The Thirty-Second AAAI Conference
on Artificial Intelligence (AAAI-18)

Adversary Is the Best Teacher: Towards Extremely Compact Neural Networks

Ameya Prabhu,* Harish Krishna,* Soham Saha
Center for Visual Information Technology
Kolhi Center for Intelligent Systems
IIIT-Hyderabad
{ameya.prabhu,harishkrishna.v,soham.saha} @ research.iiit.ac.in
asterisk indicates equal contribution

Abstract

With neural networks rapidly becoming deeper, there emerges a need for compact models. One popular approach for this is to train small student networks to mimic larger and deeper teacher models, rather than directly learn from the training data. We propose a novel technique to train student-teacher networks without directly providing label information to the student. However, our main contribution is to learn how to learn from the teacher by a unique strategy - having the student compete with a discriminator.

Introduction

The recent boom in deep neural networks has resulted in their being used for a wide variety of applications, many of which find significance when run on memory-constrained environments. Popular methods for neural network compression aim to achieve a reduction in the number of parameters while retaining state-of-the-art results. A seminal work on model compression was by Hinton et al [2] who introduced a technique in which a small *student* network learns from a large *teacher* network that is trained to saturation. The teacher network has to learn to represent complex structures in the data and in a way, pass its understanding to the student model thereby enabling it to perform better than if trained in a plain supervised fashion. Any neural network for classification in essence transforms an input from the space of the input to a point on the manifold of dimension of the n classes. The student-teacher networks try to penalize the distance between the points common to manifolds of the teacher and student networks.

Our Contributions

Our first contribution is to learning a manifold by a student neural network from a corresponding point in the manifold of its teacher network, i.e without any direct supervision. We train student networks purely with a similarity loss, and are able to achieve accuracies comparable to that of student network trained in a supervised manner. We build on this by our second and our main contribution which is learning how to learn the manifold of the student itself. We achieve this modeling the learning problem as a game between the student network and an adversary. Simply stated, the adversarial network takes a point in a manifold as its input with the objective is to classify whether this point is from the manifold of the teacher or of the student. The goal of the student network is to try fool the adversary into believing that it is the teacher, and in this process, learn the manifold. To the best of our knowledge, it is the first time an adversarial setting is used to learning how to mimic and provides significant improvement in accuracy over supervised loss functions and traditional reconstruction loss used typically in student-teacher networks.

Why is our contribution important to the community? Learning without any explicit supervision for a task *ipso facto* provides interesting properties to our approach. An example is that the learning method is domain and task independent, since instead of learning a given task, we learn a way to learn that from the teacher. Hence, it should be well suited to classification, retrieval, clustering or any other method across domains. Another interesting fact about this approach is that humans learn in a similar way too - they learn by mimicking those they consider teachers while being motivated by a critic (healthy competition). In our experiments, we are able to achieve competing accuracies with the state-of-the-art student-teacher methods in the domain usually used for benchmarking these approaches, with upto 40% lesser parameters and no supervision, when compared to the state-of-the-art compressed student-teacher networks and over 16x compression when compared to other methods.

Loss function

If $t(x)$ denotes the output of the teacher network for input x and $f(x|\theta)$ the output of a student network parameterized by θ, our loss function $\mathbf{L_T}$ is a weighted sum of the following terms:

Reconstruction loss This matches function values and forces both the networks to produce the same output,

$$\mathbf{L_R} = \frac{1}{2}||f(x|\theta) - t(x)||^2$$

Derivative loss The reconstruction loss is often insufficient; since multiple manifolds can fit through the same set of points. This problem can be relieved by augmenting the loss function with *derivative loss* i.e. penalizing the mean-

Algorithm 1 : The imitation game
Steps to train our adversarial student network

```
Reconstruction and Derivative Loss
y_stud = Student(X)                                  ▷ Forward pass
y_teach = Teacher(X)                                 ▷ Forward pass

Compute output gradients of student ∂f/∂x & teacher ∂t/∂x
Calculate L_R and L_D
L_T = α · L_R + β · L_D

if adversary == 'nil' then                           ▷ Non-Adversarial Variant
    Backpropogate L_T thru student                   ▷ Backward pass

    Repeat
If Adversarial Loss
y_inp = Random(y_stud , y_teach)
isStud , y_pred = Adversary(y_inp)                   ▷ Forward pass

Compute L_AR & L_AD and update L_T += γ · L_AR

if y_inp = y_stud then                               ▷ Student's input
    Backpropogate L_T thru student                   ▷ Backward pass
Backpropogate L_Disc thru discrimintor               ▷ Backward
```

squared-error between tangent hyperplane for the reconstruction loss to the corresponding points in the student and teacher manifolds. These are obtained by taking the derivative of the outputs with respect to x. Mathematically, the derivative square error is stated as:

$$\mathbf{L_D} = \frac{1}{2}||\frac{\partial f}{\partial x} - \frac{\partial t}{\partial x}||^2$$

Adversarial Reconstruction Loss The adversarial loss helps us learn a complicated function involved in finding distances between manifolds to the network. An interesting insight into adversarial loss functions is that they force the student output to be on some point on the manifold of the teacher output rather than the mean of all possible values, which might lie outside the manifold. It is achieved by making the student and discriminators play a game where they try to fool each other. The loss on the discriminator will be

$$\mathbf{L_{Disc}} = -logD(t(x)|y) - log(1 - D(f(x)|y))$$

where y denotes the class label and D is the discriminator. The corresponding generator loss would therefore be

$$\mathbf{L_{AR}} = -logD(f(x)|y)$$

Implementation & Experiments

We empirically demonstrate the effectiveness of our approach on CIFAR-10 dataset. We train the student network using the losses as reported in Table 1, and then test it by using cross-entropy loss between its output and target labels. Further details about the hyper parameters, model architecture of student and teacher, training procedure, etc are attached in the supplementary material.

Models	Loss Func	Accuracy
Ours	Rec	84.3%
Ours	Rec + Der	85.5%
Ours	Adv	85.8%
Ours	All Combined	**87.2%**

Table 1: This is a comparison with different loss functions applied on the model. Rec- reconstruction loss, Der- derivative loss, Adv- adversarial loss

Models	Params	Accuracy
Mimic Single	54M	84.6%
Mimic Ensemble	70M	85.8%
Urban et al.	10M	92.6%
Fitnets	2.5M	91.6%
Compressed Models		
Urban etal	1M	91.3%
Fitnet-1	250K	89.0%
Ours	**150K**	87.2%

Table 2: Comparison with state-of-the-art models

From Table 1, we observe that our method recovers accuracies comparable to the original supervised network when training using reconstruction loss only and it improves by 1.2% on the introduction of the derivative loss. However, training with purely adversarial loss, achieving 0.3% accuracy improvement over the combined reconstruction and derivative loss. Note that these accuracies are obtained by merely mimicking the teacher, i.e the student network has never seen any labels. We observe that we can combine all the three losses and achieve an accuracy of 87.2%, showing that all of the methods can work in tandem.

Conclusion

We work on the idea of student networks being able to learn exclusively from teacher networks i.e. without explicit supervision. We further show an approach to train student networks by only adversarial training. We show that our extremely compact networks are able to learn to mimic and significantly outperform the supervised approaches. We believe that the future implications of this approach seem very interesting, and we can develop significantly on settings which are able to fully exploit the strengths of this kind of training. Also, we need to rethink how big neural networks really need to be, and develop techniques to compress more information into smaller models in the light of the same. We hope we are able to create a roadmap to developing better, faster, more accurate compact models which would help deep learning models becoming ubiquitous.

References

Tangent prop: A formalism for specifying selected invariances in adaptive networks, NIPS 1992 Hinton G., Vinyals O., Dean J. Distilling the Knowledge in a Neural Network. Ph.D. diss., Deep Learning Workshop, NIPS 2014