Model Compression with Generative Adversarial Networks

Ruishan Liu*, Nicolo Fusi† and Lester Mackey†

*Stanford University, †Microsoft Research

Model Compression

Motivation: More accurate machine learning models often demand more computation and memory at test time, making them difficult to deploy on CPU- or memory-constrained devices.

Model compression trains a less expensive student model to mimic the expensive teacher model while maintaining most of the original accuracy.

Problem: The teacher's training data is typically reused for compression, leading to suboptimal performance

Our Contributions

GAN-assisted model compression (GAN-MC): We augment the compression dataset with synthetic data from a generative adversarial network (GAN).

Deep neural network GAN-MC: On CIFAR-10 image classification, GAN-MC consistently improves student test accuracy across architectures and losses.

Random Forest GAN-MC: For random forest teachers, we demonstrate 25 to 336-fold reductions in execution and storage costs with less than 1.2% loss in test performance across a suite of real-world tabular datasets.

Compression Score: We introduce a new measure for evaluating the quality of GAN-generated datasets and illustrate its advantages over the popular Inception Score on CIFAR-10.

DNN Compression

Given a compression dataset of n feature vectors paired with teacher logit vectors, $\{(x^{(1)}, z^{(1)}), ..., (x^{(n)}, z^{(n)})\}$, [1] framed the compression task as multitask regression with L^2 loss,

$$L(\theta) = ||g(x;\theta) - z||_2^2.$$

$g(x;\theta)$ is the vector of logits predicted by the student for feature vector x.

[2] introduced an alternative compression objective function, indexed by a temperature parameter $T > 0$. Specifically, the student is trained to mimic the annealed teacher class probabilities,

$$q_j(z/T) = \frac{\exp(z_j/T)}{\sum_k \exp(z_k/T)},$$

for each class j by solving a multitask regression problem with cross-entropy loss,

$$L_T(\theta) = -\sum_j q_j(z/T) \log(q_j(g(x;\theta)/T)).$$

Random Forest Compression

Focusing on the common setting of binary classification with labels in $\{0, 1\}$, we propose to train a student regression random forest to predict a teacher forest's outputted probability p of a datapoint x having the label 1.

GAN-assisted Model Compression (GAN-MC)

Main Idea

When fresh data is unavailable for model compression, we augment the compression dataset with synthetic feature vectors from a generative adversarial network (GAN) designed to approximate the training data distribution.

We use the **auxiliary classifier GAN (AC-GAN)** of [3].

The generator G produces synthetic feature vectors $X_{fake} = G(W, C)$ from random noise W and class label $C \sim p_c$

For each feature vector x, discriminator D predicts the probability of each class label $P(C \mid x)$ and of the data source being real or fake, $P(S \mid x)$ for $S \in \{real, fake\}$

Given a training dataset $\mathcal{D}_{real}$, the training objectives are the expected conditional log-likelihood of the correct source and the correct class of a feature vector:

$$L_{source} = \frac{1}{|\mathcal{D}_{real}|}\sum_{(x,c)\in\mathcal{D}_{real}} \log P(S = real \mid x) + \mathbb{E}[\log P(S = fake \mid G(W, C))]$$

$$L_{class} = \frac{1}{|\mathcal{D}_{real}|}\sum_{(x,c)\in\mathcal{D}_{real}} \log P(C = c \mid x) + \mathbb{E}[\log P(C \mid G(W, C))],$$

In the adversarial game, the generator G is trained to maximize $L_{class} - L_{source}$, and the discriminator D is trained to maximize $L_{class} + L_{source}$.

Convolutional Neural Networks on CIFAR-10

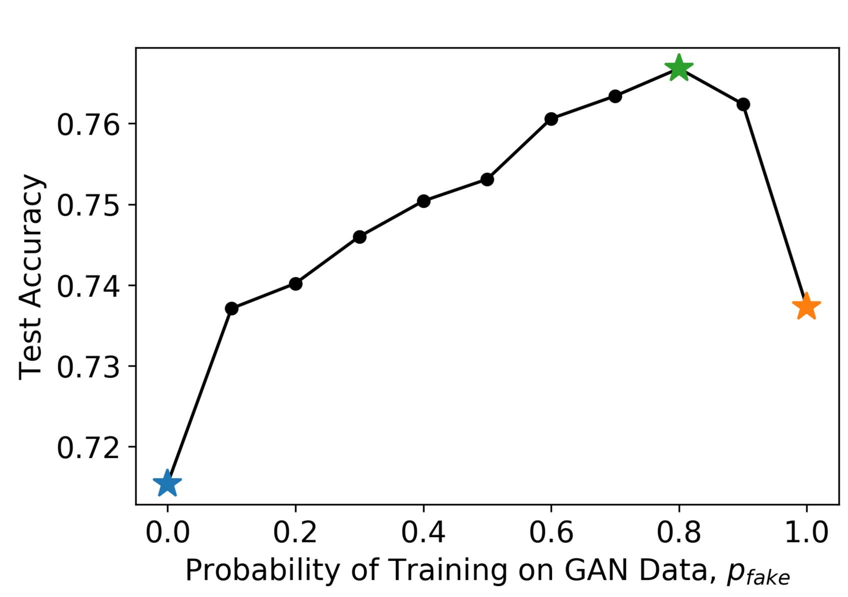


Figure: GAN-MC student accuracy using different mixtures of GAN and training data ($p_{fake} = 0 \Rightarrow$ only training data)

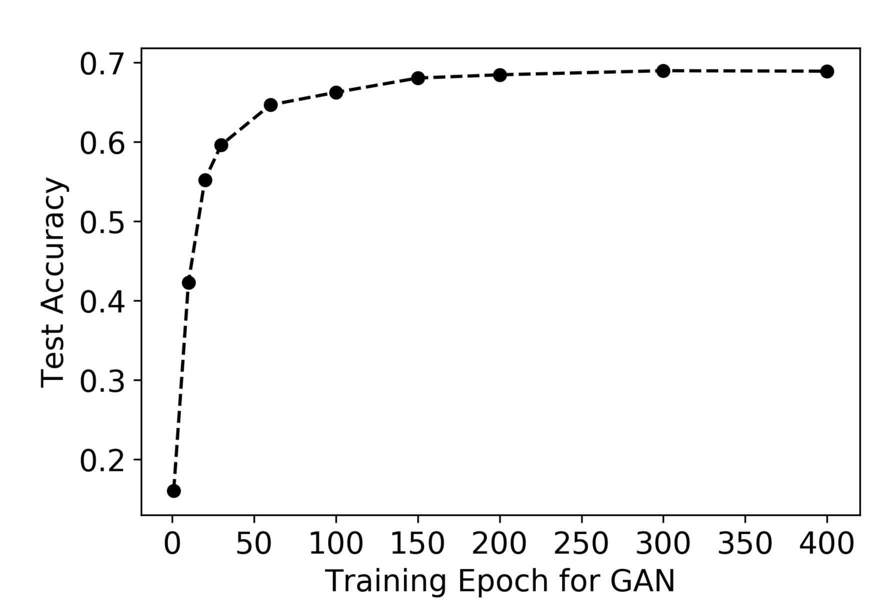


Figure: Effect of GAN quality on GAN-MC student test accuracy

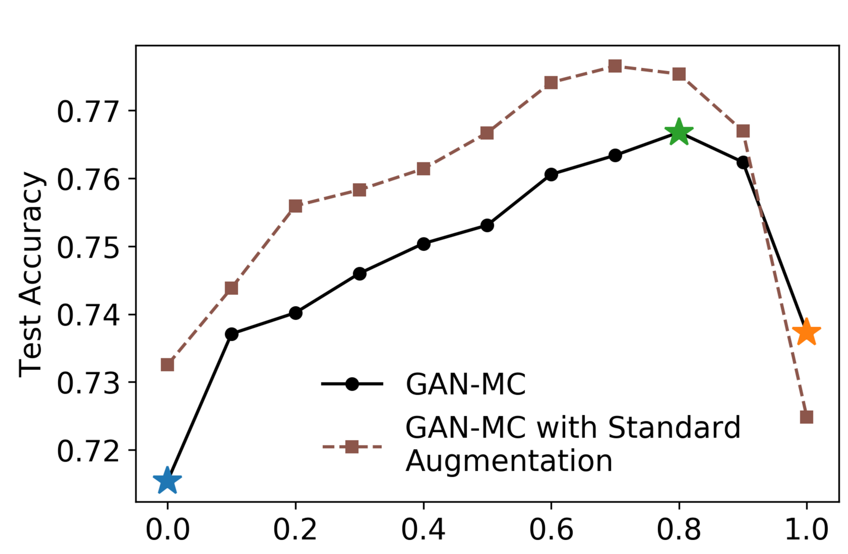


Figure: GAN-MC complements standard image augmentation

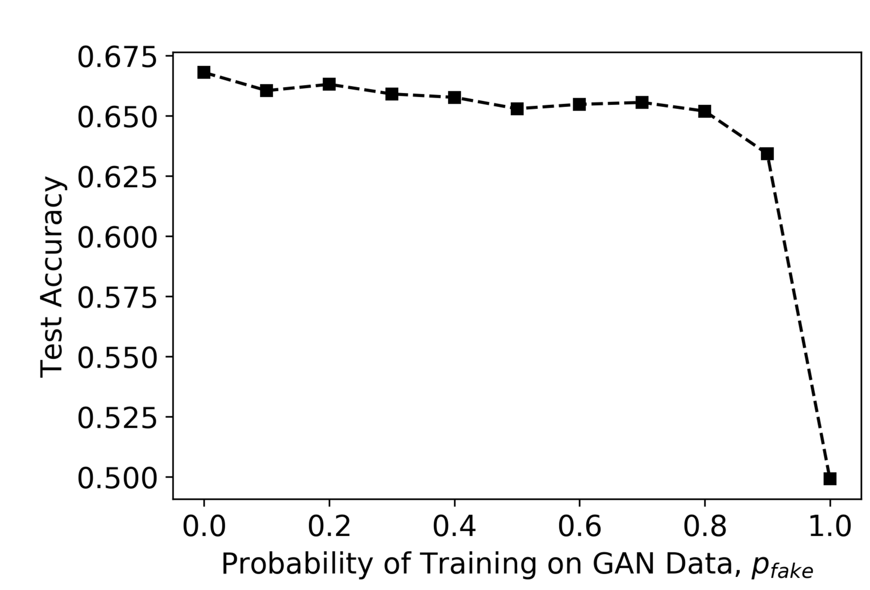
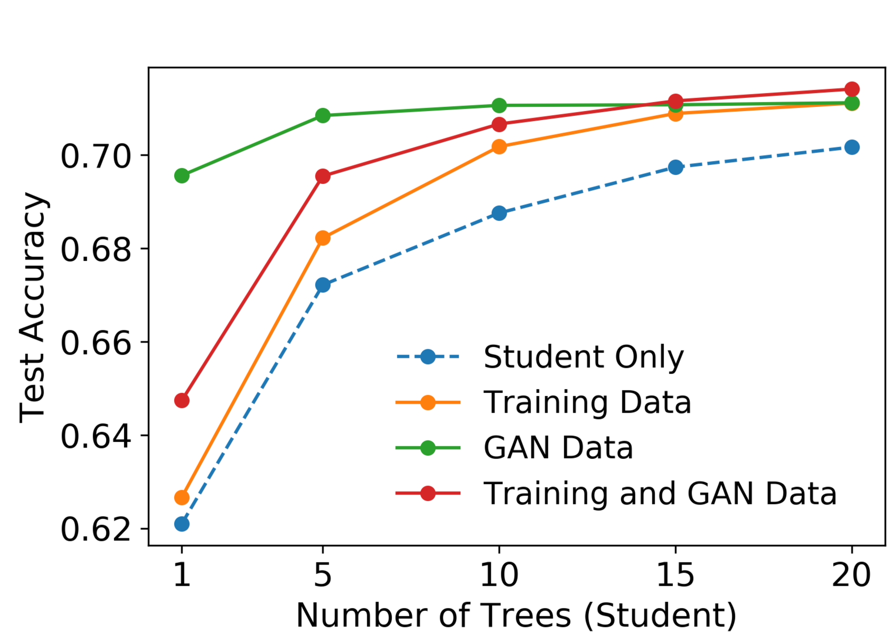


Figure: Unlike GAN-MC, augmenting original supervised learning task with GAN data impairs accuracy

	Teacher	Student	Teacher Only	Student Only	Student after Compression with Training Data	Student after Compression with Training & GAN
1	NIN	LeNet	78.1%	66.2%	71.0%	**75.3%**
2	ResNet-18	5-layer CNN	94.2%	78.8%	84.4%	**86.6%**
3	WideResNet-28-10	ResNet-18	95.8%	94.2%	94.3%	**95.0%**

Random Forest GAN-MC



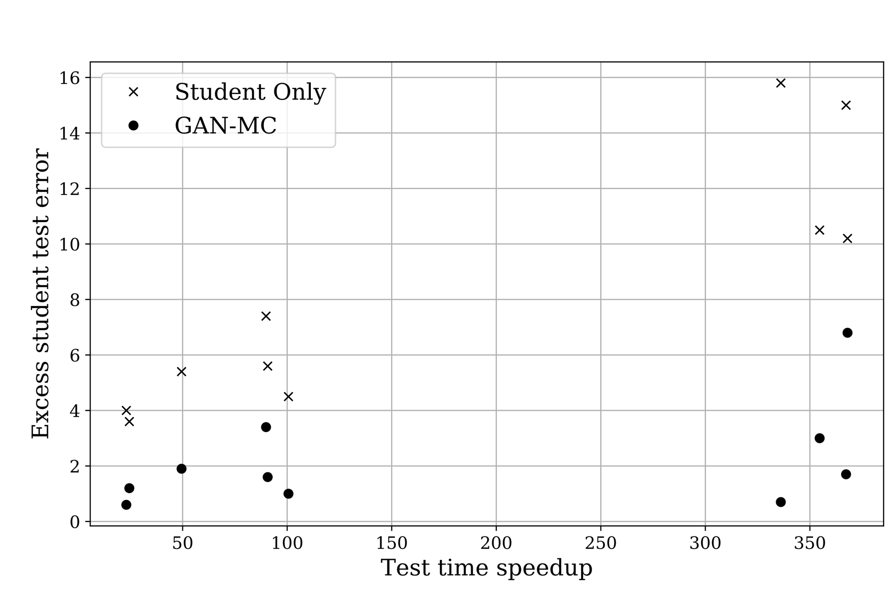


Figure: (left) Student accuracy on Higgs 100k; (right) Student error vs. speed-up across tabular datasets

Compression Score

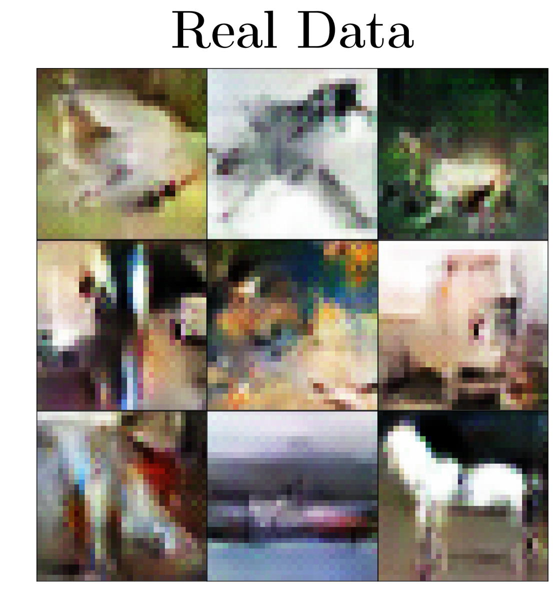
To evaluate the quality of a generated dataset $\mathcal{D}$ relative to a real dataset $\mathcal{D}_{real}$, we define a **Compression Score** based on the test accuracy $\text{acc}(\mathcal{D})$ of a student trained for one epoch with compression set $\mathcal{D}$ to mimic a teacher pre-trained on $\mathcal{D}_{real}$:

$$\texttt{CompressionScore}(\mathcal{D}; \mathcal{D}_{real}) = \frac{\text{acc}(\mathcal{D}) - \text{acc}_{\text{mode}}}{\text{acc}(\mathcal{D}_{real}) - \text{acc}_{\text{mode}}}.$$

The term acc_{mode} represents the accuracy obtained by always predicting the most common class in $\mathcal{D}_{real}$. A higher Compression Score is designed to indicate a higher quality dataset $\mathcal{D}$.

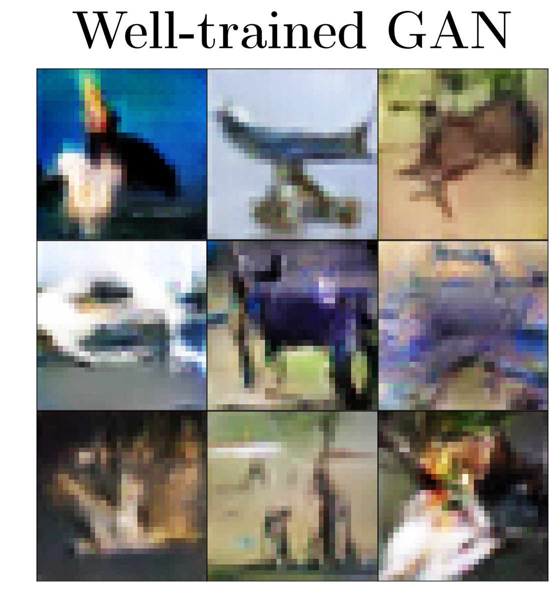
Compression vs. Inception

Real Data



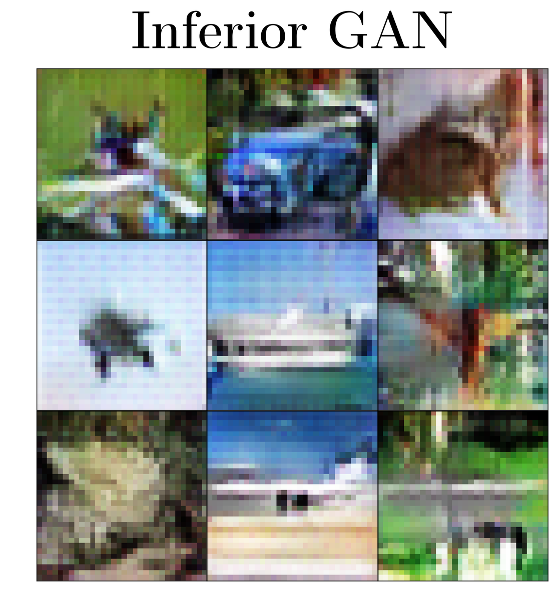
Inception: 11.2 ± 0.1
Compression: 0.994 ± 0.003

Well-trained GAN



Inception: 5.80 ± 0.06
Compression: 0.778 ± 0.002

Inferior GAN



Inception: 5.93 ± 0.06
Compression: 0.702 ± 0.002

References

[1] Jimmy Ba and Rich Caruana. Do deep nets really need to be deep? In *Advances in neural information processing systems*, pages 2654–2662, 2014.

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

[3] Augustus Odena, Christopher Olah, and Jonathon Shlens. Conditional image synthesis with auxiliary classifier GANs. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 2642–2651, 2017.

Contact

ruishan@stanford.edu,
{fusi, lmackey}@microsoft.com