GMV
Google Mobile Vision
iNatrualist Challenge at FGVC 2017

Yin Cui*                             Yang Song*                   Andrew Howard                    Chen Sun
*equal contribution
Challenge Results

- 25% better than the 2nd place relatively.
- 67.5% relative improvement over TensorFlow Inception-V3 baseline.
- 63.6% relative improvement over TensorFlow Inception-ResNet-V2 baseline.
Overview

- Validation set split: 90% validation → val; 10% validation → minival.
- Inception models trained on train + val (666k images), evaluated on minival (9.6k images).
- Using higher resolution image improves the performance.
- To deal with label imbalance, fine-tune on val further with small learning rate after training.
  - Training learns good feature, fine-tuning on validation gives the network information on the label distribution.
Implementation details

- Trained with open source TensorFlow.
- Fine-tuned from publicly available ImageNet-1k pre-trained models; about 1% worse if trained from scratch.
- 5089-way softmax without using supercategory information.
- Trained with asynchronous RMSProp on multiple GPUs and parameter servers.
- Inception-style data augmentation: random-sized cropping and aspect ratio.
- Label smoothing to smooth the target and capture class correlations implicitly.
- Exponential moving average for learnable parameters.
- 12 crops (center + 4 corners + whole image and its flip) for inference.
- Ensemble of two models: Inception-V4 and Inception-ResNet-V2
## Results

- Train on train + val. Validate on minival.

<table>
<thead>
<tr>
<th>Method</th>
<th>Top 1 (%)</th>
<th>Top 5 (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Inception-V3 (our implementation)</td>
<td>70.1</td>
<td>89.4</td>
</tr>
<tr>
<td>Inception-V3 with 448 input size</td>
<td>74.0 (+3.9)</td>
<td>91.2 (+1.8)</td>
</tr>
<tr>
<td>Inception-V3 ++: Fine-tuned on val with 560 input size further</td>
<td>77.3 (+3.3)</td>
<td>93.4 (+2.2)</td>
</tr>
<tr>
<td>Inception-V4 ++ (over Inception-V3 ++)</td>
<td>79.2 (+1.9)</td>
<td>94.6 (+1.2)</td>
</tr>
<tr>
<td>Inception-ResNet-V2 ++ (over Inception-V3 ++)</td>
<td>78.9 (+1.6)</td>
<td>94.4 (+1.0)</td>
</tr>
<tr>
<td>Inception-V4 ++ 12 crops (over Inception-V4 ++) *</td>
<td>80.8 (+1.9)</td>
<td>95.3 (+0.9)</td>
</tr>
<tr>
<td>Ensemble of Inception-V4 and Inception-ResNet-V2 12 crops **</td>
<td>81.9 (+1.1)</td>
<td>95.9 (+0.6)</td>
</tr>
</tbody>
</table>

*94.6% (5.3% error) on Kaggle public leaderboard.

**95.2% (4.8% error) on Kaggle public leaderboard.
Beyond iNaturalist

- How to evaluate the feature learned on iNaturalist?
  - Use the learned feature as an initialization for CUB-200-2011 fine-grained bird dataset.

<table>
<thead>
<tr>
<th>Network</th>
<th>Pre-trained dataset</th>
<th>FT last layer</th>
<th>FT all</th>
</tr>
</thead>
<tbody>
<tr>
<td>Inception-V3</td>
<td>ImageNet</td>
<td>64.36%</td>
<td>83.4%</td>
</tr>
<tr>
<td>Inception-V3</td>
<td>iNaturalist</td>
<td>90.57%</td>
<td>90.89%</td>
</tr>
<tr>
<td>Bilinear / Compact Bilinear CNN</td>
<td></td>
<td>N/A</td>
<td>84.1%</td>
</tr>
<tr>
<td>[Jonathan Krause et al.] Web (filtered)</td>
<td>N/A</td>
<td></td>
<td>89.0%</td>
</tr>
<tr>
<td>[Jonathan Krause et al.] L-Bird + CUB-GT</td>
<td>N/A</td>
<td></td>
<td>92.2%</td>
</tr>
</tbody>
</table>
Summary

● Powerful data (iNaturalist) produce strong features.
● Details are important:
  ○ strong data augmentation, label smoothing, exponential moving average, etc.
● Finer resolution for finer discrimination:
  ○ higher resolution induces better feature.
● Dealing with label imbalance:
  ○ fine-tune on validation set with low learning rate.

● All details and other findings not included in the talk will be summarized into an arXiv tech report. Please stay tuned.
● We plan to release code, models, etc.
Acknowledgements

- Google Mobile Vision Team
- iNaturalist community
- FGVC organizers