LIT: Learned Intermediate representation
Training for Model Compression

Animesh Koratana*, Daniel Kang*, Peter Bailis, Matei Zaharia
DAWN Project, Stanford InfoLab
http://dawn.cs.stanford.edu/
LIT can compress models up to 4x on CIFAR10: ResNet -> ResNet

This talk: achieving higher compression on modern deep networks
Deep networks can be compressed to reduce inference costs.

- Deep compression
- Knowledge distillation
- FitNets

These methods are largely architecture agnostic.
Modern networks have highly repetitive sections – can we compress them?
LIT: Learned Intermediate-representation Training for modern, very deep networks

LIT penalizes deviations in intermediate representations of architectures with the same width.
LIT: Learned Intermediate-representation Training for modern, very deep networks

Teacher model: ResNet-110
18 residual blocks
IR loss

Student model: ResNet-56
9 residual blocks
IR loss

Training only

Losses

IR comparison
KD comparison

IR loss

18 residual blocks

18 residual blocks

18 residual blocks

18 residual blocks

18 residual blocks

9 residual blocks

9 residual blocks

9 residual blocks

9 residual blocks

FC layer

FC layer

LIT uses the output of the teacher model’s previous section as input to the student model’s current section.
LIT can compress models up to 4x on CIFAR10: ResNet -> ResNet
LIT can compress StarGAN up to 1.8x

Student model outperforms teacher in Inception/FID score

<table>
<thead>
<tr>
<th>Model</th>
<th>Inception score (higher is better)</th>
<th>FID score (lower is better)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Teacher (18)</td>
<td>3.49</td>
<td>6.43</td>
</tr>
<tr>
<td>LIT student (10)</td>
<td>3.56</td>
<td>5.84</td>
</tr>
<tr>
<td>L2 student (10)</td>
<td>3.46</td>
<td>6.47</td>
</tr>
<tr>
<td>From scratch (10)</td>
<td>3.37</td>
<td>6.56</td>
</tr>
<tr>
<td>Rand init (10)</td>
<td>2.63</td>
<td>94.00</td>
</tr>
<tr>
<td>Rand init (18)</td>
<td>2.45</td>
<td>151.43</td>
</tr>
</tbody>
</table>
LIT can compress GANs up to 1.8x

Student model also outperforms teacher in qualitative evaluation
Conclusions

Neural networks are becoming more expensive to deploy

**LIT** is a novel technique that combines both:
1. Intermediate representations and
2. matching outputs

that improves training to give 3-5x compression for many tasks

ddkang@stanford.edu  
koratana@stanford.edu

Find our poster at Pacific Ballroom, #17!