Anomaly Detection With Multiple-Hypotheses Predictions - Supplementary Materials -

Anonymous Authors[1]

Definitions and more formal discussions are provided in these sections.

A. Mixture Density Network

The Mixture Density networks predict a data conditional Gaussian mixture model (GMM)in the data space. Conditioning means that each latent vector, i.e., a point on the learned manifold is projected back to a GMM in the data space.

A GMM learns from the following energy function:

$$L_{GMM}(x) = -\log \sum_h \alpha_h \mathcal{N}(x; \mu_h, \sigma_h) \qquad (1)$$

Whereby x is the input data, μ_h and σ_h parametrize the $h-th$ Gaussian distribution in the mixture. α_h are the mixing coefficients across the individual mixtures.

Contrary, a Mixture Density network hat multiple output heads (multiple-hypotheses). The framework extends the GMM-learning by the data conditioning as follows:

$$L_{MDN}(x) = E_{z_i \sim q_\phi(z_i|x)} \left[L_{GMM}(x|z_i) \right] \qquad (2)$$

whereby q_ϕ is a inference network shared by all individual mixtures. z is the latent code. The hypotheses are coupled into forming a likelihood function by the mixing coefficients α_i.

B. Multimodal learning on the flipped moon toy dataset

Fig. 1 shows the flipped half-moon dataset to demonstrate MHP-learning in contrast to unimodal output distribution learning. In this section, Fig 1 shows a qualitative evaluation of different MHP-techniques. This task is a one-to-many mapping from x to y with a discontinuity at the point $x = 0$ and $x = 0.5$.

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

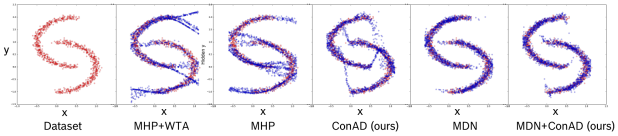

Figure 1. Flipped half-moon dataset: conditional prediction of y based on x. Red points are samples from true distribution while blue points represent samples from distributions approximations. Learning with multiple-hypotheses predictions (MHP) loss or MHP + Winner-takes-all (WTA) loss lead to support of artificial data regions. Mixture density networks and our approach ConAD reduces this effect.

When the local density function abruptly ends, MHP-techniques support artificial data regions since they are not penalized for artificial modes by the objective function as discussed before. We refer to this property as an inconsistency concerning the true underlying distribution. In contrast to that, Mixture Density Networks (MDN) and our ConADs approaches reduce the inconsistencies to the minimum.

C. One-to-many mapping tasks require multi-modality

Consider a simple toy problem with an observable x and hidden y which is to be predicted and expressed by the conditional distribution $p_{true}(y|x)$ such as in Fig. 1. Since the data conditional is multi-modal for some x, an unimodal output distribution cannot fully capture the underlying distribution. Instead, the bias-free solution for the Mean-Squared-Error-minimizer is the empirical mean $\overline{y_{x_i}}$ of $p_{train}(y|x_i)$ on the training set. However, this learned conditional density does not comply with the underlying distribution: sampled data points fall into the low-likelihood regions under $p_{true}(y|x)$. With increasing number of output hypotheses, the data modes could be gradually captured. For this task, the energy to be minimized is given by the Negative-log-likelihood of the Mixture Density Network (MDN) App. A under a Gaussian Mixture with hypotheses

h in Eq. 4 :

$$\begin{aligned} E_{MDN}(\Theta) &= -\log L(\Theta|X;Y) \\ &= -\log p_{GMM}(Y|X,\Theta) \\ &= -\sum_i \sum_h \log \alpha_h p_{\theta_h}(y_i|x_i) \end{aligned} \tag{3}$$

with

$$p_{\theta_h}(y_i|x_i,\theta_h) = \frac{1}{\sqrt{2\pi}\sigma_h} \exp -\frac{(y_i-\mu_h)^2}{2\sigma_h^2} \tag{4}$$

D. Lemma 4.1

Given a sufficient number of hypotheses H', an optimal solution Θ^* for $E_{WTA}(\Theta^*)$ is not unique (permutation is excluded). There exists a $\Theta^{'}$ with $E_{WTA}(\Theta^*) = E_{WTA}(\Theta^{'})$ which is not consistent w.r.t. the underlying output distribution $p_{train}(y_i|x_i)$.

Proof. : Suppose c is the maximal modes count of the dataset sampled from the real underlying conditional output distribution $p(y_i|x_i)$. Since $|\{(x_i,y_i)\}| < \infty \rightarrow c < \infty$.

Suppose $H = c$, then a trivial optimal solution for $E_{WTA}(\Theta_H)$ is found by centering each hypothesis μ_{ik} at a different empirical data point k $y_{ik} \sim (y_i, x_i)$ and $\sigma_{ik} \mapsto 0$. In this case $\lim_{\sigma_{ik}\mapsto 0;\forall i,k} E_{WTA}(\widehat{\Theta}_H) = 0$.

Suppose $H' > c$, then a solution $\widehat{\Theta}_{H'}$ can be formulated s.t.: $E(\widehat{\Theta}_H) = E(\widehat{\Theta}_{H'})$.

Let $\widehat{\Theta}_{H'} = \widehat{\Theta}_H \cup \widehat{\Theta}_{H+1\dots H'} = \widehat{\Theta}_H \cup \{\theta_{h+1} \dots \theta_{h'}\}$ for some **random** $\widehat{\Theta}_{H+1\dots H'}$. Due to randomness and without loss of generality, one can assume that $\forall (x_i,y_i), \forall \theta_i \in \Theta_{H+1\dots H'}$, θ_i is not the optimal hypothesis for any training point $(x_i,y_i) \in D_{train}$.

In this case due to the winner-takes-all energy formulation we have:

$$\begin{aligned} E_{WTA}(\widehat{\theta}_{H'}) &= -\sum_i \max_{1\le h\le H'} \log p_{\theta_h}(y_i|x_i) \\ &= -\sum_i \max_{1\le h\le H} \log p_{\theta_h}(y_i|x_i) \\ &= E_{WTA}(\widehat{\theta}_H) \end{aligned} \tag{5}$$

So $\widehat{\Theta}_H$ and $\widehat{\Theta}_{H'}$ with $H' > H$ are both solutions to the loss formulation and share the same energy level. The extended hypotheses can support arbitrary artificial data regions without being penalized. □

E. Lemma 4.2

$$E_{MHP}(\Theta) = -\sum_i \sum_h \log\left(p_{\theta_h}(y_i|x_i)\right) * \begin{cases} 1-\epsilon, p_{\theta_h}(y_i|x_i) \ge p_{\theta_k}(y_i|x_i), \forall k \\ \frac{\epsilon}{H-1}, \text{else} \end{cases} \tag{6}$$

Whereby x_i,y_i is corresponding input-output pairs from the training dataset, $1 \le h \le H$ is a hypothesis branch, which is generated by a parametrized neural network with the parameter set θ_h. Furthermore, ϵ is a hyperparameter used to distribute the learning signal to the non-optimal hypotheses. Θ is the collection of all θ_h.

Lemma E.1. *Similar to Lemma D, minimizing E_{MHP} in Eq. 6 might also lead to an inconsistent approximation of the real underlying output distribution.*

Proof. First, note that $0 \le \epsilon \le \frac{H-1}{H}$, since $\epsilon < 0$ would push away non-locally optimal hypotheses from the empirical solution, $\epsilon > \frac{H-1}{H}$ would penalize the best hypothesis more than others. Both are undesired properties of MHP-learning. First consider the case where $\epsilon \mapsto \frac{H-1}{H}$:

$$\lim_{\epsilon \mapsto \frac{H-1}{H}} E_{MHP}(\Theta) = \sum_i \sum_h \log\left(p_{\theta_h}(y_i|x_i)\right) * \frac{1}{H} \tag{7}$$

$$\begin{aligned} &= \frac{1}{H} \sum_h \left(\sum_i \log\left(p_{\theta_h}(y_i|x_i)\right) \right) \\ &= \frac{1}{H} \sum_h E_{\theta_h} \end{aligned}$$

$\forall \theta_h$ and training data points (x_i, y_{ik}) the optimal least-squares solution is the mean, therefore we have:

$$\begin{aligned} \theta_h^*(y_i|x_i) &= E_{y_{ik}\sim p(y|x_i)[y_i]} \\ &= \frac{1}{l} \sum_{i=1}^{l} y_i; y_{ik} \sim p(y_i|x_i) \end{aligned}$$

In this case, all hypotheses are optimized independently and converge to the same solution similar to a single-hypothesis approach. The resulting distribution is inconsistent w.r.t the real output distribution (see Fig. 1 for an example).

Now consider $\epsilon \mapsto 1$:

$$\begin{aligned}\lim_{\epsilon \mapsto 1} E_{MHP}(\Theta) &= -\sum_i \sum_h \log\left(p_{\theta_h}(y_i|x_i)\right) \\ &\quad * \begin{cases} 1; \text{if } \theta_h \text{ is best hypothesis} \\ 0; \text{else} \end{cases} \\ &= -\sum_i \max_{1 \le h \le H'} \log p_{\theta_h}(y_i|x_i) \\ &= E_{WTA}(\Theta)\end{aligned} \tag{8}$$

In this case E_{MHP} shares the same inconsistency property with E_{WTA}. Consequently, choosing $\epsilon \in [0, \frac{H-1}{H}]$ only smoothes the penalty on suboptimal hypotheses. The risk remains that distributions induced by non-optimal hypotheses are beyond the real modes of the underlying distribution.

□

F. Related works in detail

Traditional one-class learning techniques (Schölkopf et al., 2001; Tax & Duin, 2004; Liu et al., 2008; 2012; Breunig et al., 2000) often fail in high-dimensional input domains and require careful feature selection (Zong et al., 2018).

To cope with high-dimensional domains, typically a reconstruction-based approach is used. This paradigm learns the normal data distribution during training and uses the data likelihood as an anomaly score at test time. Recently, advances in generative modeling such as Generative Adversarial Network (GAN) (Goodfellow et al., 2014) and Variational Autoencoder (VAE) (Rezende et al., 2014; Kingma & Welling, 2013) are used for anomaly detection (Zong et al., 2018; Schlegl et al., 2017; Deecke et al., 2018). However, GAN and VAE approaches have limitations in anomaly detection tasks. The GAN tends to assign less probability mass to real samples, while VAE typically regresses to the conditional means. The mean regression in VAE express the model uncertainty and falsify the reconstruction-errors for unseen images.

To address model uncertainty in VAE, the decoder is given additional expressive power with multi-headed decoders. The idea is to approximate multiple conditional modes (dense data regions) by using networks with multiple heads. This leads to training of multiple networks in Multi-Choice-learning (Dey et al., 2015; Lee et al., 2017; 2016), the estimation of a conditional Gaussian Mixture model in Mixture Density Networks (MDN) (Bishop, 1994), and multiple-hypotheses predictions (MHP) (Chen & Koltun, 2017; Bhattacharyya et al., 2018; Rupprecht et al., 2016a; Ilg et al., 2018). In MDN, the mixtures are strictly coupled via mixture coefficients while mixtures in MHPs act as loosely coupled local density estimators. In MHP, only the best hypothesis branch will receive a learning signal, i.e., the one that best explains the training sample.

For anomaly detection, our model uses MHP-training with a VAE to address the model uncertainty directly. In MDN, the anomaly score is proportional to the weighted distances to all data modes, and in MHP only to closest data mode. To highlight the change in paradigm, we refer to this learning in MHP as consistency-based learning. Samples have a small effect on the loss as long as they are close to one single data mode. The learning dynamic in MHP is also different and more efficient than in MDN: the number of samples with a large loss is much lower. In this sense, we relax the learning objective from strict density-based to consistency-based learning.

This is related to the Local Outlier Factor (LOF) approach (Breunig et al., 2000), where the outlier-score only depends on the local neighborhood. In LOF, the outlier score is proportional to the mean density of neighboring points divided by the local point density. Hence, distant samples do not influence the outlier-score. Motivated by this heuristic, our model employs learning of many loosely decoupled local density estimates with MHP-learning. While LOF computes the outlier score only at test time and directly in the input space, our model first approximates the data manifold and subsequently performs anomaly detection in the input space under the learned model.

The MHP-technique has been used for uncertainty estimation in tasks like future prediction (Rupprecht et al., 2016b) or optical flow prediction (Ilg et al., 2018). In the simplest form, the multiple network heads learn from a winner-takes-all (WTA) loss, whereby only the best branch receives the learning signal. These works extended the loss with local smoothness terms (Ilg et al., 2018) or distribution of the learning signal also to the other, non-optimal branches (Rupprecht et al., 2016b) to generate diverse and meaningful hypotheses.

The major problem of MHP-approaches is that areas not supported by samples can be covered by unused hypotheses. This is fatal for anomaly detection. Therefore, our ConAD approach employs a discriminator D to assess the quality of the generated hypotheses to avoid support of non-existent data modes. To avoid mode collapse due to the GAN framework, we employ hypotheses discrimination. In the spirit of minibatch discrimination (Salimans et al., 2016), D additionally receives pair-wise distances across a batch of hypotheses. Since a batch of real samples is typically diverse, D can detect a homogeneous batch of hypotheses as fake easily.

Table 1. CIFAR-10 anomaly detection: AUROC-performance of different approaches. The column indicates which class was used as in-class data for distribution learning. Note that random performance is at 50% and higher scores are better. Top-2-methods are marked. Our ConAD approach outperforms traditional methods and vanilla MHP-approaches significantly and can benefit from an increasing number of hypotheses.

CIFAR-10	0	1	2	3	4	5	6	7	8	9	MEAN
KDE-PCA	70.5	49.3	**73.4**	52.2	69.1	43.9	77.1	45.8	59.5	49.0	59.0
KDE-ALEXNET	55.9	48.7	58.2	53.1	65.1	55.1	61.3	59.3	60.0	52.9	57.0
OC-SVM-PCA	66.6	47.3	67.5	53.0	82.7	43.8	**78.7**	53.2	72.0	45.3	61.0
OC-SVM-ALEXNET	59.4	54.0	58.8	57.5	75.3	55.8	69.2	54.7	63.0	53.0	60.1
IF	63.0	37.9	63.0	40.8	**76.4**	51.4	66.6	48.0	65.1	45.9	55.8
GMM	70.9	44.3	69.7	44.5	**76.1**	50.5	76.6	49.6	64.6	38.4	58.5
ANOGAN	61.0	56.5	64.8	52.8	67.0	59.2	62.5	57.6	72.3	58.2	61.2
ADGAN	63.2	52.9	58.0	**60.6**	60.7	**65.9**	61.1	63.0	74.4	64.4	62.
VAE	77.1	46.7	68.4	53.8	71.	54.2	64.2	51.2	**76.5**	46.7	61.0
VAEGAN	76.2	46.9	**69.7**	52.0	75.6	53.6	58.8	55.4	75.4	46.0	60.9
OC-D-SVDD	61.7	65.9	50.8	59.1	60.9	65.7	67.7	67.3	75.9	73.1	63.2
MDN-2	76.1	46.9	68.7	53.8	70.4	53.8	63.2	52.3	**76.8**	46.7	60.9
MDN-4	76.9	46.8	68.6	53.5	69.3	54.4	63.5	54.1	76.	46.9	61.0
MDN-8	76.2	46.9	68.6	53.3	70.4	54.7	63.3	53.	76.3	47.3	61.
MDN-16	76.2	47.9	68.2	52.8	70.1	54.	63.5	52.9	76.4	46.9	60.9
MHP-WTA-2	77.3	51.6	68.	55.2	69.5	54.3	64.3	55.5	76.	51.2	62.2
MHP-WTA-4	**77.8**	53.9	65.1	56.7	66.	54.2	63.5	56.3	75.2	54.1	62.2
MHP-WTA-8	76.1	56.	62.7	58.8	62.6	55.3	61.4	57.8	74.3	54.8	61.9
MHP-WTA-16	75.7	56.7	60.9	59.8	62.7	56.	61.	56.8	73.8	57.3	62.
MHP-2	75.5	49.9	67.6	54.6	69.3	54.3	63.6	57.7	76.4	50.8	61.9
MHP-4	75.2	51.	66.	56.8	67.7	55.1	64.4	56.	76.4	51.	61.9
MHP-8	75.7	54.	65.2	57.6	64.8	55.4	62.5	54.7	75.9	53.	61.8
MHP-16	75.8	53.9	64.1	58.5	64.6	55.2	62.3	54.5	75.9	53.2	61.7
MDN+GAN-2	74.6	48.9	68.6	52.1	71.1	52.5	66.8	57.7	76.5	48.1	61.6
MDN+GAN-4	76.2	50.4	69.	52.4	71.6	53.2	65.9	58.3	75.3	48.9	62.1
MDN+GAN-8	77.4	48.3	69.3	53.1	72.2	53.7	67.9	54.	76.	51.9	62.3
MDN+GAN-16	73.6	46.9	69.4	52.2	75.3	54.1	65.7	56.8	75.3	45.4	61.4
CONAD - 2 (OURS)	77.3	60.0	66.6	56.2	69.4	56.1	70.6	63.0	74.8	49.9	64.3
CONAD - 4 (OURS)	**77.6**	52.5	66.3	57.0	68.7	54.1	**80.1**	54.8	74.1	53.9	63.9
CONAD - 8 (OURS)	77.4	**65.2**	64.8	60.1	67.0	57.9	72.5	**66.2**	74.8	**66.0**	**67.1**
CONAD - 16 (OURS)	77.2	**63.1**	63.1	**61.5**	63.3	**58.8**	69.1	**64.0**	75.5	**63.7**	**65.9**

G. Detailed performance on CIFAR-10

H. Metal anomaly results

Table 2. Anomaly detection performance on Metal Anomaly dataset. Here the anomaly detection is measured by summing up reconstructions errors over all pixel positions. This consideration is rather sensitive to noise in very high-dimensional input space such as in Metal Anomaly. The best two models are marked.

	HYPOTHESES			
MODEL	1	2	4	8
MHP	79.5=VAE	**87.6**	**83.4**	79.3
MHP+WTA		85.1	**87.8**	80.0
MDN		74.6	76.5	74.3
MDN+GAN	78.2 =VAEGAN	81.0	78.1	**81.0**
CONAD		**86.7**	81.2	**81.7**

Table 3. Anomaly detection performance on Metal Anomaly dataset by summing over the 1% most-anomalous pixels for each input image. The best two models are marked.

	HYPOTHESES			
MODEL	1	2	4	8
MHP	97.7=VAE	**99.3**	**99.0**	**98.4**
MHP+WTA		99.0	**99.0**	98.1
MDN		97.0	96.0	97.5
MDN+GAN	97.8 =VAEGAN	96.6	95.1	97.8
CONAD		**99.2**	**99.0**	**98.7**

References

Bhattacharyya, A., Schiele, B., and Fritz, M. Accurate and diverse sampling of sequences based on a best of many sample objective. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 8485–

8493, 2018.

Bishop, C. M. Mixture density networks. Technical report, Citeseer, 1994.

Breunig, M. M., Kriegel, H.-P., Ng, R. T., and Sander, J. Lof: identifying density-based local outliers. In *ACM sigmod record*, volume 29, pp. 93–104. ACM, 2000.

Chen, Q. and Koltun, V. Photographic image synthesis with cascaded refinement networks. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pp. 1520–1529, 2017.

Deecke, L., Vandermeulen, R., Ruff, L., Mandt, S., and Kloft, M. Anomaly detection with generative adversarial networks. 2018.

Dey, D., Ramakrishna, V., Hebert, M., and Andrew Bagnell, J. Predicting multiple structured visual interpretations. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2947–2955, 2015.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. Generative adversarial nets. In *Advances in neural information processing systems*, pp. 2672–2680, 2014.

Ilg, E., Çiçek, Ö., Galesso, S., Klein, A., Makansi, O., Hutter, F., and Brox, T. Uncertainty Estimates with Multi-Hypotheses Networks for Optical Flow. In *European Conference on Computer Vision (ECCV)*, 2018. URL http://lmb.informatik.uni-freiburg.de/Publications/2018/ICKMB18. https://arxiv.org/abs/1802.07095.

Kingma, D. P. and Welling, M. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Lee, K., Hwang, C., Park, K., and Shin, J. Confident multiple choice learning. *arXiv preprint arXiv:1706.03475*, 2017.

Lee, S., Prakash, S. P. S., Cogswell, M., Ranjan, V., Crandall, D., and Batra, D. Stochastic multiple choice learning for training diverse deep ensembles. In *Advances in Neural Information Processing Systems*, pp. 2119–2127, 2016.

Liu, F. T., Ting, K. M., and Zhou, Z.-H. Isolation forest. In *2008 Eighth IEEE International Conference on Data Mining*, pp. 413–422. IEEE, 2008.

Liu, F. T., Ting, K. M., and Zhou, Z.-H. Isolation-based anomaly detection. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 6(1):3, 2012.

Rezende, D. J., Mohamed, S., and Wierstra, D. Stochastic backpropagation and approximate inference in deep generative models. *arXiv preprint arXiv:1401.4082*, 2014.

Rupprecht, C., Laina, I., DiPietro, R., Baust, M., Tombari, F., Navab, N., and Hager, G. D. Learning in an Uncertain World: Representing Ambiguity Through Multiple Hypotheses. *arXiv:1612.00197 [cs]*, December 2016a. URL http://arxiv.org/abs/1612.00197. arXiv: 1612.00197.

Rupprecht, C., Laina, I., DiPietro, R., Baust, M., Tombari, F., Navab, N., and Hager, G. D. Learning in an Uncertain World: Representing Ambiguity Through Multiple Hypotheses. *arXiv:1612.00197 [cs]*, December 2016b. URL http://arxiv.org/abs/1612.00197. arXiv: 1612.00197.

Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., and Chen, X. Improved techniques for training gans. In *Advances in Neural Information Processing Systems*, pp. 2234–2242, 2016.

Schlegl, T., Seeböck, P., Waldstein, S. M., Schmidt-Erfurth, U., and Langs, G. Unsupervised anomaly detection with generative adversarial networks to guide marker discovery. In *International Conference on Information Processing in Medical Imaging*, pp. 146–157. Springer, 2017.

Schölkopf, B., Platt, J. C., Shawe-Taylor, J., Smola, A. J., and Williamson, R. C. Estimating the support of a high-dimensional distribution. *Neural computation*, 13(7): 1443–1471, 2001.

Tax, D. M. and Duin, R. P. Support vector data description. *Machine learning*, 54(1):45–66, 2004.

Zong, B., Song, Q., Min, M. R., Cheng, W., Lumezanu, C., Cho, D., and Chen, H. Deep autoencoding gaussian mixture model for unsupervised anomaly detection. *International Conference on Learning Representations.*, 2018.