Knowledge Distillation via Generative Adversarial Networks

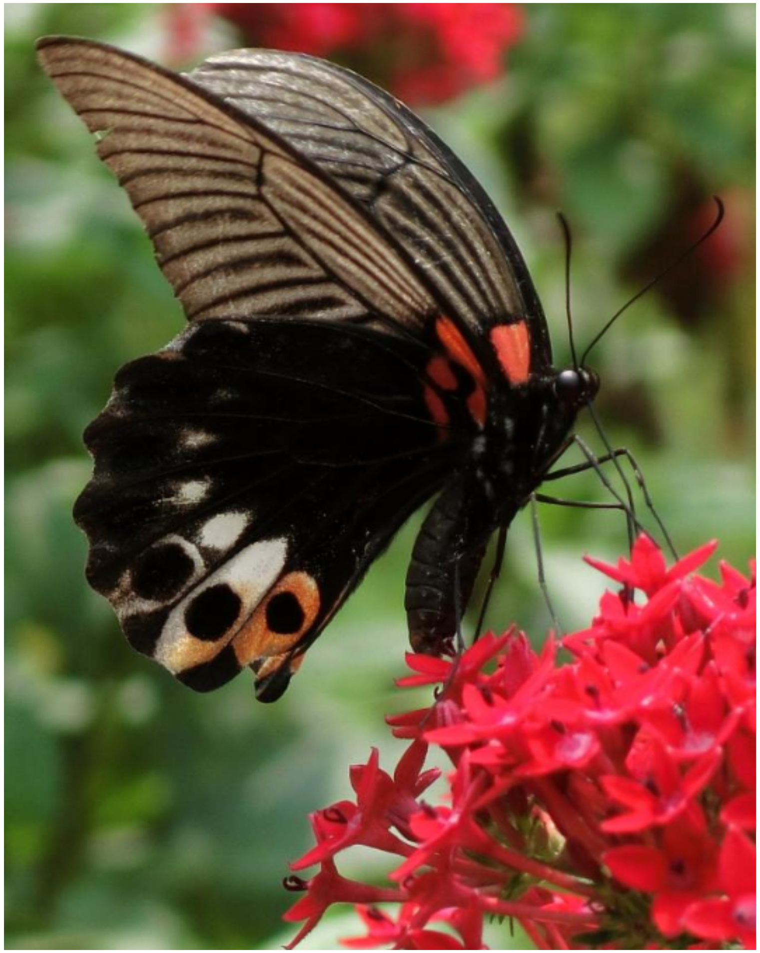
Augmented Intelligent And Interaction (AII) Workshop

李哲榮

2018/05/19

Models for Smaller Devices

- Different forms for different stages



Methods

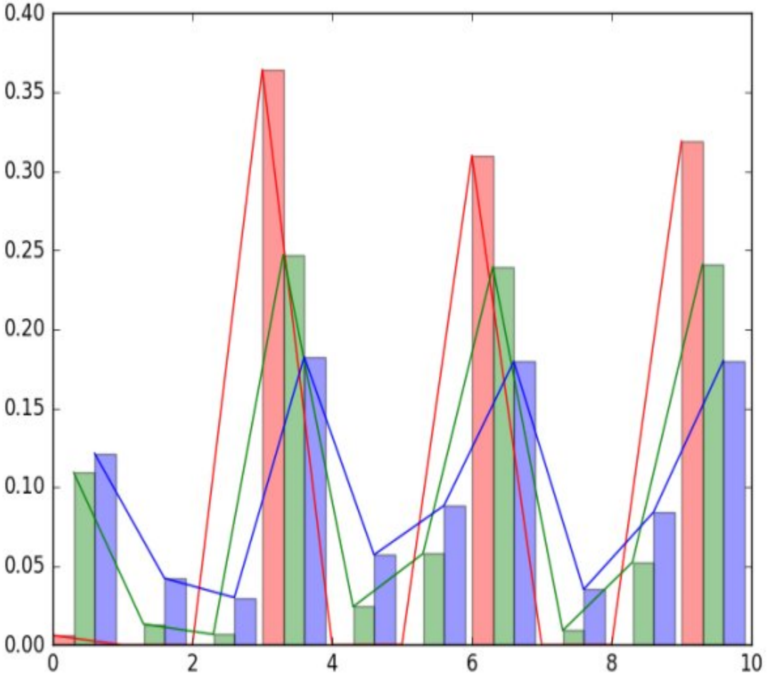
- Model compression
  – Pruning, quantization, data compression…
- Matrix/tensor decomposition
  – PCA, SVD, CPD, Sparse coding, fast convolution…
- Smaller models
  – SqueezeNet, MobileNet, …
- Architecture search
  – PPP-net, …
- Knowledge distillation

Knowledge Distillation

- Geoffrey Hinton, Oriol Vinyals, Jeff Dean (2015)
- soften output
- modify softmax

$$q_i = \frac{exp(z_i/T)}{\sum_j exp(z_j/T)}$$

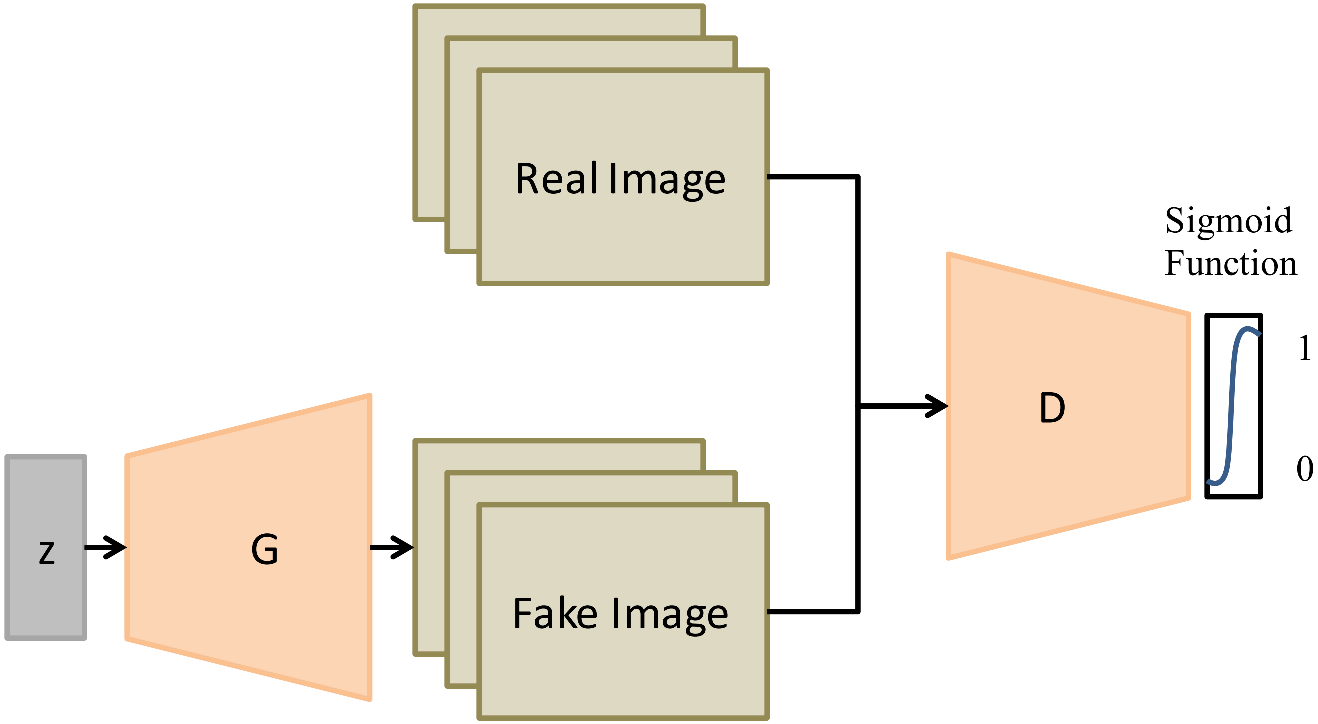
T = 5 ,25, 50



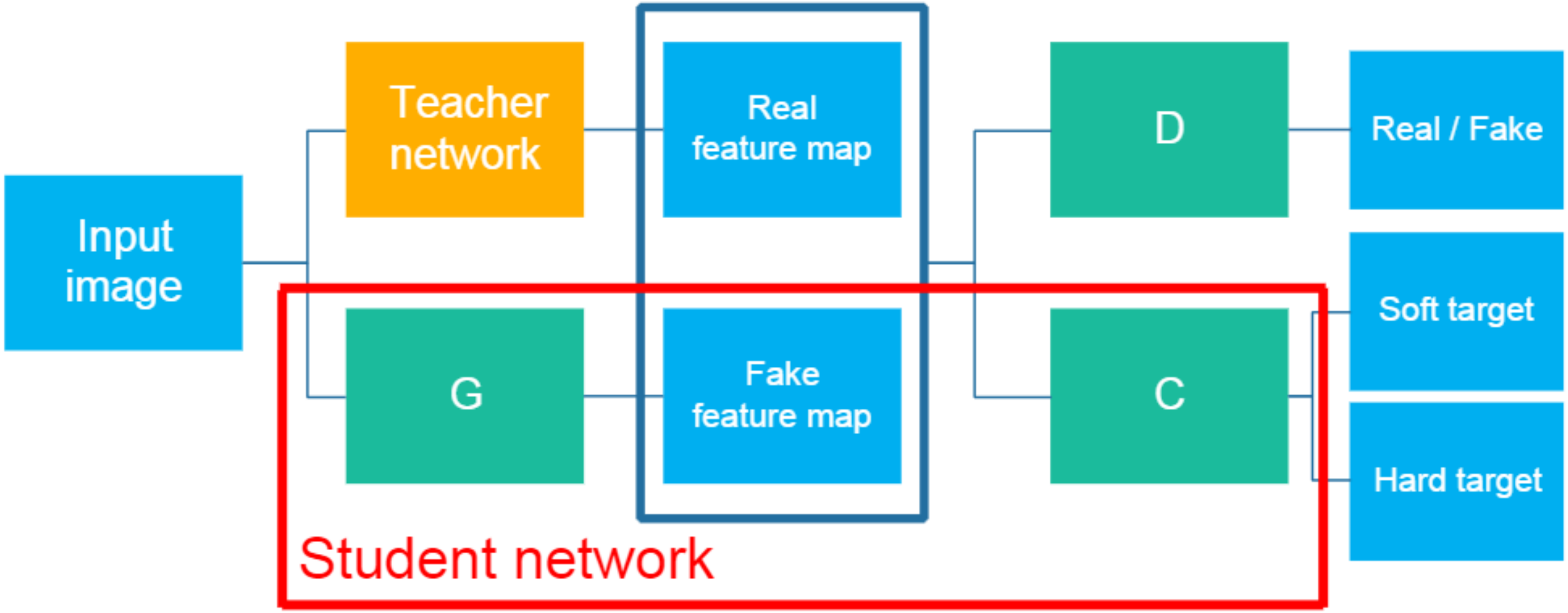
Generative Adversarial Networks

- Vanilla GAN

$$\min_{G}\max_{D} V(D,G) = \log D(x) + \log(1 - D(G(z)))$$



KDGAN


Input image
Teacher network
G
Real feature map
Fake feature map
D
C
Real / Fake
Soft target
Hard target
Student network

Experiments

Method	Accuracy
Baseline	68.53%
Logits Mimic Learning	50.95%
KD	69.14%
KDGAN	74.10%
Teacher(DenseNet-40)	74.23%

Table 1. Testing accuracy for training the student networks with 8 convolutional layers and 8M parameters by Baseline (typical training process), Logits Mimic Learning, KD, and KDGAN.

Model	No.Parameters	Accuracy	Inference time
8 conv-20M (KDGAN)	20.2M	74.36%	4.56ms
8 conv-28M (KDGAN)	28.1M	75.25%	5.7ms
MobileNet (KDGAN)	3.5M	77.20%	2.79ms
MobileNet(Baseline)	3.5M	72.99%	2.79ms
DenseNet-100(Teacher)	7.2M	77.94%	18.02ms

Table 8. Testing accuracy and inference time for training simple CNNs with 8 convolutional layers and 20.2M, 28.1M parameters, and MobileNet as student networks by KDGAN.