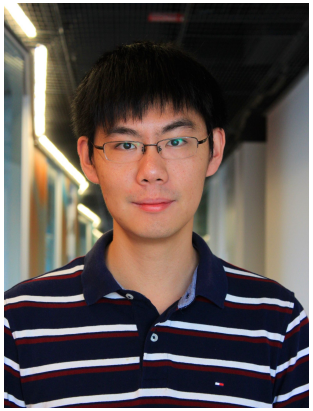
Research at Google

GMV

Google Mobile Vision

iNatrualist Challenge at FGVC 2017



Yin Cui*

Yang Song*

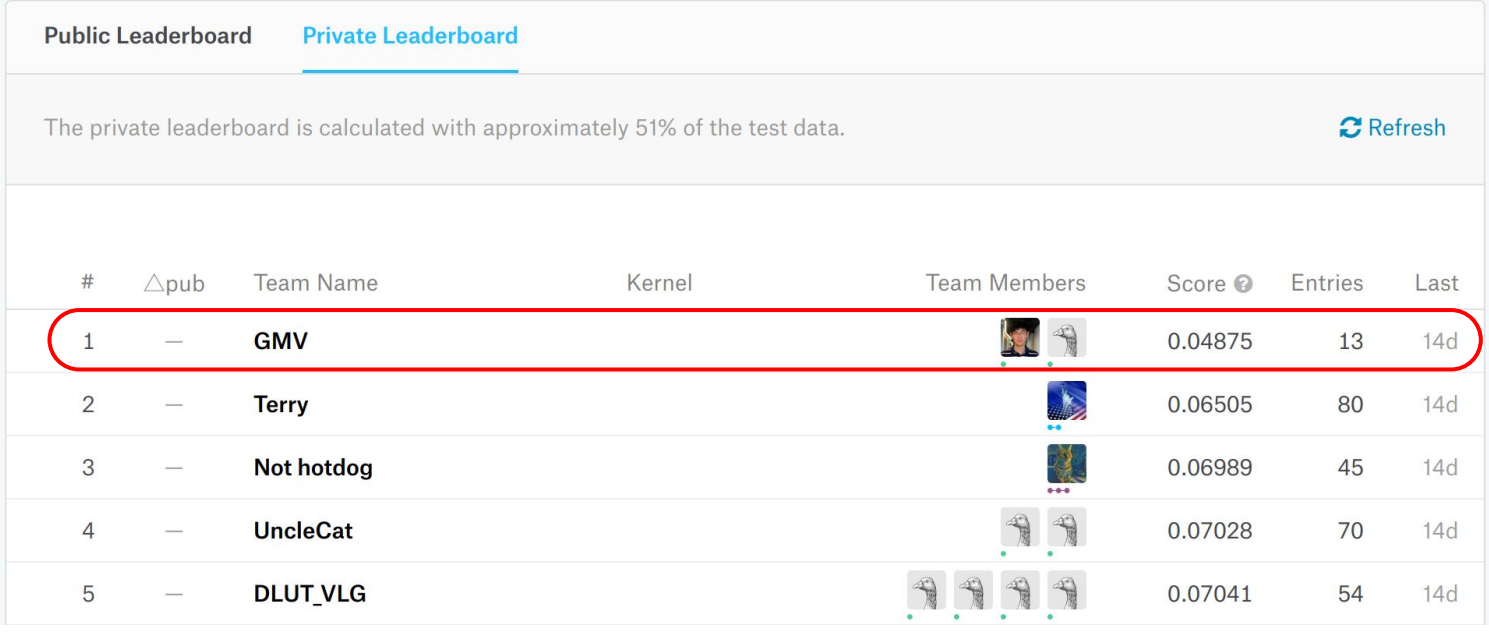
Andrew Howard

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.



Public Leaderboard | Private Leaderboard

The private leaderboard is calculated with approximately 51% of the test data. Refresh

#	△pub	Team Name	Kernel	Team Members	Score	Entries	Last
1	—	**GMV**			0.04875	13	14d
2	—	**Terry**			0.06505	80	14d
3	—	**Not hotdog**			0.06989	45	14d
4	—	**UncleCat**			0.07028	70	14d
5	—	**DLUT_VLG**			0.07041	54	14d

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.

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

* 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.

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

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