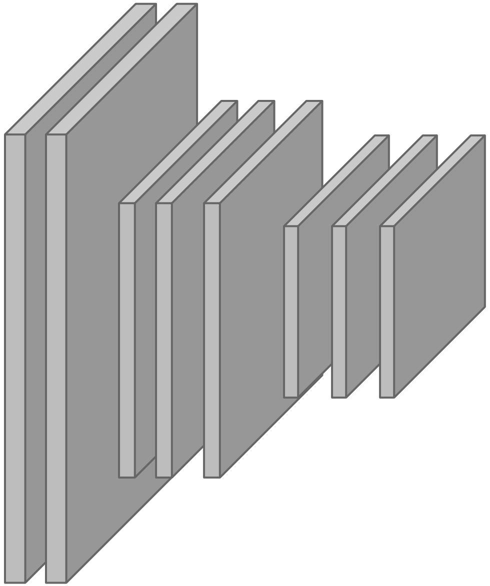
Learning from Web-scale Image Data For Visual Recognition

Chen Sun
Google Research

Recipe for Success (in Deep Learning era)

Models



Labeled Data

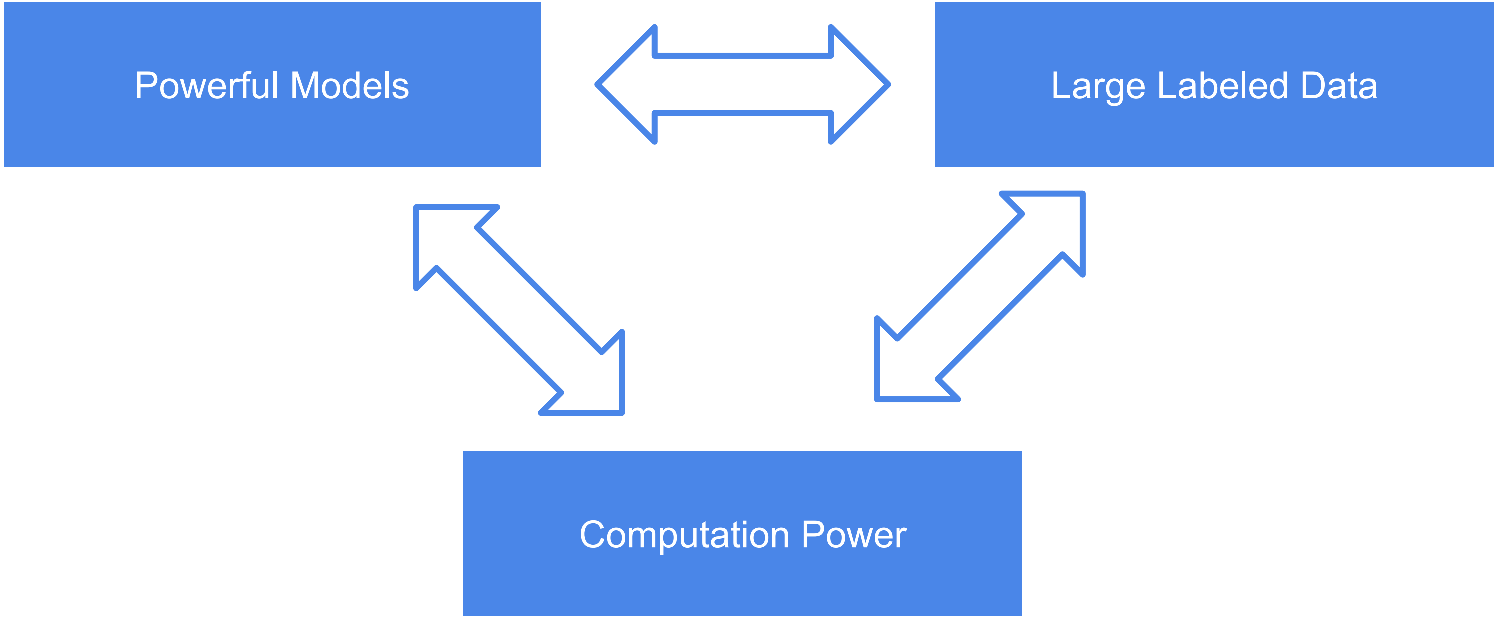
SUN

IMAGENET

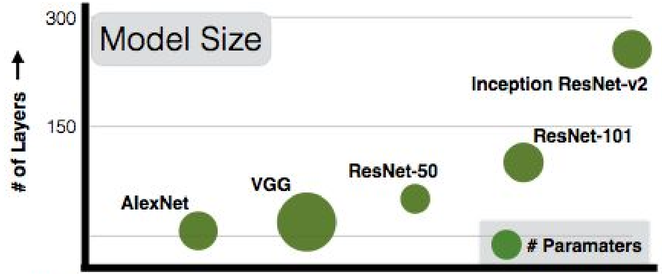
PASCAL2
Pattern Analysis, Statistical Modelling and Computational Learning

COCO
Common Objects in Context

Recipe for Success (in Deep Learning era)



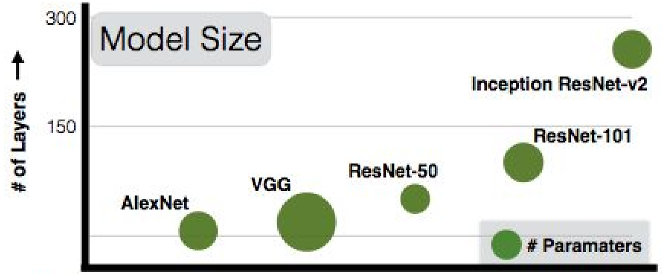
Recipe for Success (in Deep Learning era)



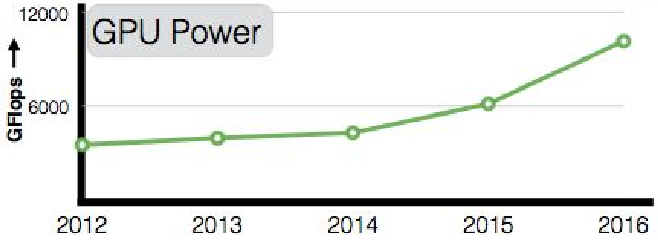
Large Labeled Data

Computation Power

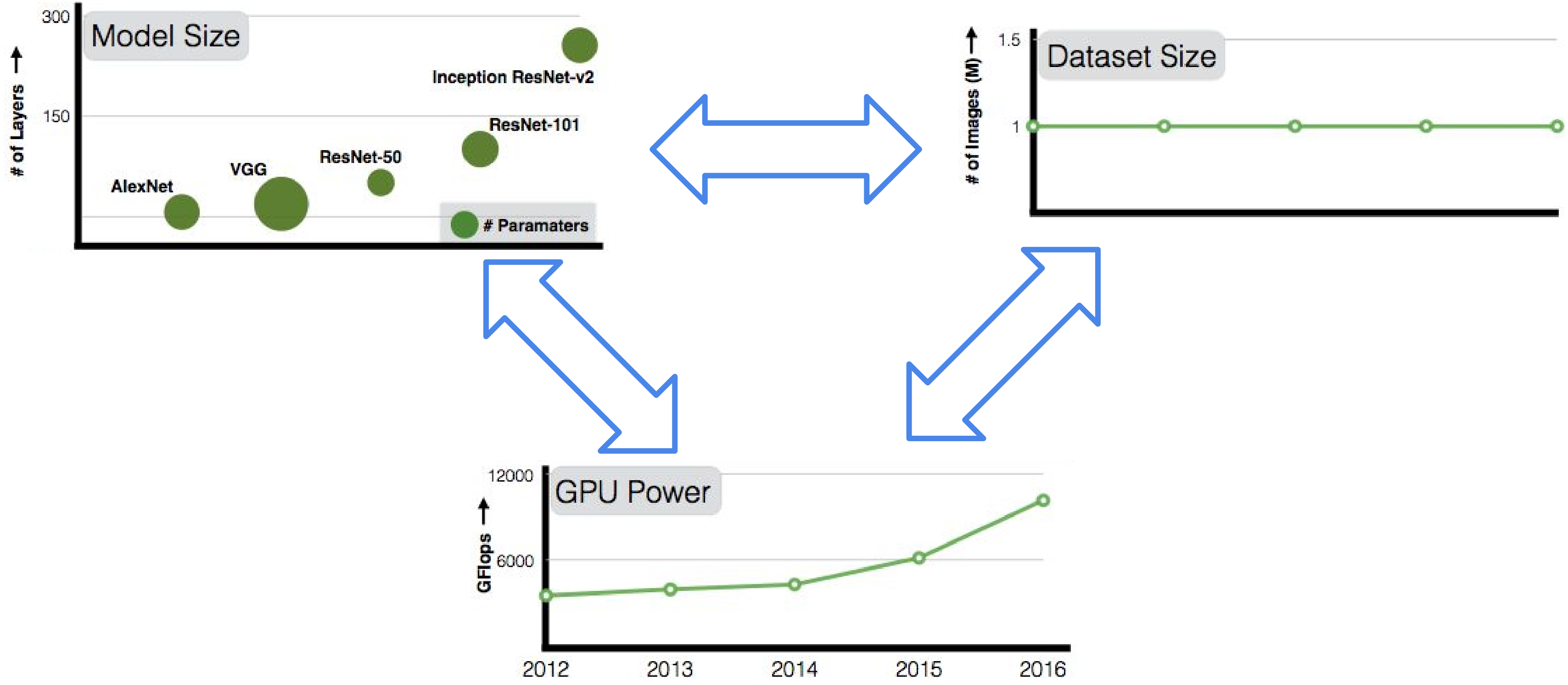
Recipe for Success (in Deep Learning era)



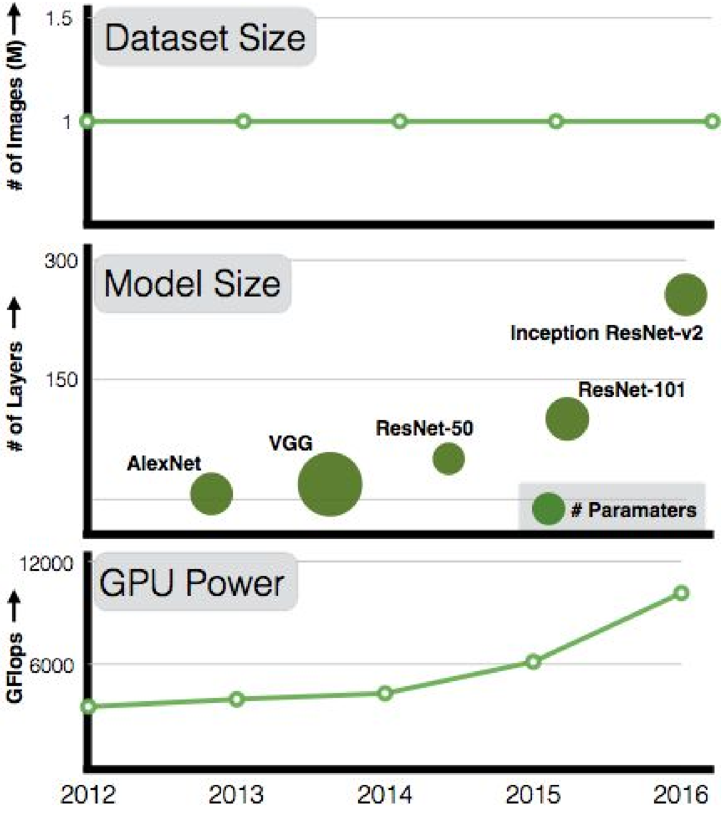
Large Labeled Data



Recipe for Success (in Deep Learning era)



Curious Case of Vision Datasets



- What happens at 300x scale of ImageNet?
- How big is big? (Plateauing effect?)
- Data Size v.s. Model size

Revisiting Unreasonable Effectiveness of Data in Deep Learning Era

Joint work with Abhinav Shrivastava, Saurabh Singh and Abhinav Gupta

ICCV 2017 (arXiv)

JFT-300M Dataset

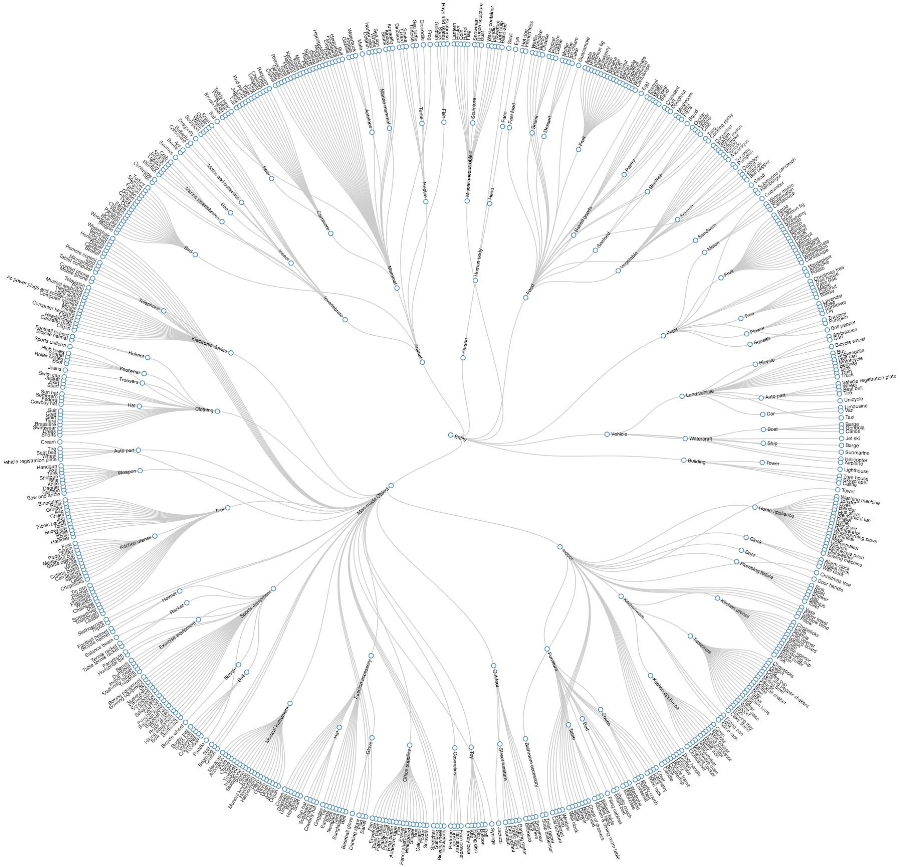
- 300M web images
- 375M image label pairs

Previous publications on JFT:

- F. Chollet, Xception: Deep learning with depthwise separable convolutions. CVPR 2017
- G. Hinton, O. Vinyals, and J. Dean. Distilling the knowledge in a neural network. NIPS 2014.

JFT-300M Dataset

- 300M web images
- 375M image label pairs
- ~ 19K categories



JFT-300M Dataset

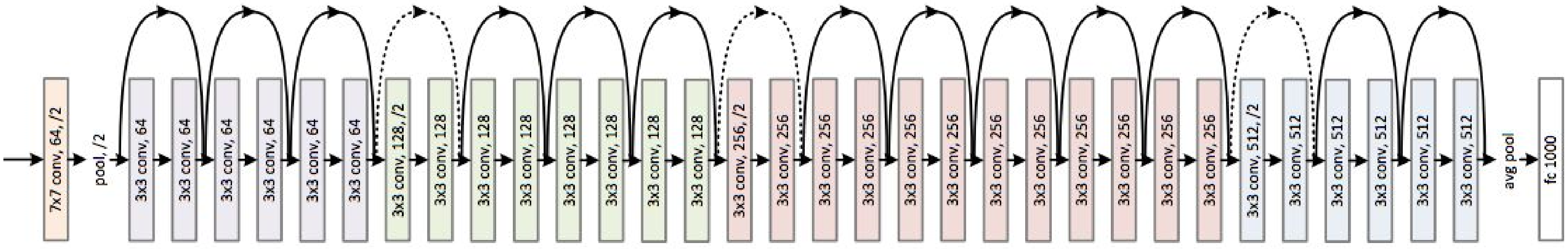
- 300M web images
- 375M image label pairs
- ~ 19K categories
- ~ 20% label noise
- Unknown recall
- Long-tail distribution

Tortoise:

V.S.

Training on JFT-300M

- Deep residual networks (ResNet-50 / 101 / 152)

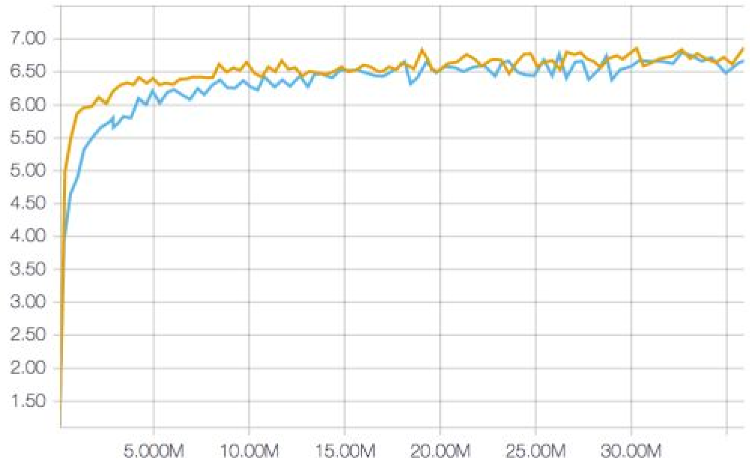


Visualization of a 34-layer ResNet

K. He, X. Zhang, S. Ren and J. Sun, Deep Residual Learning for Image Recognition, CVPR 2016.

Training on JFT-300M

- Deep residual networks (ResNet-50 / 101 / 152)
- 50 K80 GPUs for 1.5 months
- 4 epochs (ImageNet is trained for 100 epochs)
- Async SGD



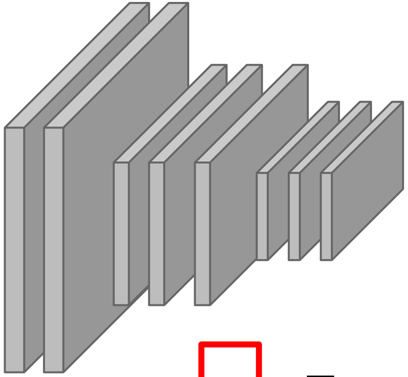
Empirical Study of JFT-300M Models

- Transfer the learned representations
  - Avoid potential bias of JFT-300M validation set
  - Common benchmark as ImageNet

Related work:
M. Oquab, L. Bottou, I. Laptev, and J. Sivic. Learning and transferring mid-level image representations using convolutional neural networks. In CVPR, 2014.

Transfer the Learned Representations

JFT 300M

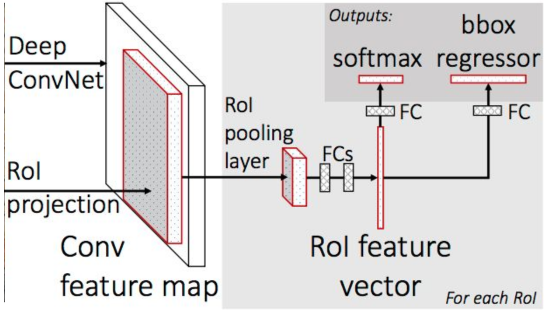


18K labels

Transfer weights

PASCAL2
Pattern Analysis, Statistical Modelling and Computational Learning

COCO
Common Objects in Context

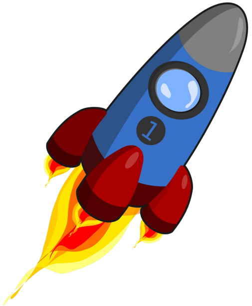


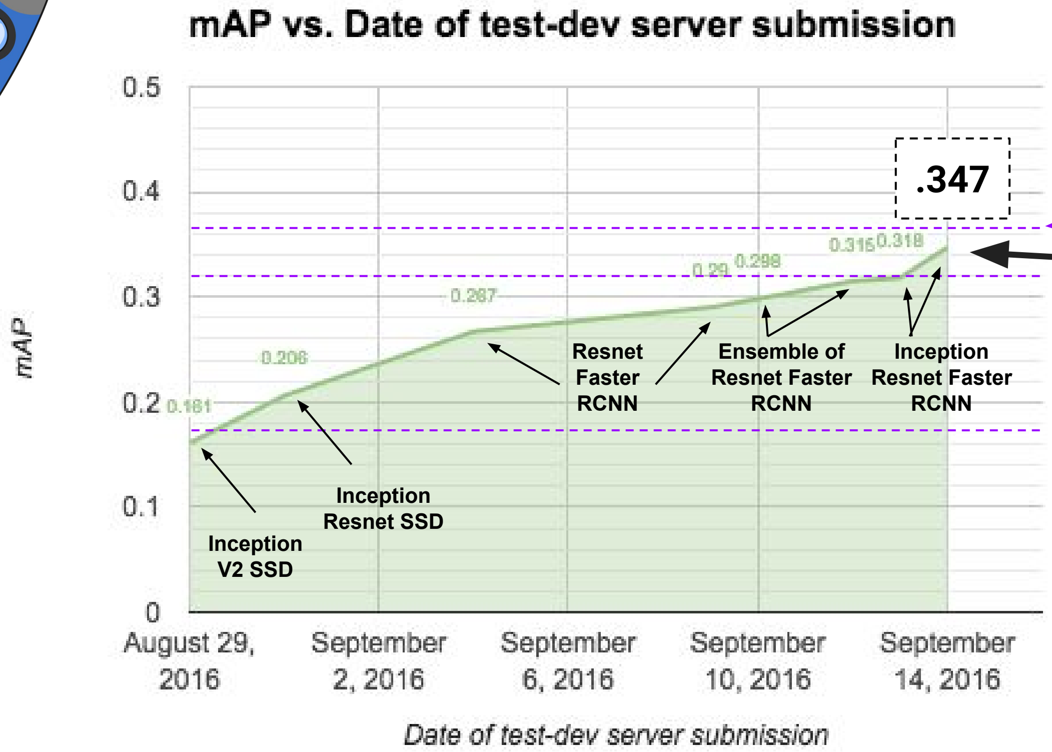
Detections

Empirical Study of JFT-300M Models

- Transfer the learned representations
  - Avoid potential bias of JFT-300M validation set
  - Common benchmark as ImageNet
- Verified on:
  - Object detection, semantic segmentation, human pose estimation
  - Frozen feature bottom v.s. Fine-tuning all layers

Better Representation Learning Helps!





Huang et al., Speed/accuracy trade-offs for modern convolutional object detectors. CVPR 2017.

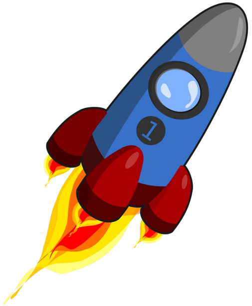
37.4% by MSRA (the best from 2015 leaderboard)

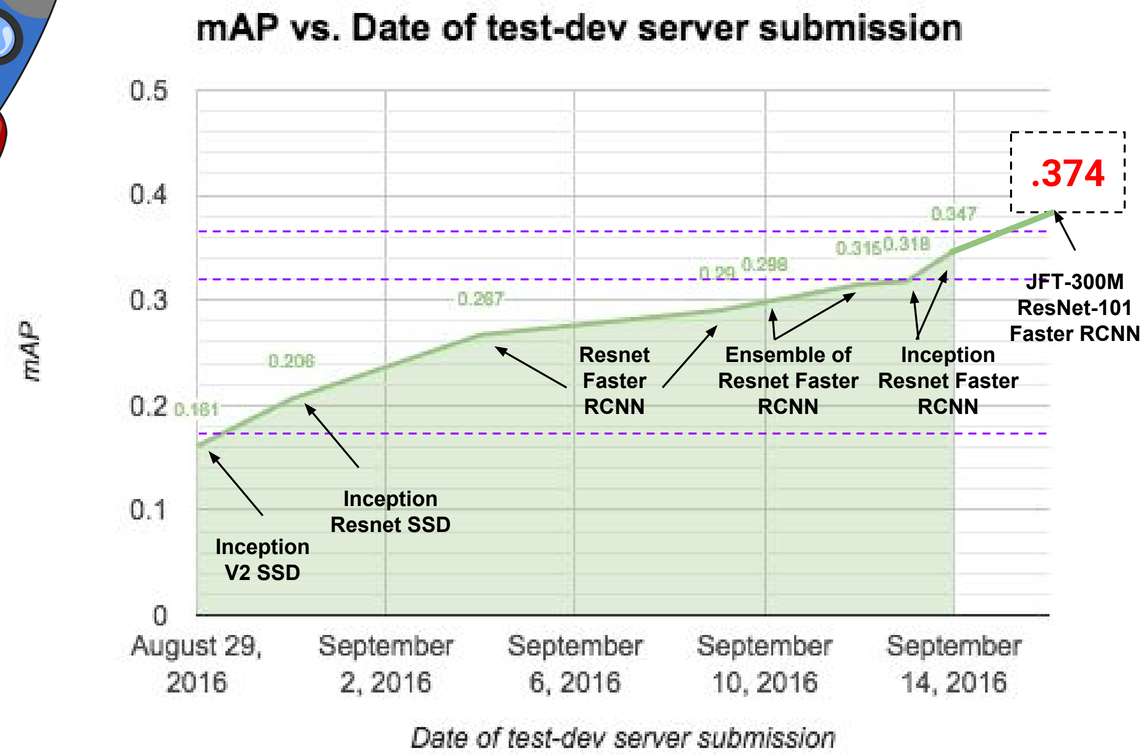
34.7: Our best single model performance before ensembling/multicrop

Best single model performance reported in literature that does not do multiscale or multicrop

(Last place from 2015 leaderboard)

Better Representation Learning Helps!





Using a JFT-300M pre-trained checkpoint to replace ImageNet ones:

- 2.7% gain over best single model
- 3.1% gain over comparable ResNet model

Better Representation Learning Helps!

Absolute gains over ImageNet pre-training:

- 2% ImageNet top-1 classification accuracy

Initialization	Top-1 Acc.	Top-5 Acc.
MSRA checkpoint [16]	76.4	92.9
Random initialization	77.5	93.9
Fine-tune from JFT-300M	**79.2**	**94.7**

Better Representation Learning Helps!

Absolute gains over ImageNet pre-training:

- 2% ImageNet top-1 classification accuracy
- 3.1% mAP COCO object detection

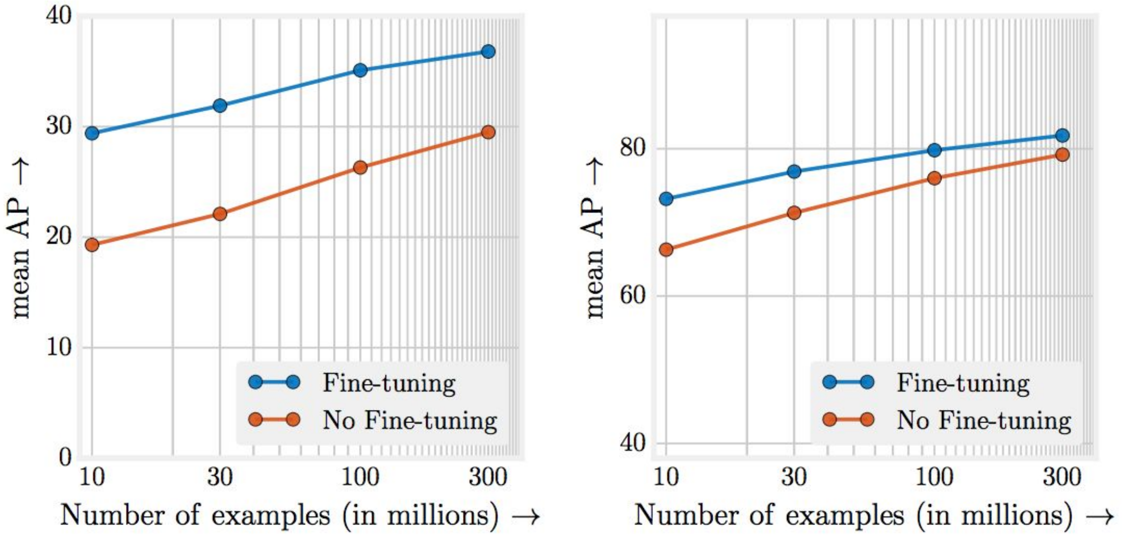
Method	mAP@0.5	mAP@[0.5,0.95]
He *et al.* [16]	53.3	32.2
ImageNet	53.6	34.3
300M	56.9	36.7
ImageNet+300M	**58.0**	**37.4**
Inception ResNet [37]	56.3	35.5

Better Representation Learning Helps!

Absolute gains over ImageNet pre-training:

- 2% ImageNet top-1 classification accuracy
- 3.1% mAP COCO object detection
- 4.8% mAP (50% IOU) VOC 07 object detection
- 3% mIOU VOC 12 segmentation
- 2% AP COCO keypoint detection

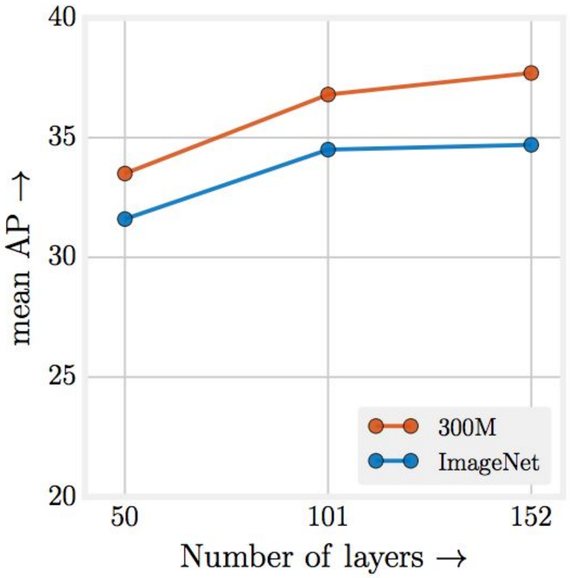
Performance v.s. Data Size



- Log-linear with number of training images
- No saturation even at 300M scale

Performance v.s. Depth

#Layers	ImageNet	300M
50	31.6	33.5
101	34.5	36.8
152	34.7	37.7



- Deeper models are better with more data

Comparison with Previous Work

- Oquab et al. showed that careful selection is needed when using more ImageNet images for training.
    - Manual selection is not needed on JFT-300M
- Joulin et al. found saturation effect at 100M scale.
    - Only uses Flickr images.
    - Shallower model: AlexNet (v.s. ResNet)

M. Oquab, L. Bottou, I. Laptev, and J. Sivic. Learning and transferring mid-level image representations using convolutional neural networks. In CVPR, 2014.
Armand Joulin, Laurens van der Maaten, Allan Jabri, Nicolas Vasilache. Learning visual features from large weakly supervised data. In ECCV, 2016.
M. Huh, P. Agrawal, and A. A. Efros. What makes imagenet good for transfer learning? arXiv:1608.08614

Just Memorizing All Test Images?

- Deduplication between JFT-300M and target test data
- 10% overlap with ImageNet validation, 4% overlap with Pascal VOC test

Just Memorizing All Test Images?

- Deduplication between JFT-300M and target test data
- 10% overlap with ImageNet validation, 4% overlap with Pascal VOC test
- **No significant change** after removing the duplicates during evaluation
- Fun fact: 1.8% overlap between ImageNet training and validation

Rethinking the principles for CNN design

- Novel architectures at 300M scale
  - Deeper models perform better on JFT-300M
  - Deeper or wider?
- Our results show the lower bound for JFT-300M's power
  - Architectures were designed for ImageNet
  - Hyperparameter search is limited

F. Chollet, Xception: Deep learning with depthwise separable convolutions. CVPR 2017

Take home messages

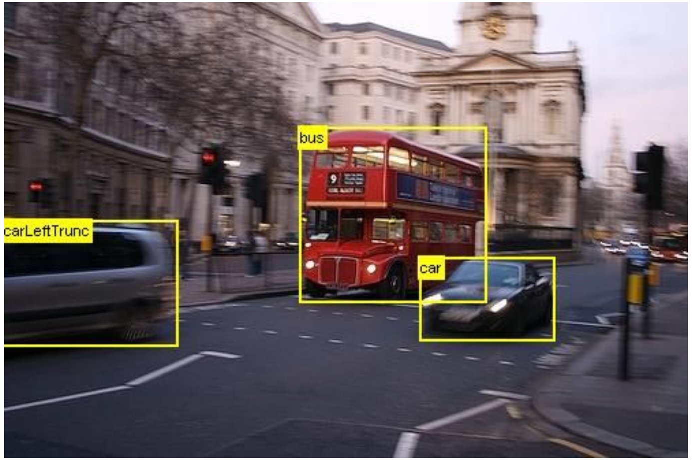
- Representation learning helps
- Performance grows log-linearly with the number of training images
- Deeper models are needed to fully utilize large-scale data

Next steps

- Further expanding the size of training data
  - 1 billion images?
- Unsupervised and semi-supervised training
- Generic representation v.s. Task specific
  - Plateauing effect for task-specific data or not?
  - Task-specific data is more difficult to obtain

Task-specific Data

COCO Dataset



VOC Dataset

Citiscape Dataset

Domain-specific (Web) Data

Figure credit: X. Chen, A. Shrivastava and A. Gupta, Enriching Visual Knowledge Bases via Object Discovery and Segmentation. In CVPR 2014.

Task-specific v.s. Domain-specific (Web)

- Task-specific data
  - Full supervision
  - Smaller scale
- Web data
  - Weak supervision
  - Large scale
  - Domain bias

Web Constraints Make Localization Easier!

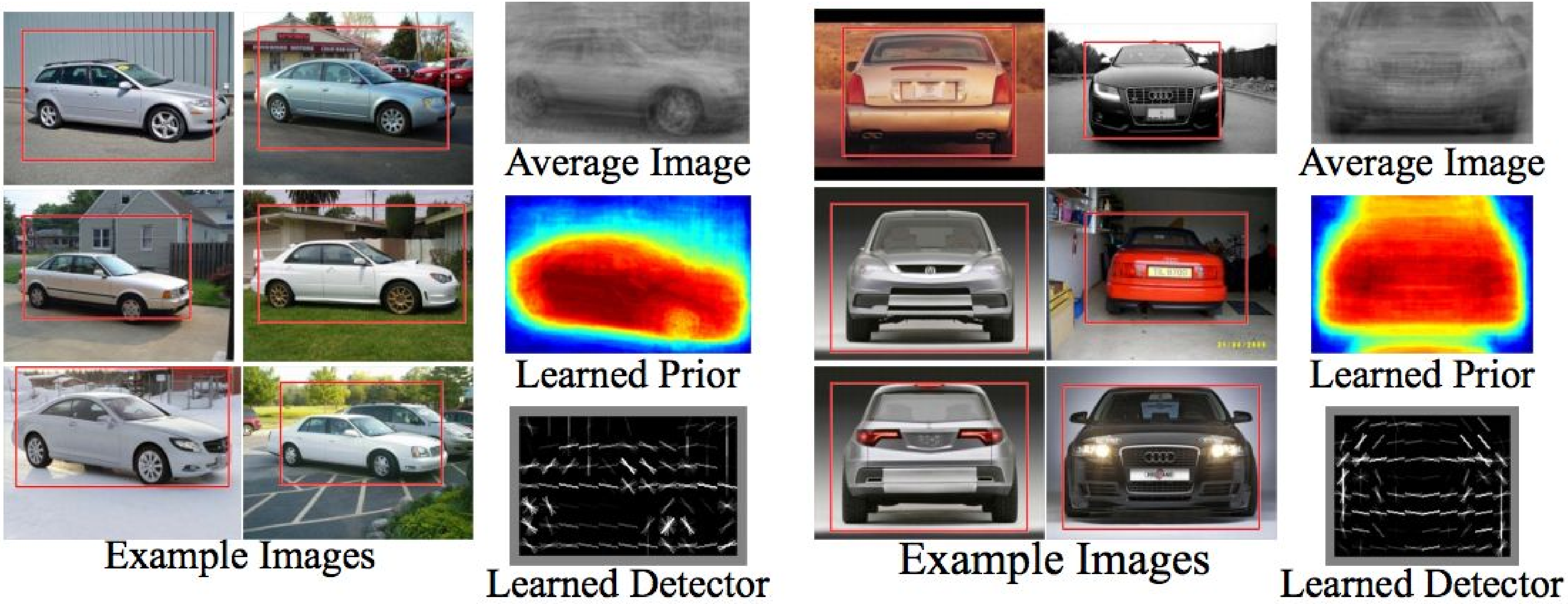


Figure credit: X. Chen, A. Shrivastava and A. Gupta, Enriching Visual Knowledge Bases via Object Discovery and Segmentation. In CVPR 2014.

Weakly-supervised Object Detection

TV, clock, book, scissors, couch, telephone, cup

Bicycle, umbrella, car, person, motorcycle

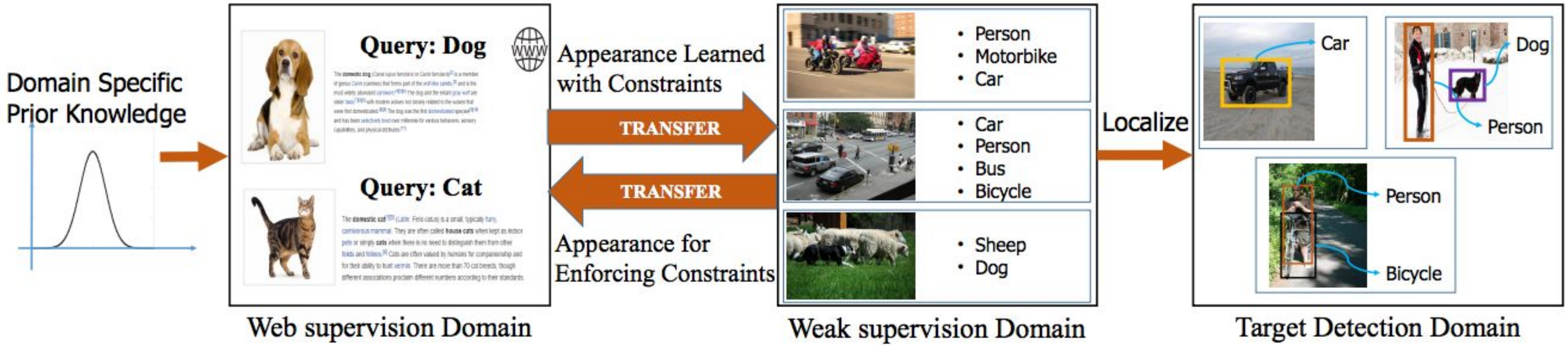
Horse, teddy bear

Donut, pizza

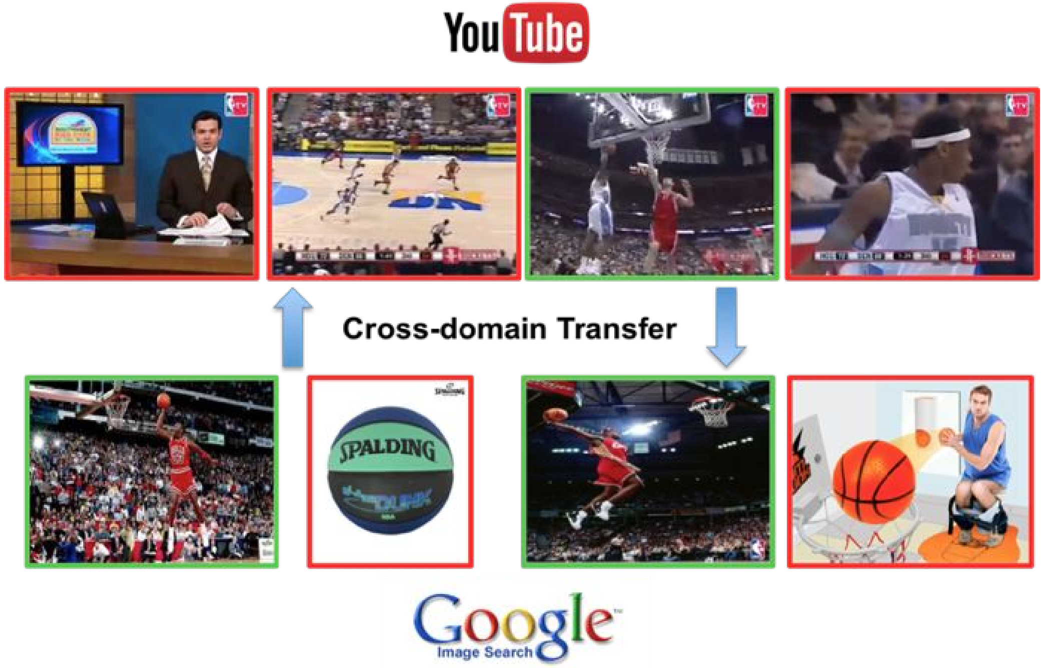
Weakly supervised object detection (WSOD):
Learn to localize objects (bounding boxes) using image-level labels

Constraint-transfer for Weakly Supervised Object Detection

Joint work with Senthil Purushwalkam and Abhinav Gupta



Domain Transfer Between (Web) Images and Videos



Temporal localization of fine grained actions in videos by domain transfer from web images.

ACM Multimedia 2015

Joint work with Sanketh Shetty, Rahul Sukthankar and Ram Nevatia.

Temporal Localization of Actions

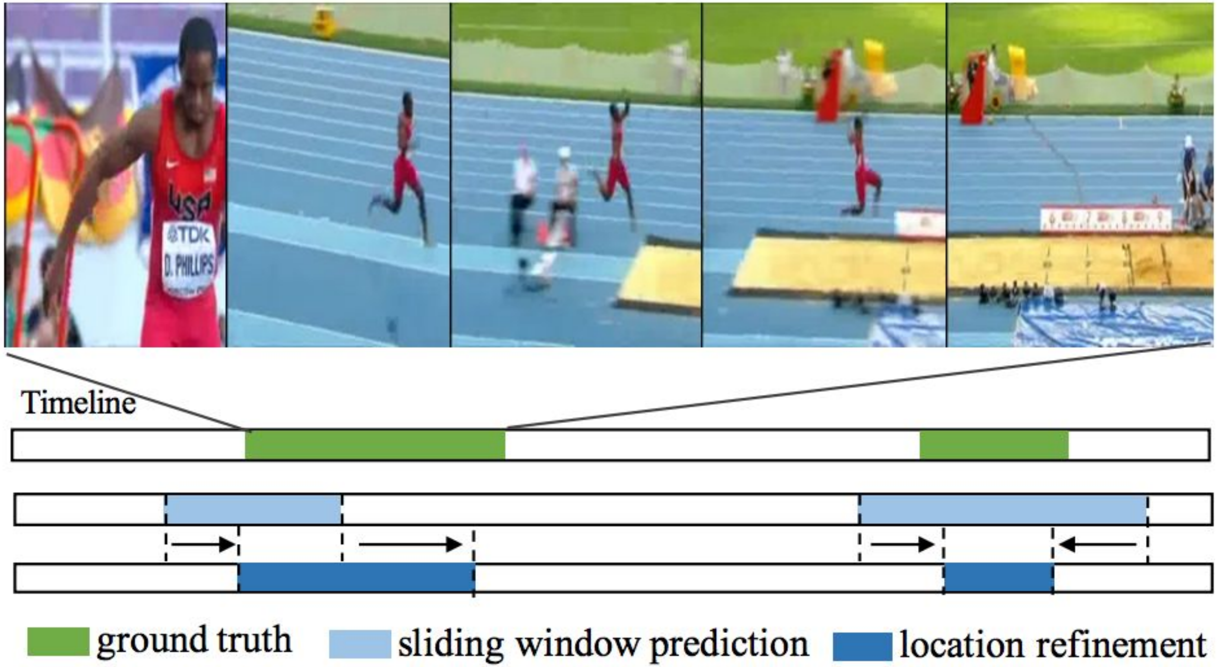


Figure credit: Gao et al., TURN TAP: Temporal Unit Regression Network for Temporal Action Proposals. In ICCV 2017.

Weakly-supervised Temporal Localization

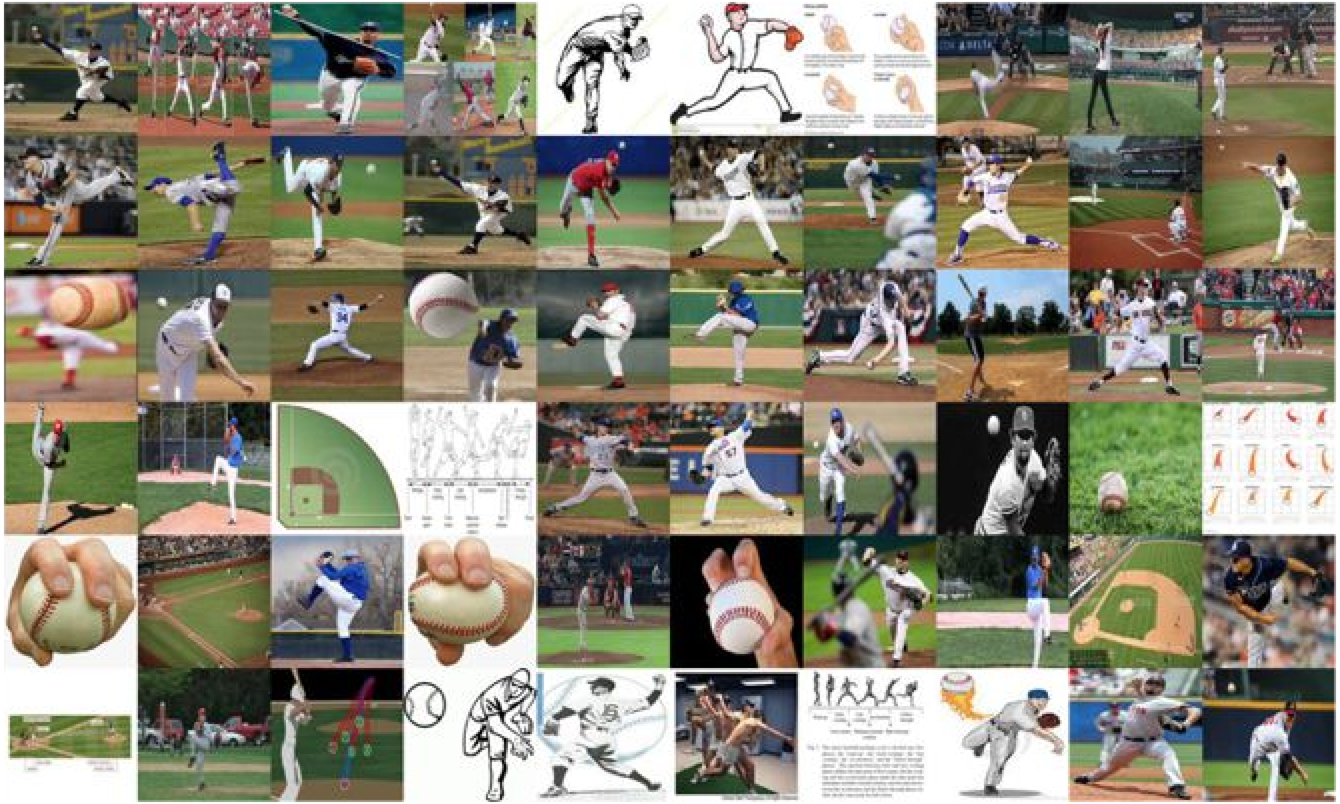
- A video typically contains multiple instances of different actions
- Only video-level labels are known, not temporal boundaries are given
- For sports, many “fine-grained” actions with similar background

Baseball

+

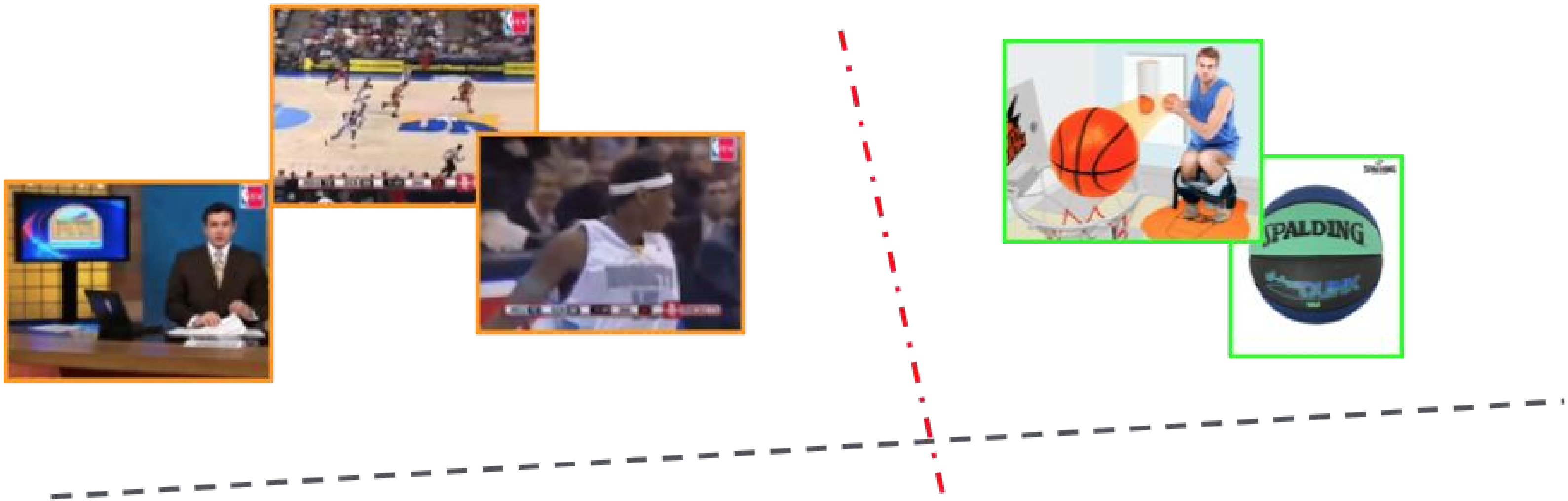
Pitch

Google

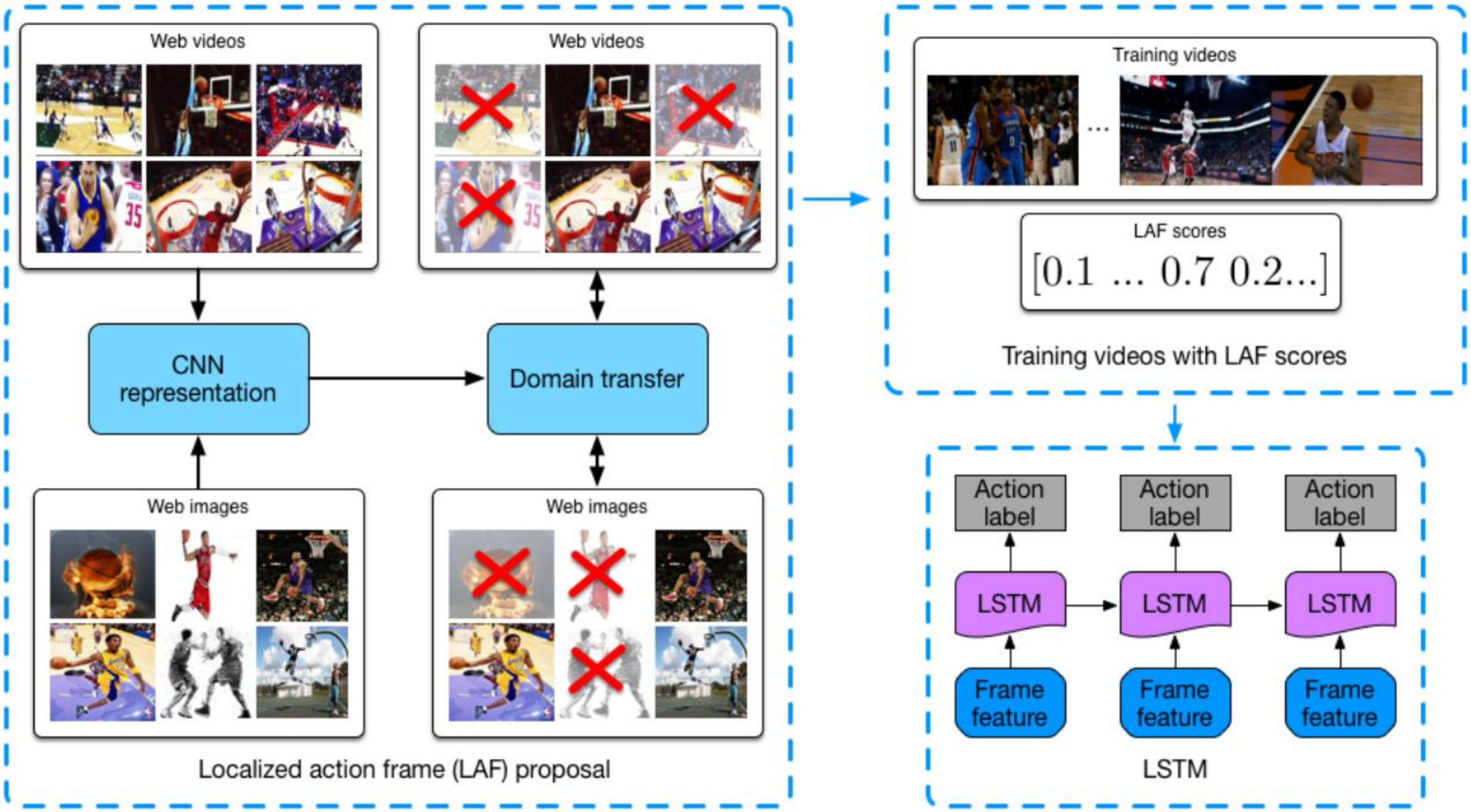



Localized action highlights
Irrelevant images

Assumption 1:
Video frames and web images which correspond to the action are visually similar



Assumption 2:
Distributions of non-action frames and web images are usually very different


Web videos
Web videos
CNN representation
Domain transfer
Web images
Web images
Localized action frame (LAF) proposal
Training videos
...
LAF scores
[0.1 ... 0.7 0.2...]
Training videos with LAF scores
Action label
Action label
Action label
LSTM
LSTM
LSTM
Frame feature
Frame feature
Frame feature
LSTM

Mutual Voting between Images and Video Frames

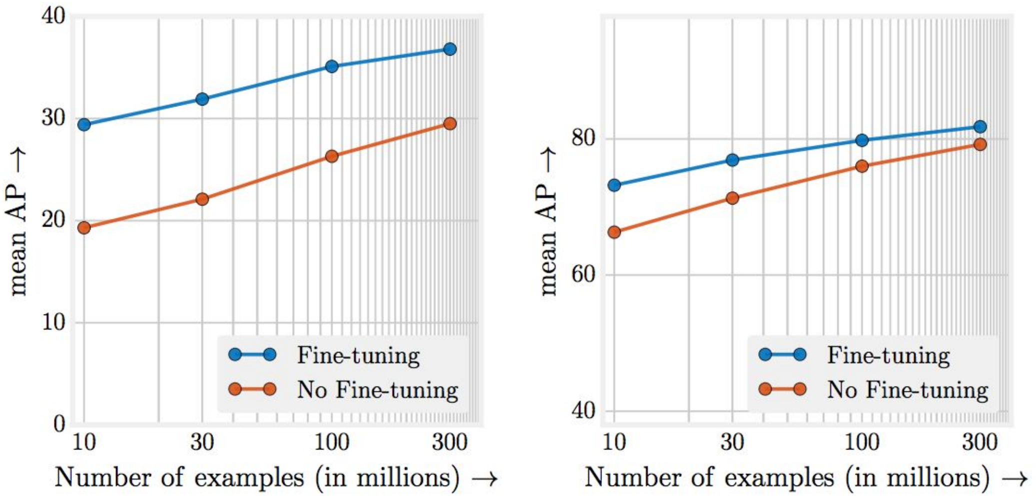
Webly-supervised Video Recognition by Mutually Voting for Relevant Web Images and Web Video Frames.

ECCV 2016

Joint work with Chuang Gan, Lixin Duan and Boqing Gong.

Three Ways to Use Web-scale Images

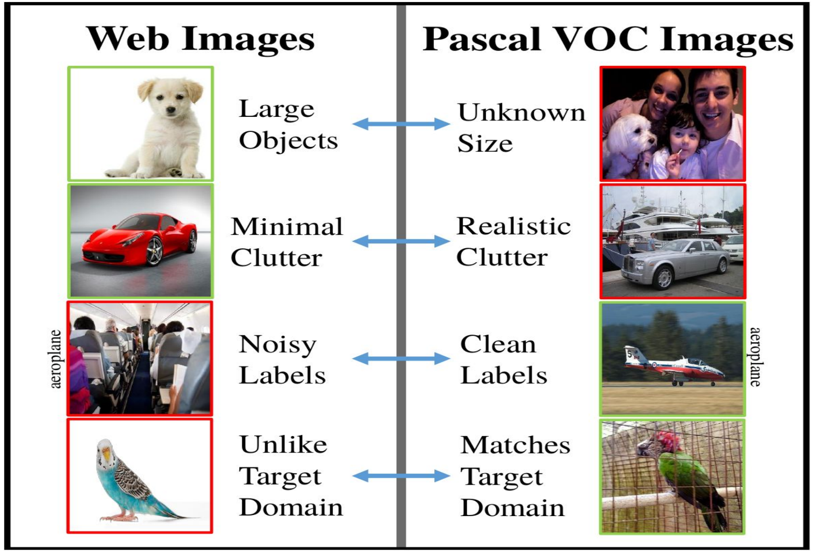
Representation Learning



Three Ways to Use Web-scale Images

Representation Learning

Cross-domain Constraints

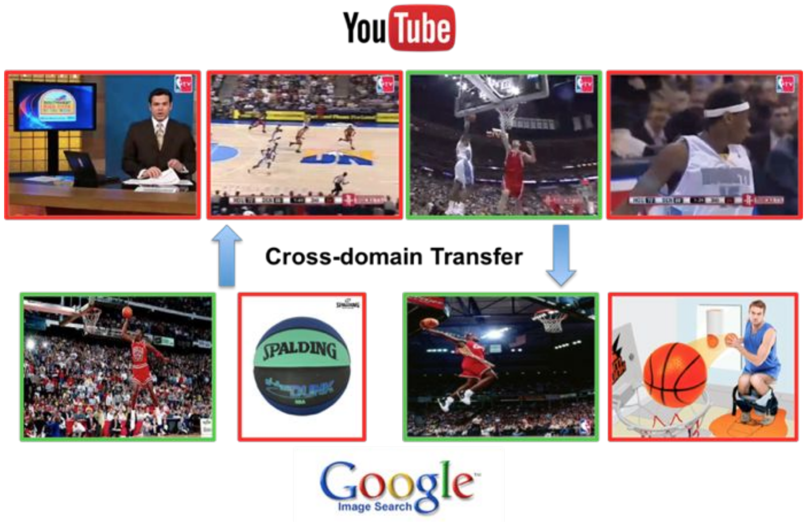


Three Ways to Use Web-scale Images

Representation Learning

Cross-domain Constraints

Cross-modal Constraints



Conclusions

- Web-scale images (300M) help visual representation learning
- Novel architectures should be explored to handle web-scale data
- Domain-specific web images provide useful constraints for weakly-supervised learning