EC-GAN: Low-Sample Classification using Semi-Supervised Algorithms and GANs

Ayaan Haque†

†Saratoga High School

Overview

- **The Problem**: Many datasets are so restricted that even finding unlabeled samples is challenging. Semi-supervised learning is a good alternative to fully-labeled datasets, but it requires unlabeled data. Moreover, current methods for GAN-based semi-supervised learning employ multi-tasking when it may not be applicable.
- **The Solution**: One of the most effective methods of improving deep learning models is increasing the size of datasets. Using a Generative Adversarial Network [1], additional artificial data can be generated and fed to classifiers as unlabeled supplemental data.
- **EC-GAN**: We use GANs and semi-supervised algorithms to produce unlabeled artificial data for classification, in essence increasing the size of datasets. We importantly separate the tasks of classification and discrimination, challenging the popular multi-tasking framework.

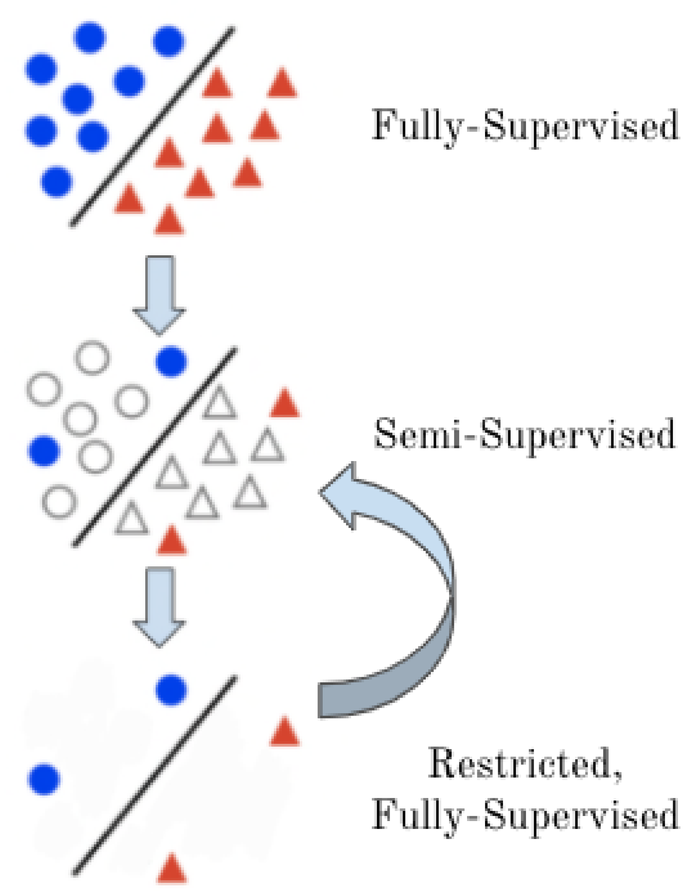

Fig. 1: Comparison of fully-supervised learning, semi-supervised learning, and restricted, fully-supervised learning. We address restricted, fully-supervised learning by changing the problem to a semi-supervised problem.

EC-GAN Generations

Fig. 2: Real images (left) compared to EC-GAN generated images (right).

- EC-GAN accurately produces realistic images on both benchmark and real datasets, as many prevalent features are visible, making them well-suited to be used as classification data.

Methods

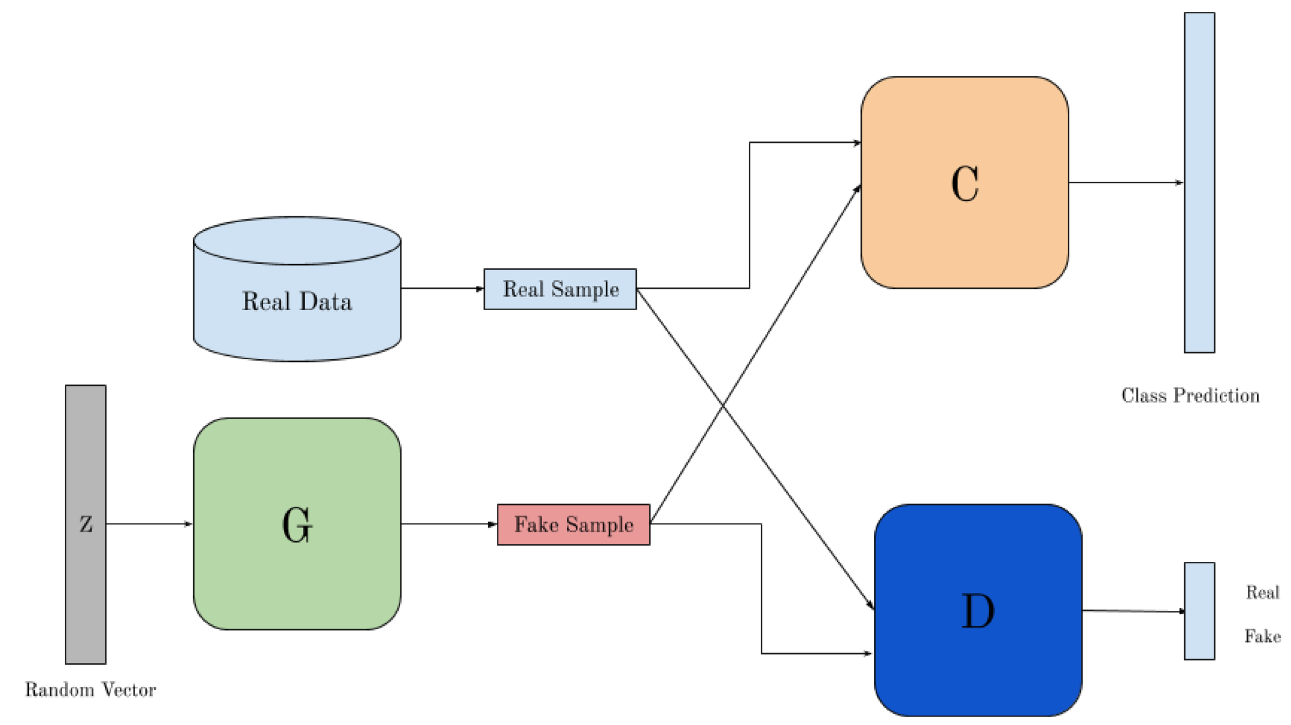

Fig. 3: Illustration of EC-GAN. The classifier is trained on real and fake images, generator and discriminator produce realistic images, and each model is optimized for a single task.

- EC-GAN model consists of **three separate networks**: a generator, a discriminator, and a classifier (Figure 3).
- At every training iteration, the generator is given random noise vectors (z) and generates new images, while the discriminator predicts between real and fake images. Simultaneously, a classifier is trained in supervised fashion on the labeled data.
- **Generated images** are subsequently fed to the classifier without labels, **increasing** the size of the dataset. Classifier is trained in **semi-supervised** fashion to learn from generated images.

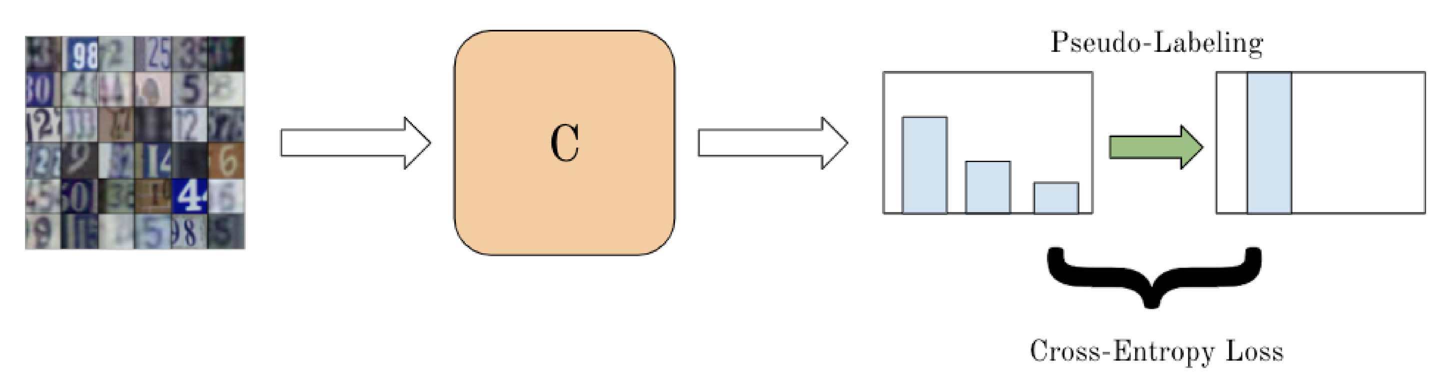

Fig. 4: Pseudo-labeling uses the predicted class of an unlabeled data as the label, assuming the prediction is confident above a threshold t.

- To create labels for the artificial samples, we use a **pseudo-labeling** (Figure 4) scheme which assumes a label based on the most likely class prediction according to the current state of the classifier [3].
- Loss of classifier (L_C), discriminator (L_D), and generator (L_G) can be written respectively as follows:

$$L_C(x,y,z) = CE(C(x),y) + \lambda CE(C(G(z)), argmax(C(G(z))) > t) \quad (1)$$

$$L_D(x,z) = BCE(D(x),1) + BCE(D(G(z)),0) \quad (2)$$

$$L_G(z) = BCE(D(G(z)),1) \quad (3)$$

- Each model has its own loss as opposed to a singular loss for multi-tasking, but all loses are intertwined, preserving a mutually-beneficial arrangement.
- The pseudo-labels are only retained if the probability is above a specific **confidence threshold** "t" (Equation 1), ensuring only accurate GAN images are used for classification.

Implementation Details

- Academic benchmark dataset SVHN (development and testing), 73,257 training images, 26,032 validation images, 32x32 size
- Real-world dataset for Pneumonia classification chest X-Ray dataset, 5,863 total images, <10% of SVHN, resized to 64x64 [2]
- Varied dataset sizes for experiments, comparisons against SOTA and baselines

Results

Dataset Size (%)	EC-GAN (%)		Shared DCDiscriminator (%)	
	Classifier	GAN	Classifier	GAN
10	88.63	91.15	83.54	86.17
15	90.88	92.21	85.20	88.72
20	92.61	93.40	86.77	89.39
25	92.89	93.93	87.58	87.93
30	93.12	94.32	87.78	90.62

Tab. 1: EC-GAN is compared to the shared architecture method on SVHN at different dataset sizes.

Dataset Size (%)	EC-GAN (%)	
	Classifier	GAN
25	94.37	96.48
50	95.24	97.83
75	95.64	97.40
100	96.42	97.99

Tab. 2: The conditional version of EC-GAN is tested on the X-ray dataset and compared against a baseline.

- EC-GAN performs on par and occasionally better than the shared architecture in small datasets (Table 1), matching state-of-the-art performance
- On both an academic and real-world datasets (Table 2), EC-GAN significantly improves accuracy metrics compared to baselines

Conclusion

- EC-GAN is a semi-supervised generative model which improves classification through the use of artificial data and pseudo-labeling. Our competing framework yields results that match the state-of-the-art.
- Our future work aims to integrate the classifier in into the adversarial framework as well as using new semi-supervised algorithms, potentially leveraging Conditional-GANs for labeling.

References

[1] Ian J. Goodfellow et al. "Generative Adversarial Nets". In: *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*. 2014, pp. 2672–2680.

[2] Daniel S. Kermany, K. Zhang, and M. Goldbaum. *Labeled Optical Coherence Tomography (OCT) and Chest X-Ray Images for Classification.* 2018.

[3] Dong-Hyun Lee. "Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks". In: *Workshop on challenges in representation learning, ICML*. Vol. 3.