CROSS-MODALITY DISTILLATION: A CASE FOR CONDITIONAL GENERATIVE ADVERSARIAL NETWORKS

Siddharth Roheda[⋆] *Benjamin S. Riggan*[†] *Hamid Krim*[⋆] *Liyi Dai*[‡]

[⋆]Department of Electrical and Computer Engineering, North Carolina State University, Raleigh, NC, USA
[†]U.S. Army Research Laboratory, Adelphi, MD, USA
[‡]U.S. Army Research Office, 800 Park Offices Dr., Durham, NC, USA

ABSTRACT

In this paper, we propose to use a Conditional Generative Adversarial Network (CGAN) for distilling (i.e. transferring) knowledge from sensor data and enhancing low-resolution target detection. In unconstrained surveillance settings, sensor measurements are often noisy, degraded, corrupted, and even missing/absent, thereby presenting a significant problem for multi-modal fusion. We therefore specifically tackle the problem of a missing modality in our attempt to propose an algorithm based on CGANs to generate representative information from the missing modalities when given some other available modalities. Despite modality gaps, we show that one can distill knowledge from one set of modalities to another. Moreover, we demonstrate that it achieves better performance than traditional approaches and recent teacher-student models.

Index Terms— Missing Modalities, Generative Adversarial Networks, Target Detection, Multi-Modal Fusion

1. INTRODUCTION

Some sensor measurements are often noisy, missing, or unusable in unconstrained surveillance settings, and sometimes there are limitations on the types of sensors that are deployed. In this scenario, it is common to ignore such sensors, with a potential negative impact on performance relative to the ideal scenario (i.e. all sensors are available and functional). We consider exploiting prior knowledge about the relationship between the two sets of modalities, so that a detection system can safeguard a high detection accuracy.

There has recently been an interest in this type of knowledge transfer. Shao et al. [1] proposes to use multiple deep auto-encoders with a bagging strategy and Robust PCA to address missing modalities for image classification. Hoffman et al. [2] introduces hallucination networks that can address a missing modality at test time by distilling knowledge into the network during the training. Furthermore, [3] shows that considering interactions between modalities may lead to a better representation. Recent work in Domain Adaptation [4, 5, 6] addresses differences between source and target domains. In these works, the authors try to learn a set of intermediate domains (represented as points on the Grassman manifold in [4, 5] and by dictionaries in [6]) between the source domain and the target domain. These approaches to cross-modal inference do not explore the distillation of information between modalities with different phenomenologies (e.g., imaging and acoustic data). In this paper, we try to explore the possibility of such a transfer of knowledge by leveraging information from seismic and acoustic sensors while training a model that takes in images for classification.

Our Contributions: We propose to use Conditional Generative Adversarial Networks (CGANs) [7] to generate representative information from the missing sensor modalities (e.g. temporal data) while conditioned on the available modalities (e.g. spatial data). This, in effect, distills knowledge from the missing modalities into the model trained for dealing with available modalities.We modify the generator cost function so that the model learns to extract realistic and discriminative information corresponding to the missing modalities. Furthermore, we compare our approach to traditional approaches (i.e. ignore missing modalities) and recent distillation approaches. We also evaluate various cost functionals in order to better understand the impact of different loss functions and regularizations.

2. BACKGROUND AND RELATED WORK

2.1. Teacher-Student Distillation Model

This model has been largely used for compressing large neural networks (i.e. teacher) into smaller ones (i.e. students), while achieving a similar performance. In [8], the output from the teacher was used as the target probability distribution for the student. Hence, the student was able to achieve similar performance to the teacher. Further improvements on the technique introduced in [8] are proposed in [9]. In [10], a type of teacher-student model plus additional privileged information is used. This kind of distillation was then extended to training networks to deal with missing modalities in [2], where L^2 loss (hallucination loss) is used to train a hallucination network.

2.2. Generative Adversarial Networks

Generative Adversarial Networks (GANs) were recently introduced in [11]. GANs are composed of two parts: a generator and discriminator. The generator aims to confuse the discriminator by randomly synthesizing a realistic sample from some underlying distribution. The adversarial discriminative network aims to differentiate between actual samples and generated ones. When the generator and discriminator networks were alternatively trained, the generator was able to successfully generate a random, but realistic sample from the distribution of the training data. Later, Conditional Generative Adversarial Networks (CGANs) in [7] were trained to generate samples from a class conditional distribution. Here, the input to the generator was replaced by some useful information rather than random noise. Therefore, the objective of the generator was to generate realistic data, given the conditional information. CGANs have been used to generate random faces given attributes [12] and to produce relevant images given a text descriptions [13].

2.3. Noise Models

Past works have explored various noise models for CGANs. Without a random process (i.e. noise), the generator would produce deterministic results. In [14] a Gaussian input noise is provided to the generator. But often, the generator can learn to ignore this noise [15, 16]. Another noise model that has been used is dropout noise in [16], where dropout is applied to several layers of the generator.

3. PROBLEM FORMULATION

Consider m sensors surveilling an area of interest. We wish to detect the presence of a target, given the data collected by these sensors. Let the observations for the i^{th} sensor be denoted by, $\boldsymbol{a}_i = \{a_{ij}\}_{j=1,...,n}$, where, a_{ij} corresponds to the signal value at time j. Now, assume that observations from k sensors, $\boldsymbol{A} = \{\boldsymbol{a}_1, \boldsymbol{a}_2, ..., \boldsymbol{a}_k\}$, are missing or unusable during the testing phase, while the rest of the sensor observations, $\boldsymbol{B} = \{\boldsymbol{a}_{k+1}, \boldsymbol{a}_{k+2}, ..., \boldsymbol{a}_m\}$, are usable for target detection. Let $M : \boldsymbol{A} \rightarrow F_M$ be a mapping from the data (available for training), $\boldsymbol{A}$, to a feature space, F_M. Let $C_{\boldsymbol{A}}(F_M) = \boldsymbol{w}_{\boldsymbol{A}}^T F_M + b_{\boldsymbol{A}}$ be a scoring function gives a detection score, where, $\boldsymbol{w}_{\boldsymbol{A}}$ and $b_{\boldsymbol{A}}$ is the weight vector and the bias for a classifier trained to detect a person. Then, the probability of detection is $P_{\boldsymbol{A}} = S(C_{\boldsymbol{A}}(F_M))$, where, $S(x) = 1/(1 - e^{-x})$, is the sigmoid function. Also consider another mapping, $G : \{z, \boldsymbol{B}\} \rightarrow F_G$, that maps random noise, z, and the usable data, $\boldsymbol{B}$, to a feature space, F_G.

Our goal is to then use the available sensor data, $\boldsymbol{B}$, in order to predict features that would have been generated by $\boldsymbol{A}$, and hence improve detection performance with the usable data. That is, we would like to find the mapping, G, such that $F_G = G(z, \boldsymbol{B}) \approx F_M = M(\boldsymbol{A})$.

4. PROPOSED METHOD

Fig. 1 shows a high level block diagram for learning the mapping G. The left side of the diagram depicts how $\boldsymbol{A}$ is mapped to the feature space, F_M, by the mapping M. These features are then used for classification purposes, and a linear classifier, with scoring function, $C_{\boldsymbol{A}}(.)$, is trained using F_M.

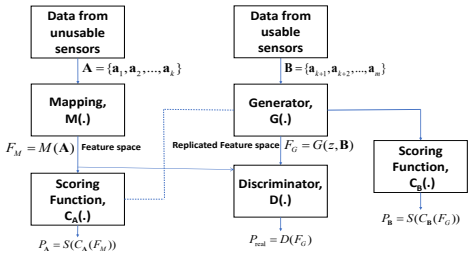

Fig. 1: A high level block diagram for training the generator

On the right side of the diagram, we have a CGAN network that we use to replicate the feature space, F_M. The generator from the CGAN network is used as the mapping, G, and takes the data from available modalities as the input. This generative model is pitted against an adversary: a discriminator function, $D(.)$, that tries to discriminate between the generated feature space, F_G, and the targeted feature space, F_M. The generator tries to produce a feature space, F_G, that replicates F_M as closely as possible and causes misclassification by the discriminator. Then using the CGAN formulation as defined in [7, 16], we have,

$$\min_{G(.)} \max_{D(.)} V(D, G), \tag{1}$$

where

$$V(D,G) = \mathbb{E}_{\boldsymbol{B} \sim p_{data}(\boldsymbol{B}), M(\boldsymbol{A}) \sim p_{data}(M(\boldsymbol{A}))}[logD(\boldsymbol{B}, M(\boldsymbol{A}))] + \mathbb{E}_{\boldsymbol{B} \sim p_{data}(\boldsymbol{B}), z \sim p_z(z)}[log(1 - D(\boldsymbol{B}, G(z, \boldsymbol{B})))], \tag{2}$$

and $\mathbb{E}(.)$ is the expectation operator with respect to the observed data. This, in effect, minimizes some distance loss function, $\mathcal{L}(F_G, F_M)$, such that, F_G and F_M are close to each other, leading to performance comparable to that achievable prior to missing $\boldsymbol{A}$. While it is reasonable to expect good classification performance from such an attempt, we found that the discriminative information is suboptimal. This results in the discriminator adapting to "erroneous real features" and lead to an undesirable outcome. This may be attributed to a convergence to the wrong stable point of the objective functional. As an illustration, when considering an image patch in which very few pixels correspond to the target (see Fig. 2), the generator may be easily confused. Therefore, we require additional criteria to guide the generator to a better convergence point.

We remedy this issue by letting the classifier (with the scoring function $C_{\boldsymbol{A}}(.)$) influence $G(.)$. This classifier is useful for guiding the generator to produce more discriminative features derived from $\boldsymbol{B}$. This causes the generator to

produce features that lie on the correct side of the decision boundary. Adding this classifier influence does not have a large impact on training time as the classifier is pre-trained. The cross-entropy loss for a classifier with a scoring function, $C(.)$, given a set of features, $X = \{\boldsymbol{x}_i\}$, with classification labels, $L = \{l_i\}$, where index i represents the i^{th} sample, is given as,

$$C_{Loss}(X, L) = \sum_i -l_i log S(C(\boldsymbol{x}_i)). \quad (3)$$

We further incorporate the minimization of some measure of distance between the generated features and the real features. For our implementation, we minimize the L^2 loss: $L^2(F_M, F_G) = \sum_i (f_{M_i} - f_{G_i})^2$, where f_{M_i} and f_{G_i} denote samples from F_M and F_G, respectively. This loss encourages the generator to produce a representation corresponding to the conditional input. Since we used the classifier with scoring function $C_{\boldsymbol{A}}(.)$ to influence the training of the generator, $G(.)$, it is reasonable to expect its good performance in classifying features generated by $G(.)$. F_G is unfortunately not an exact copy of F_M, albeit close. So, a better classifier for F_G may exist, and is shown as the scoring function $C_{\boldsymbol{B}}(.)$ in the block diagram in Fig. 1. We use dropout noise as the input noise to the generator, which is discussed in [16]. Let U be the set of class labels for the training samples, $C_{\boldsymbol{A}_{Loss}}(F_G, U)$ be the classification loss for the generated features using the pre-trained classifier with scoring function, $C_{\boldsymbol{A}}(.)$, and $L^2(F_M, F_G)$ be the L^2 loss between the targeted features and the generated features. Then, we formulate the following optimization problem to efficiently replicate F_M,

$$\min_{G(.)} \max_{D(.)} J(D, G) = V(D, G) + \alpha C_{\boldsymbol{A}_{Loss}}(F_G, U) + \beta L^2(F_M, F_G) \quad (4)$$

Where, α and β are scaling coefficients for controlling influence of the pre-trained classifier and the L^2 loss. For our implementation, we use $\alpha = 10^{-3}$, and $\beta = 10^{-5}$.

5. EXPERIMENTS AND RESULTS

5.1. Experimental Setup

For our experiments we used pre-collected data from a network of seismic sensors, acoustic sensors, and video cameras deployed in a field, where people were walking around in specified patterns. Details about this sensor setup and experiments can be found in [17]. This dataset has been previously used for target detection in [18, 19], where the authors focus on detection/classification using seismic and acoustic sensors only. Some data samples from the dataset can be seen in Fig. 2. It can be seen that while the response from seismic and acoustic sensors has significant discriminative information, the same cannot be said for the images from the video camera. So, in an ideal case, we would use the responses from the seismic and acoustic sensors to detect target presence. We assume those sensors are unavailable during the testing phase.

Fig. 2: A sample video frame with a small human target (left), and samples from seismic (top right) and acoustic (bottom right) sensors

5.2. Implementation Details

To simplify the generator conditioned on images, we feed image patches of size 64x64 instead of the entire 720x1080 image. This makes the conditioning on patches of the image, and the generator generates features that are similar to those from the network trained over the seismic and acoustic data. An image patch including a person is labeled as a target and correlated with 1 second long (4096 samples) seismic and acoustic data from the closest node to the person. On the other hand, the person free patches, are labeled as non-target and correlated with 1 second long (4096 samples) seismic and acoustic data from the farthest node to the person.

The mappings, $M(.)$ and $G(.)$ are approximated by neural networks. The network structure used is the same across all algorithms that are compared. The structures are identical for the mapping of seismic, acoustic data into the feature space and for generation of features from image patches, except for the fact that the former uses 1-d convolutional filters, while the latter uses 2-d convolutional filters. A 3 layered CNN structure is used, with 2x2 max-pooling after the first and last convolutional layers. The first layer uses a filter size of 3 and the last two layers approximate a layer of filter size 5 by using two layers of filter size 3. As discussed in [21], using two layers of filter size 3 is computationally faster, with an equivalent output to using a single layer with filter size 5. The convolutional layers are followed by a fully connected layer that transforms the output of the convolutional layers into a feature vector. The fully connected layer generates a feature vector with 50 values, before being fed into a classification layer. This is the feature vector we try to predict using the CGAN based algorithm we propose.

5.3. Results

The results for the proposed algorithm can be seen in Fig. 3, in comparison to a network that was trained over just video frames. It can be seen that the network trained over video frames only biases toward the horizon in the image, where the person is mostly found to be walking, but this issue is resolved when cross-modal distillation is performed, and a more focused detection is achieved. Fig. 4 shows results on a video frame captured from a different angle, and shows that the proposed algorithm is not over trained, and is robust to different perspective.

Further, Fig. 5 shows the output from the convolutional

Method	Avg Accuracy (using $C_A(.)$)	Avg Accuracy (using $C_B(.)$)	Avg F-score (using $C_A(.)$)	Avg F-Score (using $C_B(.)$)
Seismic, Acoustic Sensors[1]	95.93%	-	0.91	-
Video Frames	-	76.67%	-	0.53
Hallucination Networks (Teacher-Student Model)	-	82.40%	-	0.63
Cross-Modal Distillation (CGAN)	67.10 %	67.73%	0.59	0.60
Cross-Modal Distillation (CGAN + Classifier Loss)	90.93%	91.33%	0.81	0.83
Cross-Modal Distillation (CGAN+ Classifier Loss + L^2)	92.18%	**93.52%**	0.87	**0.88**
Cross-Modal Distillation (CGAN + L^2)	89.21%	90.89%	0.73	0.76
Cross-Modal Distillation (Classifier Loss + L^2)	85.97%	86.01%	0.70	0.72
Cross-Modal Distillation (Distance Correlation [20] + Classifier Loss)	87.19%	91.11%	0.77	0.81
Cross-Modal Distillation (CGAN + Distance Correlation [20] + Classifier Loss)	90.21%	90.9%	0.80	0.80

Table 1: Performance Comparison [2]

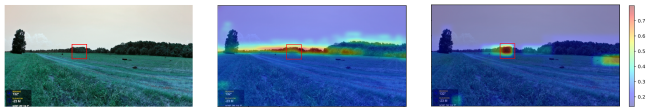

Fig. 3: Original video frame (left), Detection Probability when network was trained using just the video frames (center), Detection Probability when network was trained using proposed Cross-Modal distillation (right). Red box marks true location of target

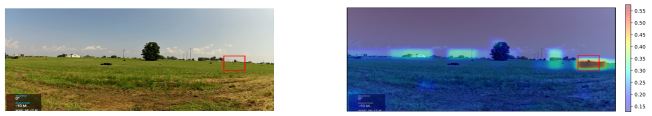

Fig. 4: Detection results on a video frame captured from a different angle shows robustness of the proposed algorithm

layer when a patch from the input video frame was fed in. It can be seen that the network trained using just the video frames is highly responsive along the edge between trees and grass. While, this is a feature, it is not of interest here. On the other hand, the network trained using cross-modal distillation is highly responsive around the head of the person in the image, which is a much more interesting feature for our purpose. Table 1 presents the detection accuracies for different combinations of cost functions. Accuracies and F-scores are reported for both cases of classifiers, i.e. the pre-trained classifier, $C_A(.)$, and the finetuned (on the generated feature space) classifier, $C_B(.)$. Detection performance achieved by the seismic and acoustic sensor data is optimal. But, when these sensors are damaged or not available, we must use the available video frames for target detection. On comparing the performance of classifier based on only video frames with other distillation techniques, it is clear that some transfer of knowledge definitely takes place, and helps us build a model with improved detection performance. Furthermore, our approach, which is based on using CGANs along with additional losses, is able to improve the knowledge transfer compared to teacher-student model based distillation. From Table 1 we can see that the combination of CGAN conditioned on video frames, a classifier influence, and an L^2 loss yields the best detection performance, which is very close to that achieved by the seismic and acoustic sensors.

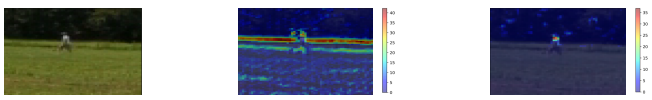

Fig. 5: Positive Patch (left), Output from convolutional layer trained using video frames only (center), Output from convolutional layer trained using proposed Cross-Modal distillation (right)

6. CONCLUSION

In this paper we proposed a technique for cross-modal distillation that uses Conditional Generative Adversarial Networks (CGANs) in order to predict features from modalities that may be unusable for testing/implementation due to damage or other potential limitations. This cross-modal distillation allows us to leverage information from these missing modalities in order to guide the model being trained to learn highly discriminative features. Experiments show that detection using modalities with low discriminative information (due to low resolution of target in our case) can be significantly improved by distilling knowledge from a modality with higher discriminative power. This algorithm improves upon the performance using recent Teacher-Student model for distillation, and hence provides robustness in presence of limited sensing modalities.

[1]This is the ideal performance, which is not achievable since we assume these sensors are missing.

[2]Cells with a "-" cannot be reported due to incompatibilities.

7. REFERENCES

[1] Ming Shao, Zhengming Ding, and Yun Fu, "Sparse low-rank fusion based deep features for missing modality face recognition," in *Automatic Face and Gesture Recognition (FG), 2015 11th IEEE International Conference and Workshops on*. IEEE, 2015, vol. 1, pp. 1–6.

[2] Judy Hoffman, Saurabh Gupta, and Trevor Darrell, "Learning with side information through modality hallucination," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 826–834.

[3] Sarah Rastegar, Mahdieh Soleymani, Hamid R Rabiee, and Seyed Mohsen Shojaee, "Mdl-cw: A multimodal deep learning framework with cross weights," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 2601–2609.

[4] Raghuraman Gopalan, Ruonan Li, and Rama Chellappa, "Domain adaptation for object recognition: An unsupervised approach," in *Computer Vision (ICCV), 2011 IEEE International Conference on*. IEEE, 2011, pp. 999–1006.

[5] Jingjing Zheng, Ming-Yu Liu, Rama Chellappa, and P Jonathon Phillips, "A grassmann manifold-based domain adaptation approach," in *Pattern Recognition (ICPR), 2012 21st International Conference on*. IEEE, 2012, pp. 2095–2099.

[6] Jie Ni, Qiang Qiu, and Rama Chellappa, "Subspace interpolation via dictionary learning for unsupervised domain adaptation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 692–699.

[7] Mehdi Mirza and Simon Osindero, "Conditional generative adversarial nets," *arXiv preprint arXiv:1411.1784*, 2014.

[8] Cristian Bucilu, Rich Caruana, and Alexandru Niculescu-Mizil, "Model compression," in *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2006, pp. 535–541.

[9] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean, "Distilling the knowledge in a neural network," *arXiv preprint arXiv:1503.02531*, 2015.

[10] Vladimir Vapnik and Rauf Izmailov, "Learning using privileged information: similarity control and knowledge transfer.," *Journal of machine learning research*, vol. 16, no. 20232049, pp. 55, 2015.

[11] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio, "Generative adversarial nets," in *Advances in neural information processing systems*, 2014, pp. 2672–2680.

[12] Jon Gauthier, "Conditional generative adversarial nets for convolutional face generation," *Class Project for Stanford CS231N: Convolutional Neural Networks for Visual Recognition, Winter semester*, vol. 2014, no. 5, pp. 2, 2014.

[13] Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee, "Generative adversarial text to image synthesis," *arXiv preprint arXiv:1605.05396*, 2016.

[14] Xiaolong Wang and Abhinav Gupta, "Generative image modeling using style and structure adversarial networks," in *European Conference on Computer Vision*. Springer, 2016, pp. 318–335.

[15] Michael Mathieu, Camille Couprie, and Yann LeCun, "Deep multi-scale video prediction beyond mean square error," *arXiv preprint arXiv:1511.05440*, 2015.

[16] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros, "Image-to-image translation with conditional adversarial networks," *arXiv preprint arXiv:1611.07004*, 2016.

[17] Sylvester M Nabritt, Thyagaraju Damarla, and Gary Chatters, "Personnel and vehicle data collection at aberdeen proving ground (apg) and its distribution for research," Tech. Rep., ARMY RESEARCH LAB ADELPHI MD SENSORS AND ELECTRON DEVICES DIRECTORATE, 2015.

[18] Kyunghun Lee, Benjamin S Riggan, and Shuvra S Bhattacharyya, "An accumulative fusion architecture for discriminating people and vehicles using acoustic and seismic signals," in *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*. IEEE, 2017, pp. 2976–2980.

[19] Kyunghun Lee, Benjamin S Riggan, and Shuvra S Bhattacharyya, "An optimized embedded target detection system using acoustic and seismic sensors," .

[20] Gábor J Székely, Maria L Rizzo, et al., "Brownian distance covariance," *The annals of applied statistics*, vol. 3, no. 4, pp. 1236–1265, 2009.

[21] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna, "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 2818–2826.