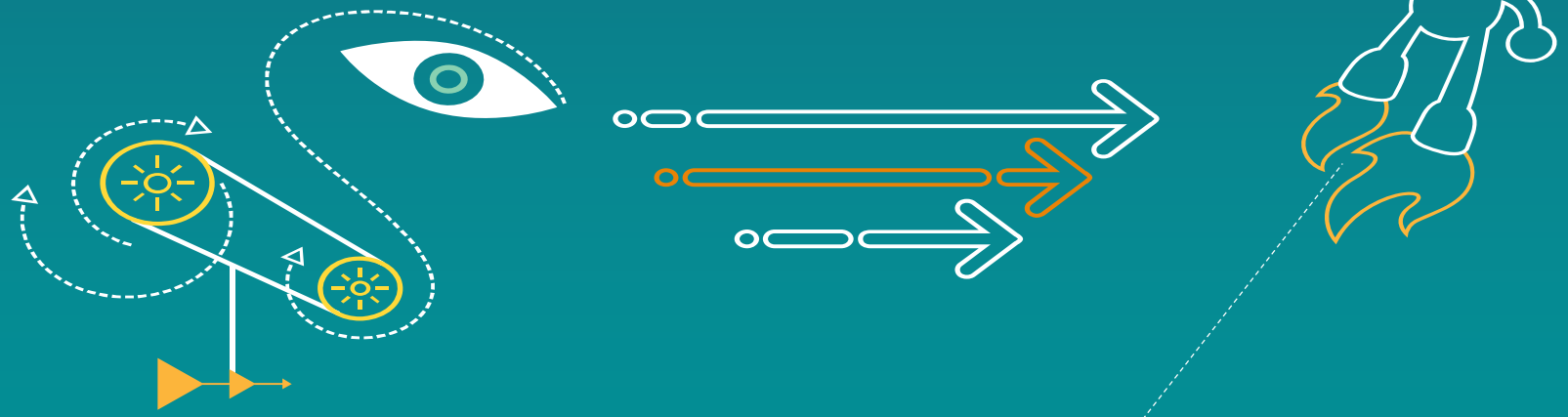
Daniel Fontijne (Engineer, Senior Staff, QTI), **David Julian** (Engineer, Principal, QTI), **Koen E. A. van de Sande** (Engineer, Staff, QTI), **Anthony Sarah** (Engineer, Sr. Staff/Manager, QTI), **Harro Stokman** (Director, Product Management, QTI), **R. Blythe Towal** (Engineer, Staff, QTI), **Cees G. M. Snoek** (Engineer, Principal, QTI)

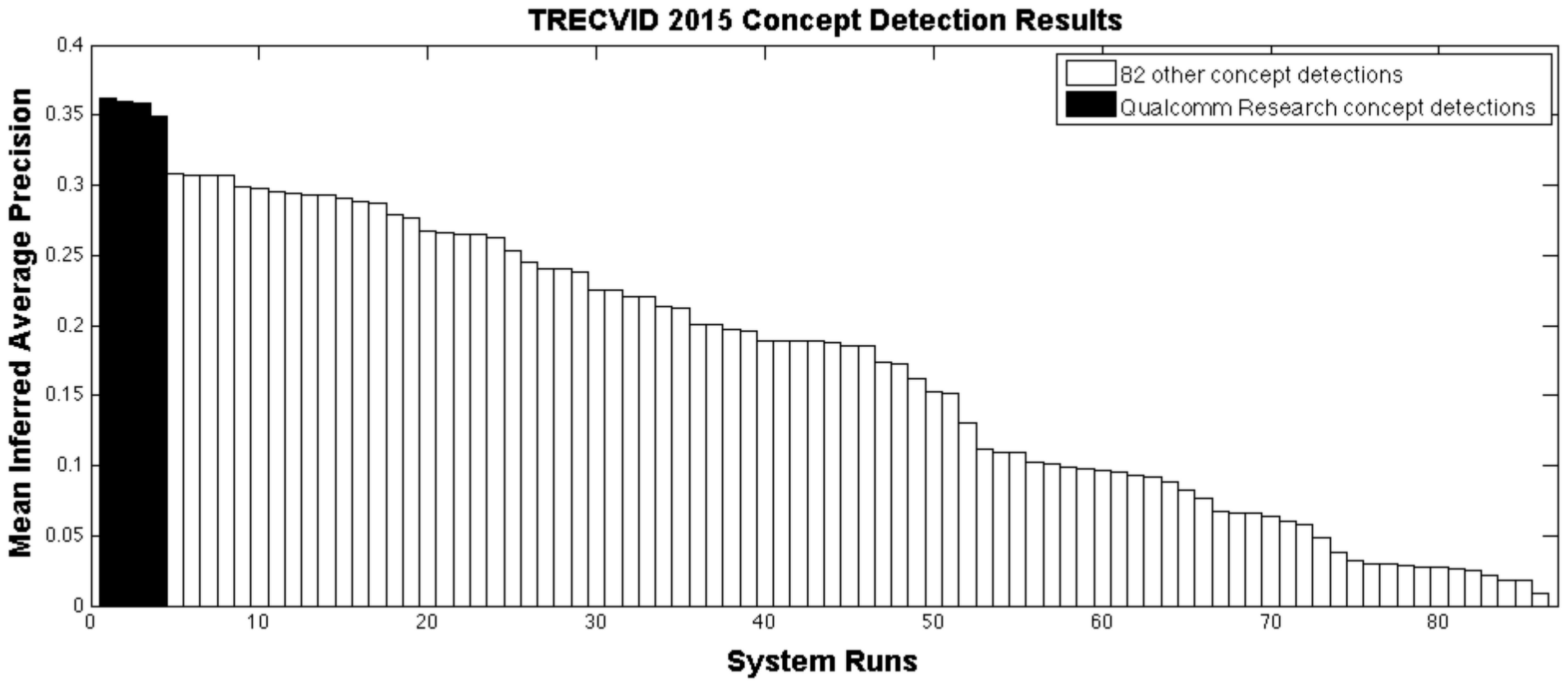
Qualcomm Research Deep Net for Video Concept Detection

QUALCOMM TECHNOLOGIES, INC.

November 16, 2015



Summary



The Qualcomm Research system is deep learning only

Inspiration from ImageNet

Very deep convolutional neural networks

Inception

Small 1x1 convolutions

Convolution stride of two or one

ReLU non-linearity

Four max-pool layers

One fully connected layer

Dropout

Nine inception modules

Batch normalization

Szgedy et al. CVPR 2015

VGGNet

Small 3x3 convolutions

Convolution stride of one

ReLU non-linearity

Five max-pool layers

Three fully-connected layers

Dropout

Simonyan & Zisserman. ICLR 2015

Batch normalization

$$\mu_{\mathcal{B}} \leftarrow \frac{1}{m}\sum_{i=1}^{m} x_i \qquad \text{// mini-batch mean}$$

$$\sigma_{\mathcal{B}}^2 \leftarrow \frac{1}{m}\sum_{i=1}^{m}(x_i - \mu_{\mathcal{B}})^2 \qquad \text{// mini-batch variance}$$

$$\widehat{x}_i \leftarrow \frac{x_i - \mu_{\mathcal{B}}}{\sqrt{\sigma_{\mathcal{B}}^2 + \epsilon}} \qquad \text{// normalize}$$

$$y_i \leftarrow \gamma \widehat{x}_i + \beta \equiv \mathrm{BN}_{\gamma,\beta}(x_i) \qquad \text{// scale and shift}$$

- Address covariate shift per layer
- Normalize the activations in each layer within a mini-batch
- Learn the mean and variance of each layer as parameters
- Multi-layer CNN's train faster with fewer data samples
- Employ faster learning rates and less network regularizations.
- Achieves state-of-the-art on ImageNet, post-competition

Approach

High-level overview

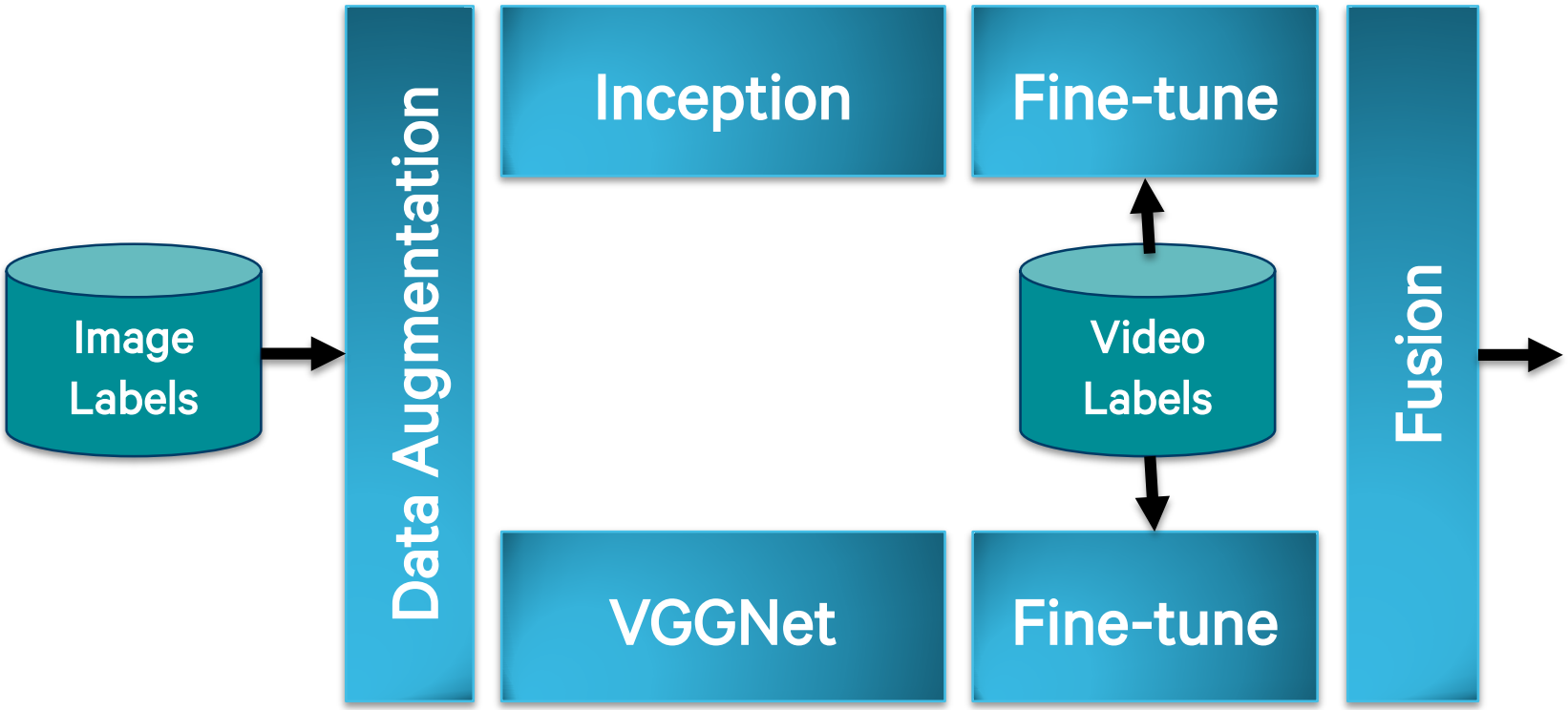


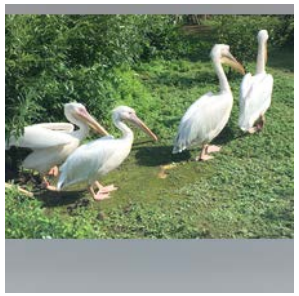
Image labels

- All models are pre-trained on ImageNet
  - 1,000 standard ImageNet categories
  - 1,024 categories better matching the video concepts
  - 2,048 same as above, plus 1,024 random categories
  - 4,096 same as above, plus more random categories

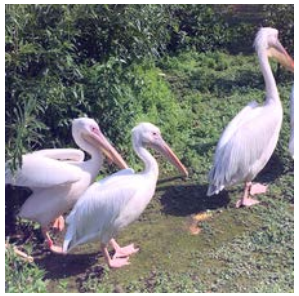
Data augmentation

Adding color casting and vignetting to default translation and mirroring

Original



Translate/Mirroring



Color casting

Vignetting

All augmentations

Fine-tune

- Inception networks typically have an average pooling on top, making them less suited for domain transfer
  - We add an ‘Alex-style’ fully connected head on the one-but-last layer

- We fine tune the fully connected layers with video labels
  - For both VGGNet and Inception

Video labels

- Common annotation effort finished in 2013 Ayache & Quénot, ECIR 2008

- Deep learning profits from more labeled data
  - Relied on Euvision annotations from 2014
  - Hired annotators to correct and supplement

Fusion

- Our models exploit diversity in
  - Networks
  - Image labels
  - Augmentations
  - Video labels

- We have a total of 63 models available for fusion
  - Non-weighted late fusion
  - Weighted late fusion

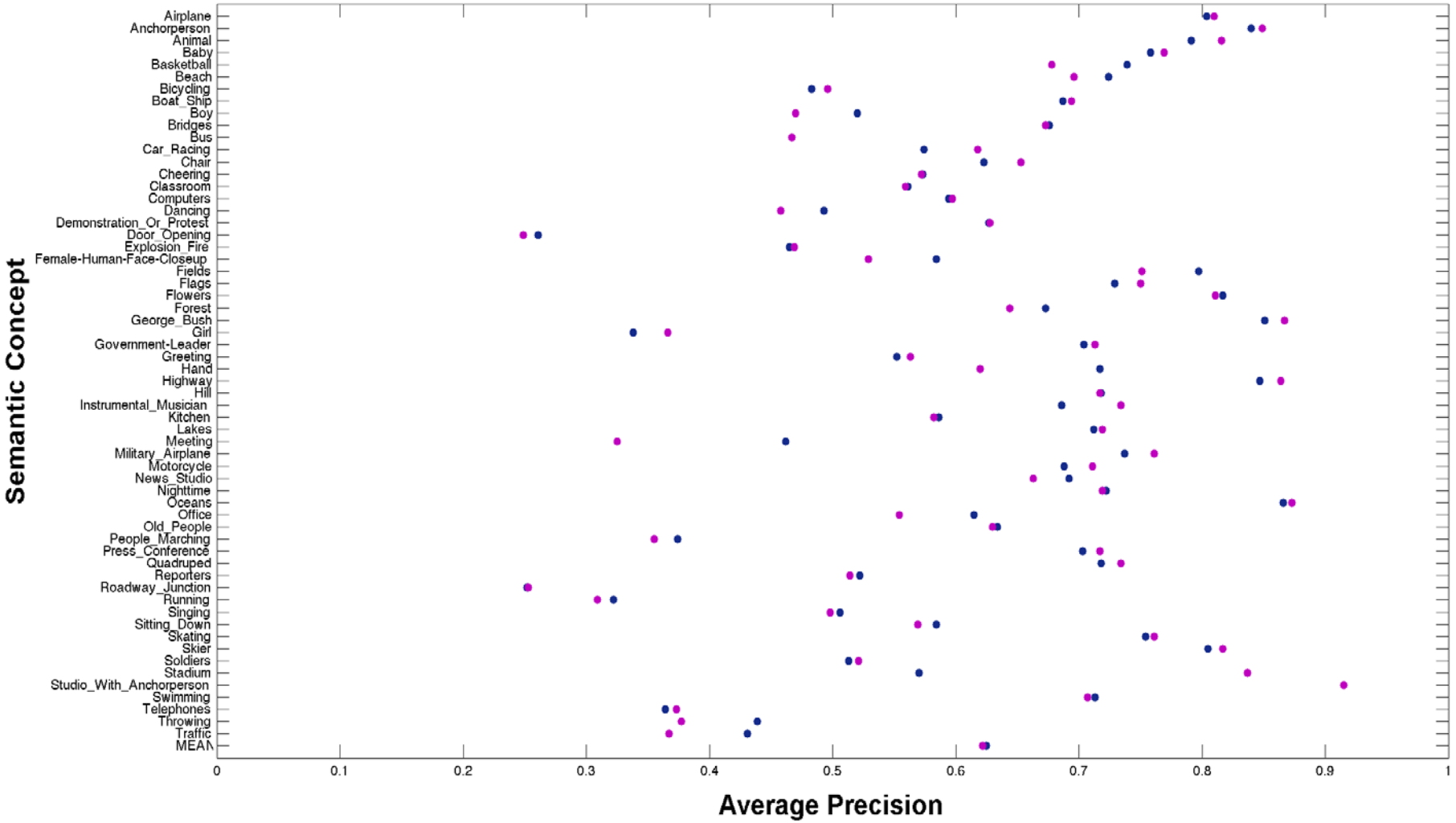
Experiments

Internal validation set

- Training set
  - 2012devel
  - 2013test
  - 2014test
- Validation set
  - 2012test

MediaMill TRECVID 2014 Baselines	mAP
Single deep network	56.0
Seven deep networks	58.0
Seven deep networks, plus color Fisher vector	60.0

Value of annotations



- 2014 video labels
- 2015 video labels

Additional annotations do not necessarily improve the detection

Value of image labels

Pre-training for single inception model	mAP
1,000 ImageNet baseline	62.2
1,024 ImageNet for TRECVID	61.7
2,048 ImageNet for TRECVID + Random	63.1
4,096 ImageNet for TRECVID + Random	62.3

Default 1,000 ImageNet categories not necessarily best

Value of additional data augmentations

	Default Augmentation	Additional Augmentation
Inception	62.3	63.1
VGGNet	61.1	61.5

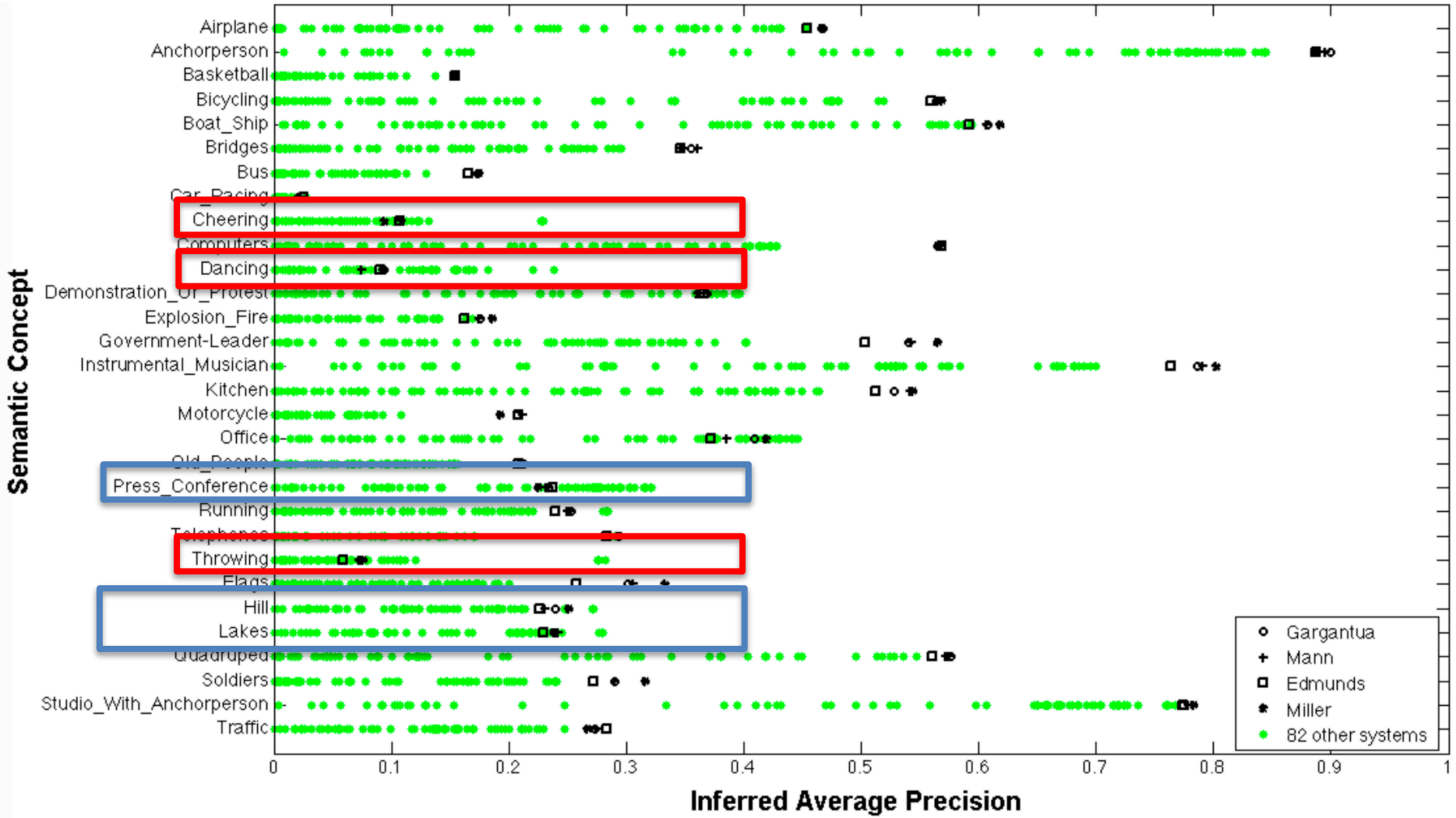
Additional augmentations give a small but consistent improvement

Value of fusion

Runs	Fusion	Internal mAP	TRECVID mAP
Gargantua	Non-weighted fusion – all 63 networks	66.9	36.0
Mann	Weighted fusion – all 63 networks	**67.3**	35.9
Edmunds	Non-weighted fusion – 32 networks	66.9	34.9
Miller	Non-weighted fusion – 7 networks	66.5	**36.2**

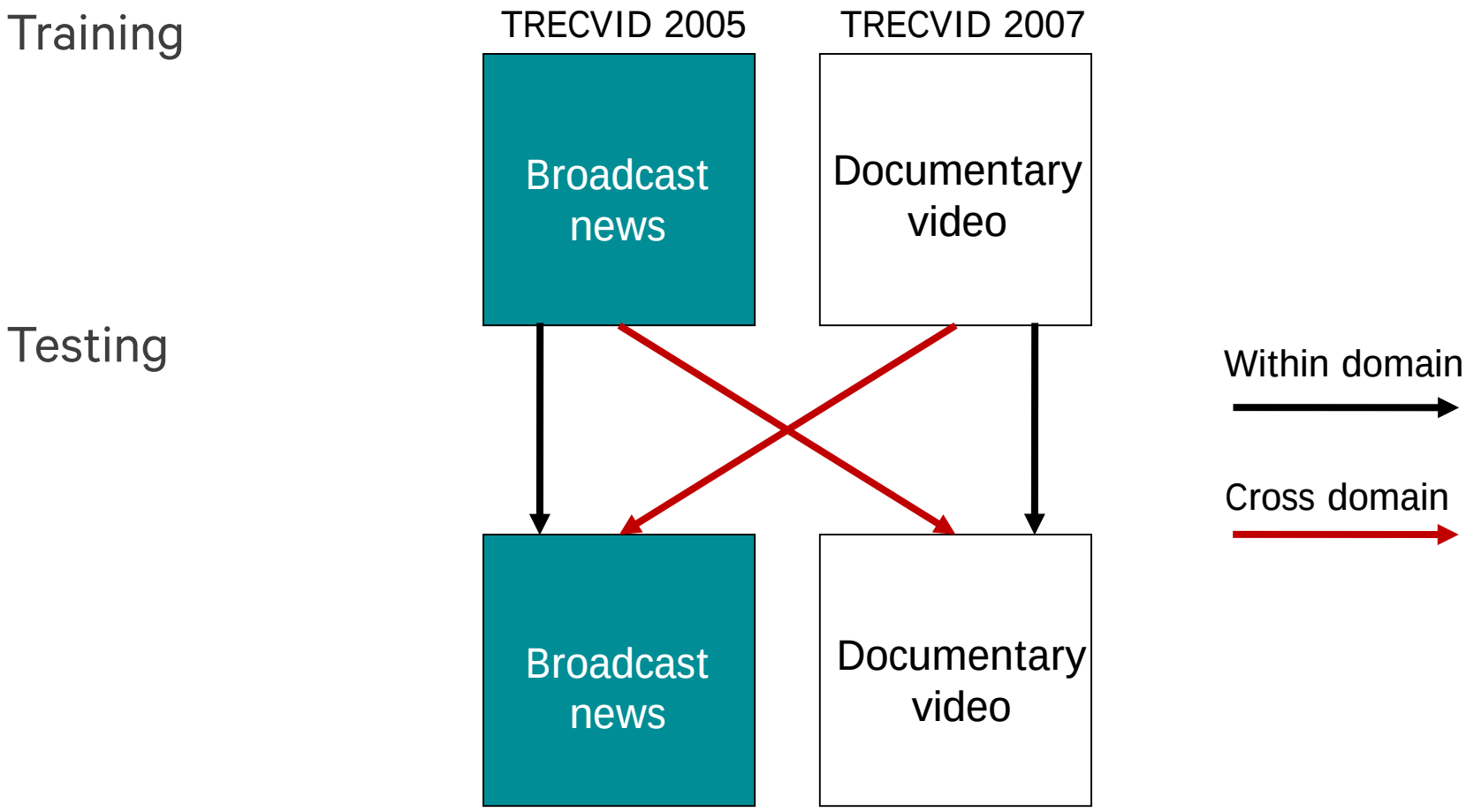
Seven diverse models fused without weights is good choice

Great for objects, ok for scenes, poor for actions

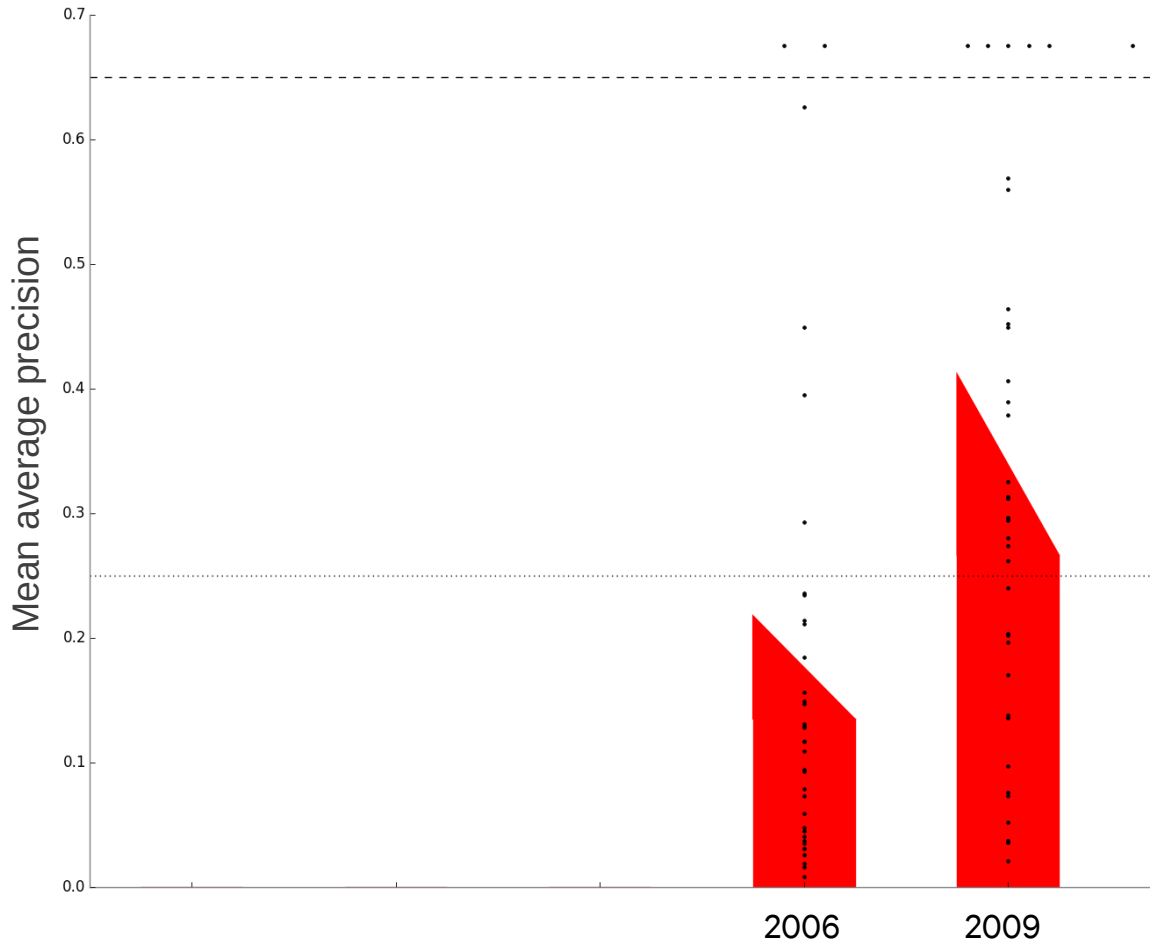


10-year progress

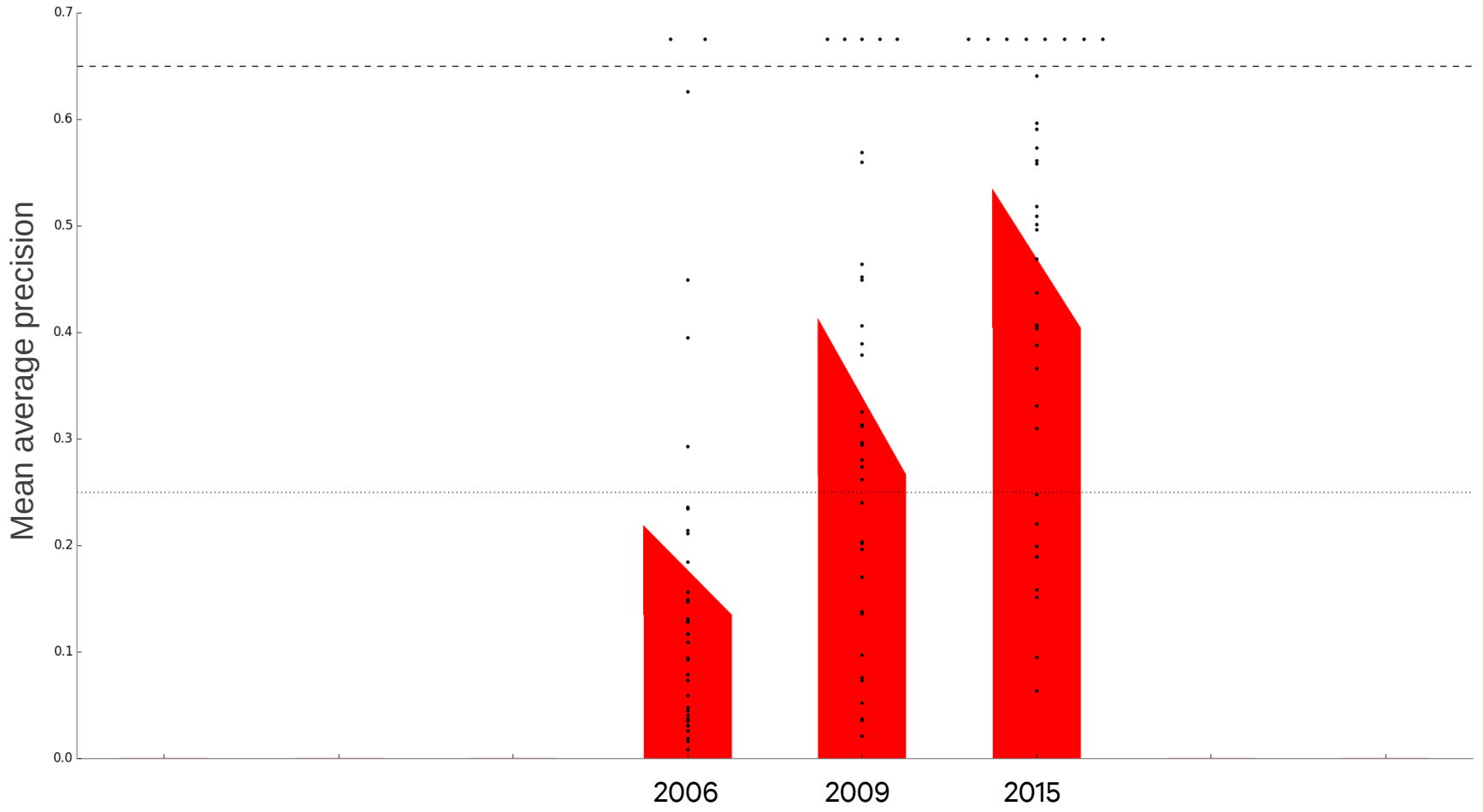
Four video data set mixtures



2006-2009: Performance doubled in just three years

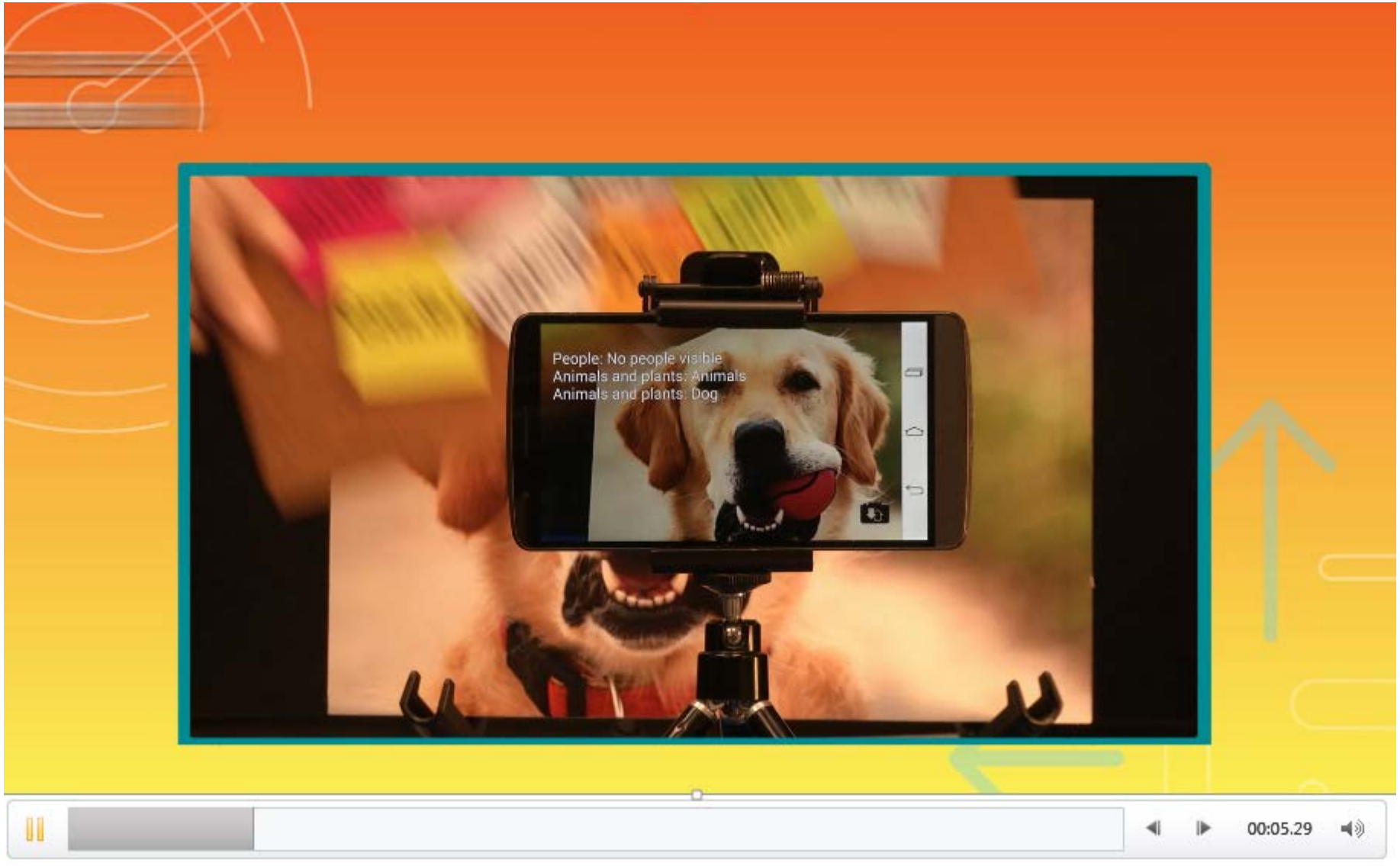


2009-2015: same jump by deep learning



Concept detection on mobile

Qualcomm Zeroth provides on-device deep learning solution



Conclusions

- Deep learning for images leading in video as well
- Technology available on mobile

- TRECVID instrumental in decade of concept detection progress
- Time for a new challenge!

Thank you

Follow us on:

For more information on Qualcomm, visit us at:
www.qualcomm.com & www.qualcomm.com/blog