Generative Well-intentioned Networks

Justin Cosentino, Jun Zhu

Department of Computer Science, Tsinghua University

Motivation

- Representing uncertainty is crucial across a variety of domains, such as autonomous vehicles, medical diagnoses, and anomaly detection
- Uncertain, incorrect labelling can result in costly errors and at times it is better to abstain from or reject a query rather than make a mistake
- Many rejection-based algorithms do not offer insights on how to reevaluate uncertain inputs to increase confidence and generate fewer rejections in the future

Generative Well-intentioned Networks

The GWIN framework includes three core components:

1. A pretrained, certainty-based classifier C that emits a prediction y_i' with certainty c_i when labeling a new observation $\boldsymbol{x_i}$
2. A rejection function $r : \{(c, y')\} \rightarrow \{\text{reject}, y'\}$ that allows the classifier to reject an instance rather than predicting its label
3. A conditional generative network G that transforms an observation $\boldsymbol{x_i}$ and noise vector $\boldsymbol{z}$ to a new representation $\boldsymbol{x_i'}$

Wasserstein GWIN with Gradient Penalty

We present an implementation of this framework: the Wasserstein GWIN.

Classifier and Reject Function

We experiment with two classifiers: a simple LeNet-5 Bayesian neural network and an Improved BNN with a more complicated architecture. We use Monte-carlo methods to sample from the networks and generate predictive probabilities.

We use a simple reject rule for each (y_i', c_i) pair and a rejection bound τ:

$$r(c_i, y_i') = \begin{cases} y_i', & \text{if } c_i \geq \tau \\ \text{reject}, & \text{otherwise.} \end{cases} \tag{1}$$

Wasserstein GWIN with Gradient Penalty (WGWIN-GP)

The Wasserstein GWIN is based on the Wasserstein GAN with Gradient Penalty (WGAN-GP) [1]. The WGWIN's generator and critic are conditioned on the input image [2] and the class label [3], respectively. We modify the generator's loss function to take into account the classifier's loss on the transformed observations. The critic is only trained on high-certainty images that the classifier labels correctly.

Loss with Transformation Penalty

The new WGWIN-GP loss function builds on top of the WGAN-GP loss function and penalizes decreases in classifier loss due to WGWIN-GP transformation:

$$L = \underbrace{\mathop{\mathbb{E}}_{\boldsymbol{x'} \sim \mathbb{P}_g}[D(\boldsymbol{x'}, y)] - \mathop{\mathbb{E}}_{\boldsymbol{x} \sim \mathbb{P}_c}[D(\boldsymbol{x}, y)]}_{\text{WGAN Loss}} + \underbrace{\lambda_{GP} \mathop{\mathbb{E}}_{\hat{\boldsymbol{x}} \sim \mathbb{P}_{\hat{\boldsymbol{x}}}}[(||\nabla_{\hat{\boldsymbol{x}}} D(\hat{\boldsymbol{x}}, y)||_2 - 1)^2]}_{\text{WGAN-GP Penalty}} + \underbrace{\lambda_{Loss} \mathop{\mathbb{E}}_{\boldsymbol{x'} \sim \mathbb{P}_g}[\text{Loss}(C(\boldsymbol{x'}))]}_{\text{Transformation Penalty}} \tag{2}$$

where $\text{Loss}(C(\boldsymbol{x'}))$ denotes the classifier's loss given the transformed image and the correct class.

TL;DR

We propose GWINs, a novel **framework** combining a **fixed, certainty-based classifier** with a **reject option** and a **conditional generative network**.

The conditional generative network learns the distribution of observations that the classifier labels **correctly and with high certainty**.

During inference, the classifier can **reject uncertain observations**. The generative network **transforms** low-certainty queries rejected by the classifier to high-certainty representations that are then **relabeled** by the classifier.

The capability of a Wasserstein GAN (WGAN)-based proof of concept is assessed using benchmark classification datasets and shows that GWINs improve, and rarely worsen, the accuracy of rejected image classification.

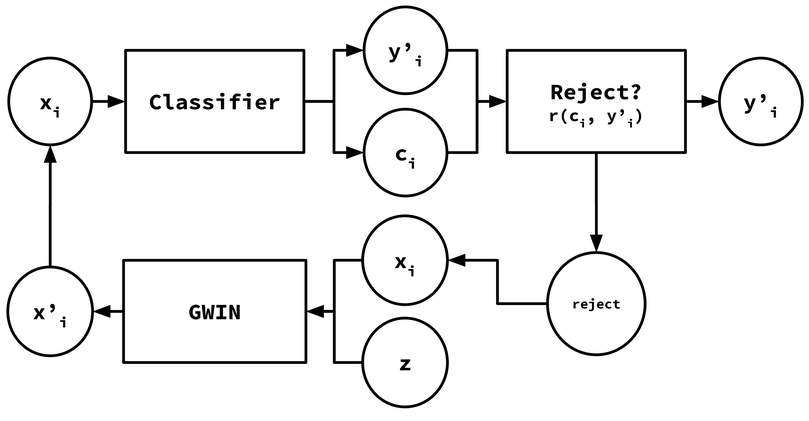


Figure 1. The inference process for some new observation $\boldsymbol{x_i}$. If classifier C labels the input y_i' with certainty c_i and rejects the query, the conditional GWIN translates the given query to the classifier's confident distribution. The transformed query $\boldsymbol{x_i'}$ is then relabeled.

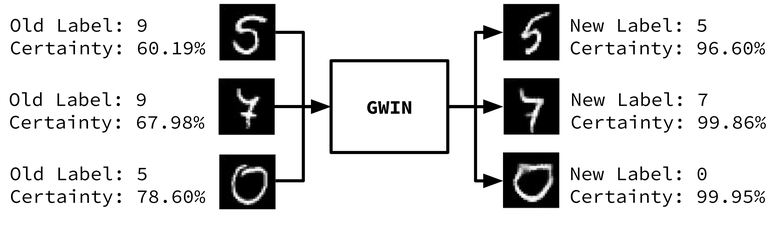
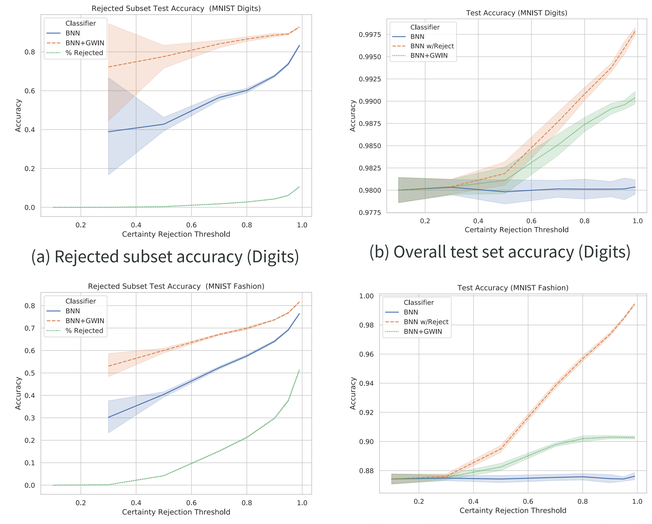


Figure 2. A visual representation of the GWIN transformation using example images from the MNIST Digits dataset. With a certainty threshold of $\tau = 0.8$, the classifier rejects the observations on the left, which would had been labeled incorrectly were the classifier forced to predict. These observations are then transformed into the representations on the right. When relabeling the generated images, the classifier labels correctly with high-certainty.

Experimental and Results



(a) Rejected subset accuracy (Digits)

(b) Overall test set accuracy (Digits)

(c) Rejected subset accuracy (Fashion)

(d) Overall test set accuracy (Fashion)

Figure 3. Test set accuracy for MNIST Digits and Fashion using GWIN transformation for varying certainty thresholds τ.

τ	% Reject	BNN Acc.	BNN+GWIN Acc.	Rejected Acc. Δ	Overall Acc. Δ	% Error Δ
0.70	1.83	54.48 ± 2.21	85.07 ± 2.63	30.59 ± 2.64	0.56 ± 0.06	-27.55 ± 2.66
0.80	2.74	58.91 ± 1.49	86.30 ± 1.85	27.39 ± 2.03	0.75 ± 0.06	-36.36 ± 1.93
0.90	4.39	68.79 ± 2.38	86.95 ± 0.97	18.16 ± 2.55	0.80 ± 0.13	-40.26 ± 4.19
0.95	6.04	73.48 ± 1.66	89.34 ± 0.85	15.86 ± 2.07	0.96 ± 0.13	-47.45 ± 4.09
0.99	11.00	83.54 ± 0.88	92.55 ± 0.49	9.02 ± 0.94	0.99 ± 0.10	-49.45 ± 3.16
0.70	15.25	52.08 ± 1.55	66.95 ± 0.67	14.87 ± 1.78	2.27 ± 0.30	-18.08 ± 1.98
0.80	21.21	57.87 ± 0.89	69.16 ± 0.47	11.29 ± 0.87	2.39 ± 0.19	-19.25 ± 1.32
0.90	30.29	64.14 ± 0.66	73.18 ± 0.73	9.04 ± 0.83	2.74 ± 0.29	-21.63 ± 1.85
0.95	37.30	68.93 ± 0.49	76.06 ± 0.43	7.14 ± 0.61	2.66 ± 0.25	-21.15 ± 1.61
0.99	51.97	76.55 ± 0.30	81.34 ± 0.26	4.79 ± 0.34	2.49 ± 0.19	-19.94 ± 1.30

Table 1. Test set accuracy for MNIST Digits (top) and Fashion (bottom) on rejected observations using GWIN transformation for the given certainty threshold τ. *BNN* and *BNN+GWIN* denote accuracy for the rejected subset using only the BNN and the BNN with GWIN reformulation, respectively. With no rejections ($\tau = 0$), the BNN had an accuracy of 98.0% on Digits and 87.4% on Fashion.

Conclusions

GWINs show potential for improving rejection-based classifier accuracy using certainty estimates. Since the rejected sample size is small relative to all possible queries, we must improve accuracy when the rejection bound τ is very high. Ask about future work!

[1] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville. Improved training of wasserstein gans. In *Advances in Neural Information Processing Systems*, 2017.

[2] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets, 2014.

[3] Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. Generative adversarial text to image synthesis. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48*, 2016.