game.

Trust: The extent to which participants trusted each of the other players was measured using a single-item repeated measure, which was asked after the ice-breaker, and then every even-numbered round during the game. The item read: Please rate how much you trust each player. Are they trustworthy or suspicious? A rating of 5 would mean they seem honest, reliable and truthful and 1 would mean you thought they were dishonest, unreliable and deceitful (1 Not at all to 5 Very much; Mean = 3.29, SD = 1.36). Because participants responded to this item three to five times about each of the other players, we chose to use a single item in order to avoid fatigue.

Dominance: The extent to which participants found other players to be dominant was measured using a single-item repeated measure (after the icebreaker and each of the even numbered rounds). Participants read the following text: Please rate how dominant each player is. Are they active and forceful or passive and quiet? A rating of 5 would mean you thought they were assertive, active, talkative, and persuasive. A score of 1 would mean you thought they were unassertive, passive, quiet and not influential. We got the statistical of Dominance as Mean = 3.28 and SD = .87.

Previous Game Experience: Participants' previous experience playing similar games was evaluated after the completion of the post-game measures. Participants indicated that they had or had not played a similar game. In this study, 54.2% said they had not played a similar game before.

4.2 Results of Spy Detection

In the experiments, 280 players' videos are collected for Spy/Villager prediction, including 110 spies and 170 villagers, where the players with different culture background are mixed. We segment the video clip of the first game round for training and testing. The total length is 84000 seconds (1400min). We randomly select the 10%/20% videos as testing data and the rest as the training, where there is no duplicate players appearing in both training and validation set. For each setting, the experiments are conducted 5-times cross validation and the results are reported in Table 1 and 2. Those two tables shows the Spy/Villager predication accuracy with two different frame sampling mechanisms, random sampling and attention guided frame sampling.

Random Sampling During training, given a video file, we random sample 16 frames as the input of C3D and each frame is re-sized to 112x112. The temporal order is kept among the selected frames. The prediction accuracy is shown in Table 1, where the C3D model performs better than the random guess of 60% (170/280), in the margin of $\sim 3\%$ and $\sim 5\%$.

Attention-guided Sampling The crucial difference between our Spy/Village prediction and the use of conventional image or video classification is that even with the label (spy or villager), a human has a hard time to explain the reason why the players are classified as 'Spies'. In most of these cases, spies and villagers have very similar behaviours, which means the data is not discriminative. As the high accuracy of spy prediction is one of our goal, finding where and when the players show the visual cues for 'being a spy' is the goal of our work. As in Table 1, the trained deep neural net model demonstrates better performance than random guess, which motivates us to interpret the model so to understand what visual patterns make the model predict the players as 'Spies'. Given the trained C3D model, we apply the proposed harmonic attention mechanism to compute the frame importance via averaging the attention maps. Instead of random sampling the frames in equal probabilities, we sample the frames according to the importance, leading to higher chances to sample the frames with more contribution to Spy/Villager prediction.

We keep all the parameters the same for training models with the two different frame sampling approaches. Table 1 shows our classification results when we randomly selected the video clips from the data. Table 2 shows the classification results when we retrain the model based on our attention discovered video frames. The results show clearly that the model trained with attention guided frame sampling outperforms the one with random sampling in the notable margins, $\sim 2\%$ and $\sim 4\%$ in testing/training splitting of 1:9 and 2:8 respectively. The performance boosting validates the effectiveness of our attention mechanism to identify the potential frames where spies and villagers show notable discriminative visual signals so that it is easier for training a model with better accuracy. Besides the quantitative results, we also apply the attention map to identify important pixels and visualize them in the next subsection.

4.3 Attention and Deception Cues

In Figure 3 we show promising qualitative results on the fact that our attention NN is capable of discovering cues related to what is known from communication theory for deception. In Figure 3 we show some Facial Action Units related to spies extracted from the discovered frames and the respective probabilities, i.e., AUs:13,20,24,45. The players showing such AUs are more likely to be classified as Spies. According to the communication theory, AUs 20 and 45 are related to deception, which is consistent to our expectation that spies

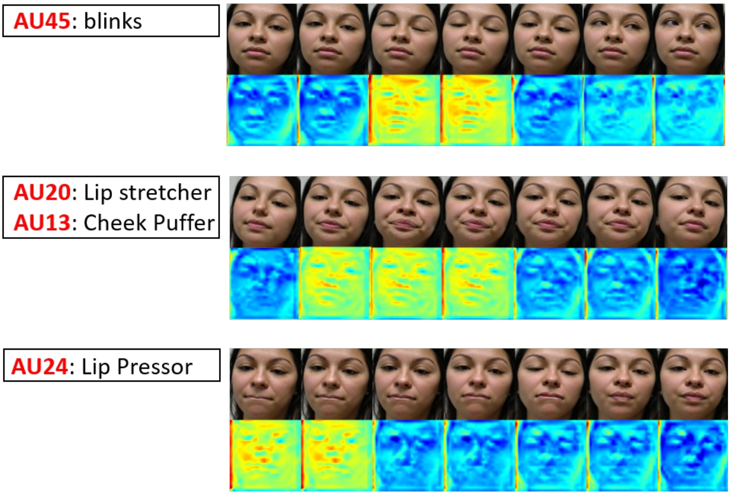

Figure 3: We compare what the model attends to known Action Units which are useful in communication research regarding face and head. The attended facial cues are coded as facial action units (AU). All examples are from spies.

are more willing to lie, but not always. In Figure 4 we show the ability of our network to attend to different cues for a spy and a villager which are also consistent with the current communication theory [6, 8] on deception. Figure 4 clearly shows cues and respective pixel probabilities, which are related to deception such as eyes closed, fake smiles, changes in lips. In particular, we show the comparison of model attention between spies and villagers. For example, our approach can attend to small facial movements related to deception like eye blinking in the bottom left case (spies). At the top row (spies and villagers), the model detects the fake and real smile so as to classify the two type of players role, correctly. These initial encouraging results show that we can extract cues and AUs related to communication and deception theory without using a prior known cues. They provide cues which are human interpretable and can be used in many other types of applications.

5 Conclusion

In this paper, we presented a novel attention-based neural network (NN) that discovers through learning in a video sequence the most discriminative frames and related pixel probabilities and AUs that contributed the most to the final class inference of the neural net. We applied our method to facial videos of a variant of the Resistance game collected in various countries where the players assume the roles of deceivers (spies) vs truth-tellers (villagers). We demonstrated for the first time that it is possible to discover the frames and AUs that contributed the most to the NN's class decision on several hours of video testing. The results are consistent with the current communication theory on nonverbal communication and can be used in future studies to discover static and dynamic relationships among cues and AUs currently not known.

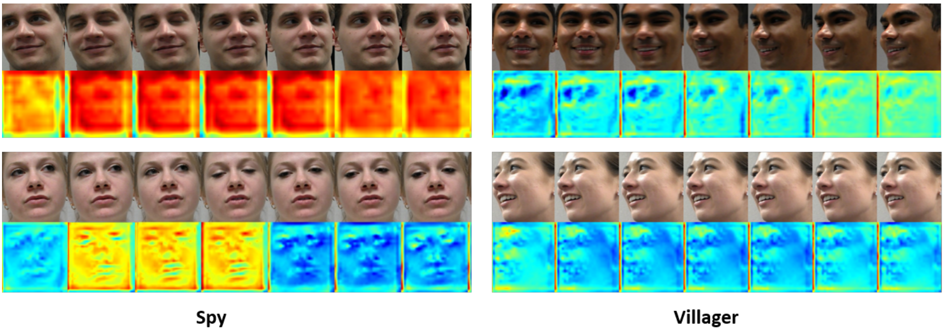

Figure 4: What the model attends to for "Spy" vs "Villager". We show the comparison of attention maps between spy and villager. The model can attend to small facial movements related to deception like eye blinking in the bottom left row (spies). And at the top row (spies and villagers), the model detects fake and real smiles so as to classify the two type of players role, correctly.

6 Acknowledgments

This work was funded partly by ARO-MURI-68985NSMUR and NSF 1763523, 1747778, 1733843, 1703883 grants to Dimitris N. Metaxas.

References

[1] Guerrero L Floyd K Burgoon, J. K. Nonverbal communication. In *Allyn Bacon.*, 2010.

[2] J. K. Burgoon. Nonverbal measurement of deceit. In *In V. Manusov (Ed.), The sourcebook of nonverbal measures: Going beyond words. Hillsdale, NJ: Erlbaum.*, pages 237–250, 2005.

[3] J. K. Burgoon. When is deceptive message production more effortful than truth-telling? a baker's dozen of moderators. In *Frontiers in Psychology*, page 6, 2015.

[4] Jensen M. L. Kruse J. Meservy T. O. Nunamaker J. F. Jr. (2007). Burgoon, J. K. Deception and intention detection. In *Handbooks in Information Systems: National Security*, pages Vol 2,193–214, 2007.

[5] Metaxas D. Bourlai T. Elkins A Burgoon, J. K. Social signals of deception and dishonesty. In *Social signal processing.Cambridge, UK: Cambridge University Press*, pages 404–428, 2017.

[6] Proudfoot J. G. Wilson D. Schuetzler R. Burgoon, J. K. Patterns of nonverbal behavior associated with truth and deception: Illustrations from three experiments. In *Journal of Nonverbal Behavior*, pages 38, 325–354., 2014.

[7] Aditya Chattopadhay, Anirban Sarkar, Prantik Howlader, and Vineeth N Balasubramanian. Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 839–847. IEEE, 2018.

[8] Lindsay J.J. Malone B.E. Muhlenbruck L. Charlton K. DePaulo, B.M. and H. Cooper. Cues to deception. In *Cues to deception. Psychological Bulletin, 129, 1*, pages 74–118, 2003.

[9] Alexey Dosovitskiy and Thomas Brox. Inverting visual representations with convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4829–4837, 2016.

[10] Friesen W. V. Ekman, P. Unmasking the face. a guide to recognizing emotions from facial clues. In *Englewood Cliffs, NJ: Prentice Hall.*, pages 71, 2, 197–214., 1975.

[11] Zafeiriou S. Pantic M. Burgoon J. K. Elkins, A. Unobtrusive deception detection. In *In R. Calvo, S. K. D'Mello, J. Gratch, A. Kappas (Eds.), The Oxford handbook of affective computing. UK: Oxford University Press.*, 2015.

[12] Dumitru Erhan, Yoshua Bengio, Aaron Courville, and Pascal Vincent. Visualizing higher-layer features of a deep network. *University of Montreal*, 1341(3):1, 2009.

[13] Aravindh Mahendran and Andrea Vedaldi. Understanding deep image representations by inverting them. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5188–5196, 2015.

[14] Twyman N.W. Burgoon J. K. Nunamaker J. F. Diller C. B. R. Pentland, S.J. A video-based screening system for automated risk assessment using nuanced facial features. In *Journal of Management Information Systems*, pages 970–993, 2017.

[15] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, Dhruv Batra, et al. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *ICCV*, pages 618–626, 2017.

[16] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*, 2013.

[17] Jost Tobias Springenberg, Alexey Dosovitskiy, Thomas Brox, and Martin Riedmiller. Striving for simplicity: The all convolutional net. *arXiv preprint arXiv:1412.6806*, 2014.

[18] Min Sun, Ali Farhadi, and Steve Seitz. Ranking domain-specific highlights by analyzing edited videos. In *European conference on computer vision*, pages 787–802. Springer, 2014.

[19] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 4489–4497, 2015.

[20] Dowdall J. Shastri D. Pavlidis I.T. Frank M.G. Tsiamyrtzis, P. and P. Ekman. Imaging facial physiology for the detection of deceit. In *International Journal of Computer Vision*, pages 71, 2, 197–214., 2007.

[21] Elkins A. Burgoon J. K. Nunamaker J. F. Jr. Twyman, N. W. A rigidity detection system for automated credibility assessment. In *Journal of Management Information Systems*, pages 31, 173–201, 2014.

[22] Huan Yang, Baoyuan Wang, Stephen Lin, David Wipf, Minyi Guo, and Baining Guo. Unsupervised extraction of video highlights via robust recurrent auto-encoders. In *Proceedings of the IEEE international conference on computer vision*, pages 4633–4641, 2015.

[23] Ting Yao, Tao Mei, and Yong Rui. Highlight detection with pairwise deep ranking for first-person video summarization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 982–990, 2016.

[24] Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *European conference on computer vision*, pages 818–833. Springer, 2014.

[25] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2921–2929, 2016.

[26] Zhang D. Sung Y. Zhou, L. The effects of group factors on deception detection performance. In *Small Group Research*, pages 44, 272–297, 2013.