Google AI

The 2nd YouTube-8M Large-Scale Video Understanding Workshop

Joonseok Lee joonseok@google.com
Walter Reade inversion@google.com

September 9, 2018

ECCV 2018

Organizers

General Chairs

Paul Natsev

Rahul Sukthankar

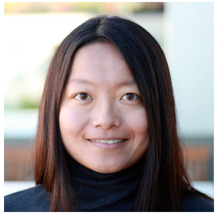
Program Chairs

Joonseok Lee

George Toderici

kaggle

Julia Elliott

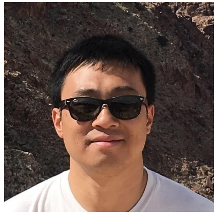

Wendy Kan

Sohier Dane

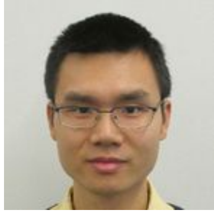
Walter Reade

Challenge Organizers

Sami Abu-El-Haija


Ke Chen

Nisarg Kothari


Hanhan Li

Sobhan Naderi Parizi

Balakrishnan Varadarajan

Joe Ng

Javier Snaider

Agenda (Morning)

Time	Content	Presenter
9:00 - 9:05	Opening Remarks	Paul Natsev
9:05 - 9:30	**Overview of 2018 YouTube-8M Dataset & Challenge**	Joonseok Lee, Walter Reade
Session 1		
9:30 - 10:00	**Invited Talk 1**: Human action recognition and the Kinetics dataset	Andrew Zisserman
10:00 - 10:30	**Invited Talk 2**: Segmental Spatio-Temporal Networks for Discovering the Language of Surgery	Rene Vidal
10:30 - 10:45	*Coffee Break*	
Session 2		
10:45 - 12:00	**Oral Session 1** ● Building a Size Constrained Predictive Model for Video Classification ● Temporal Attention Mechanism with Conditional Inference for Large-Scale Multi-Label Video Classification ● Label Denoising with Large Ensembles of Heterogeneous Neural Networks ● NeXtVLAD: An Efficient Neural Network to Aggregate Frame-level Features for Large-scale Video Classification ● Non-local NetVLAD Encoding for Video Classification	● Next top GB model (#1) ● KANU (#5) ● Samsung AI Moscow (#2) ● PhoenixLin (#3) ● YT8M-T (#4)
12:00 - 1:00	*Lunch on your own*	

Agenda (Afternoon)

Time	Content	Presenter
Session 3		
1:00 - 1:30	**Invited Talk 3**: Learning video representations for physical interactions and language-based retrieval	Josef Sivic
1:30 - 2:00	**Invited Talk 4**: Towards Video Understanding at Scale	Manohar Paluri
2:00 - 2:30	**Context-Gated DBoF Models for YouTube-8M**	Paul Natsev
2:30 - 3:45	**Poster Session**	Participants
3:45 - 4:00	*Coffee Break*	
Session 4		
4:00 - 4:45	**Oral Session 2** ● Learnable Pooling Methods for Video Classification ● Training compact deep learning models for video classification using circulant matrices ● Axon AI's Solution to the 2nd YouTube-8M Video Understanding Challenge	● Deep Topology ● Alexandre Araujo (#36) ● Axon AI (#17)
4:45 - 5:00	Closing and Award Ceremony	Paul Natsev

Introduction

Joonseok Lee (joonseok@google)

What is Video Understanding?

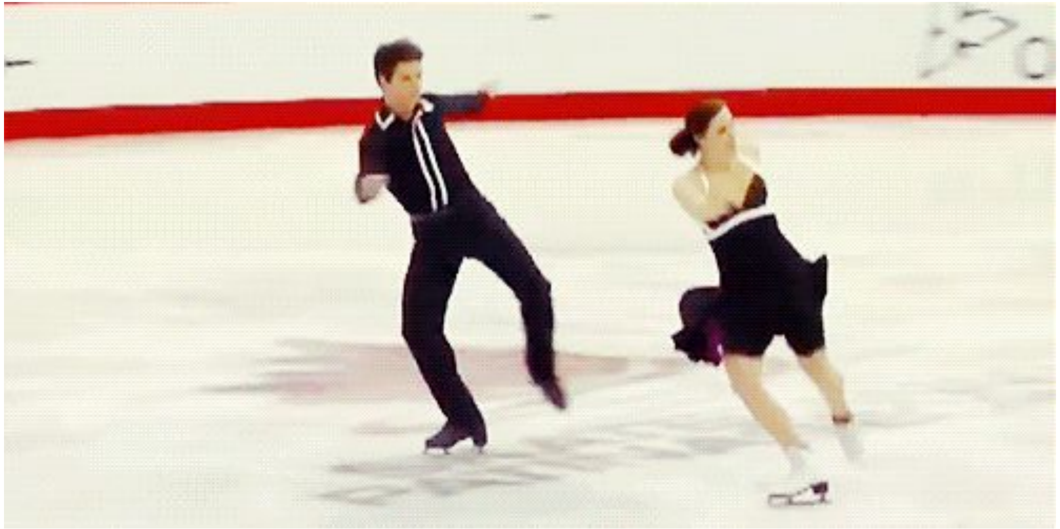
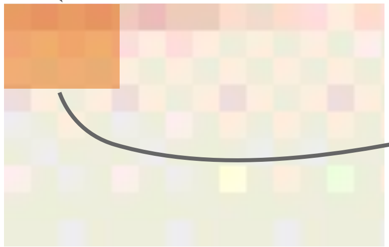


Figure skating   Winter sports   Ice rink   Pair skating

What is Video Understanding?

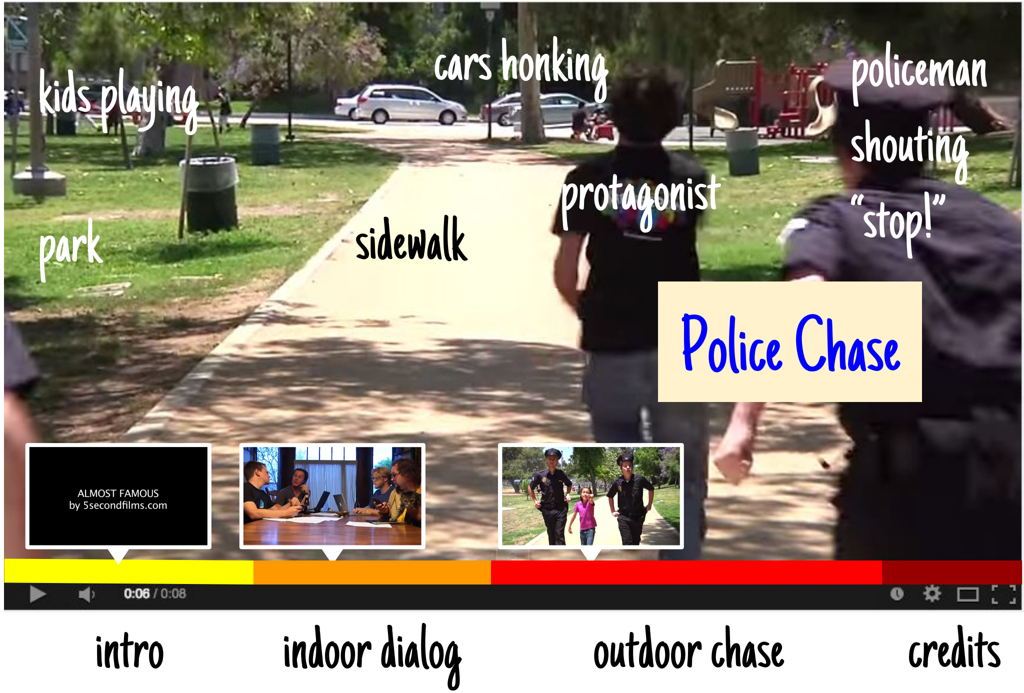


{(238, 204, 187), (238, 187, 187), …
(255, 221, 221), (255, 238, 204), …
(255, 238, 221), (238, 238, 221), …
:

$\xrightarrow{f}$

Figure skating
Winter sports
Ice rink
Pair skating

The Multiple Shades of Video Understanding



Describing the **content:**
what is visible/audible?

Inferring the **central topics:**
what is the story about?

Describing the **structure & style:**
how is the story told?

Inferring **creator / viewer intent:**

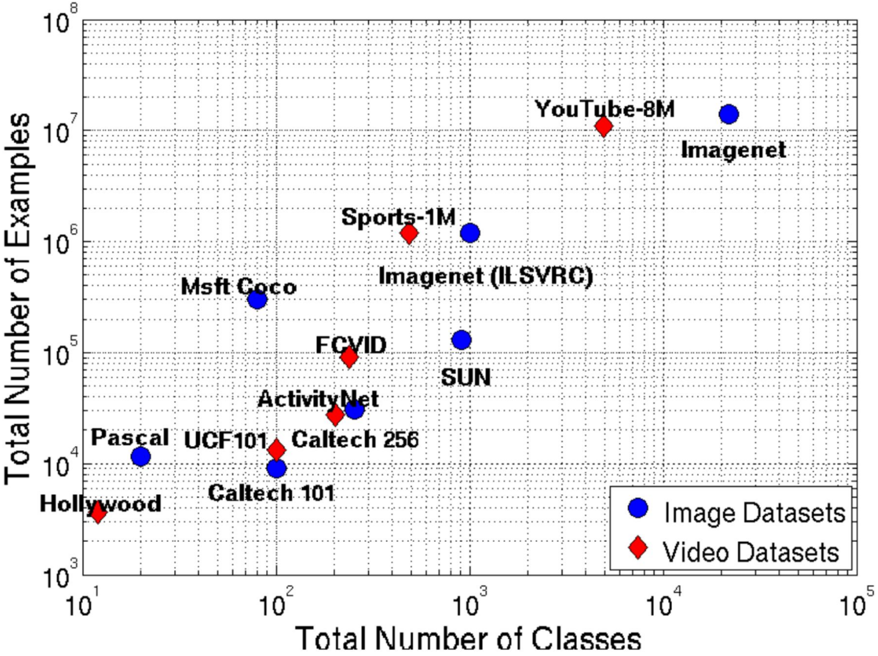
- **why capture** this video?
- **why watch** this video?

YouTube-8M: Primary Motives

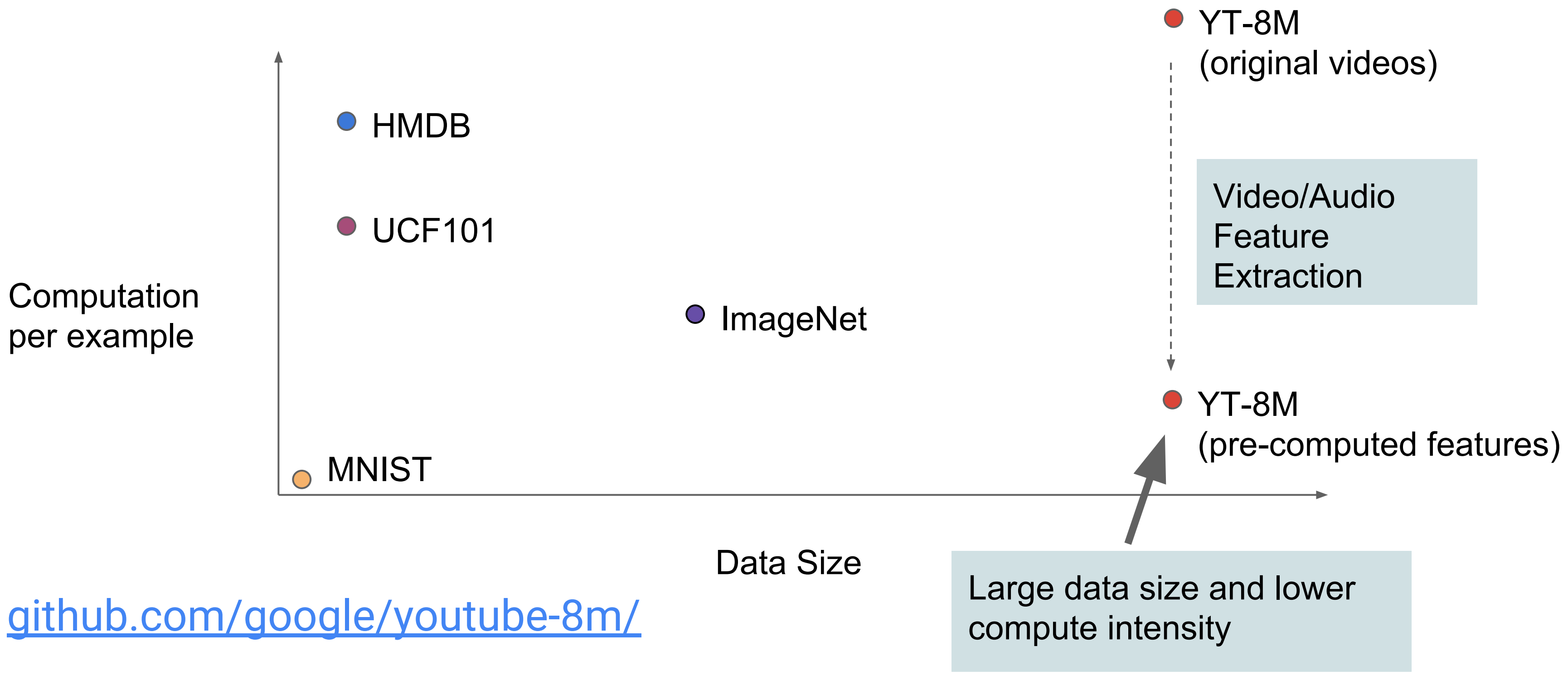
- Help advance the state-of-the-art in Video Understanding
  - By providing a large, free, realistic, labeled video dataset
  - Hoping that we can collaborate with the research community to reach better-than-human performance on Video Classification, similar to Image Classification tasks.
- Establishing a representative sample of YouTube
  - The YouTube corpus is HUGE - slow to train on
  - It is faster for us to continuously test our ideas on a smaller yet representative dataset.

Challenges in Creating Video Dataset

- File sizes are larger than images.
  - More expensive to download, store, and train from.
- Video labels are more expensive to obtain.
  - Requiring annotators to watch the video and listen to audio stream.
- Therefore, existing video datasets tend to be small.



YouTube-8M: TensorFlow Framework Design



YT-8M
(original videos)

HMDB

UCF101

Video/Audio
Feature
Extraction

Computation
per example

ImageNet

YT-8M
(pre-computed features)

MNIST

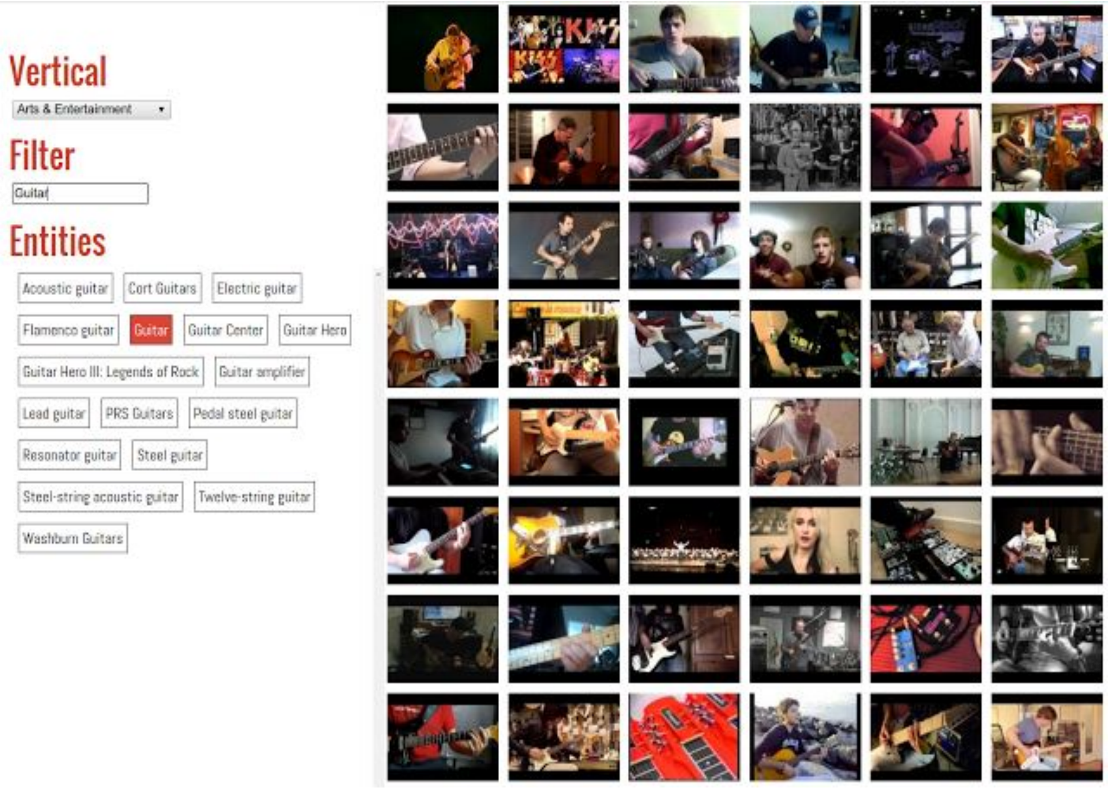
Data Size

Large data size and lower compute intensity

github.com/google/youtube-8m/

YouTube-8M: The Dataset (v3)

- 6.1M videos
- 350,000 hours
- 2.6B audio/visual features
- 3,862 classes
- 3.0 labels/video



Vertical

Arts & Entertainment ▾

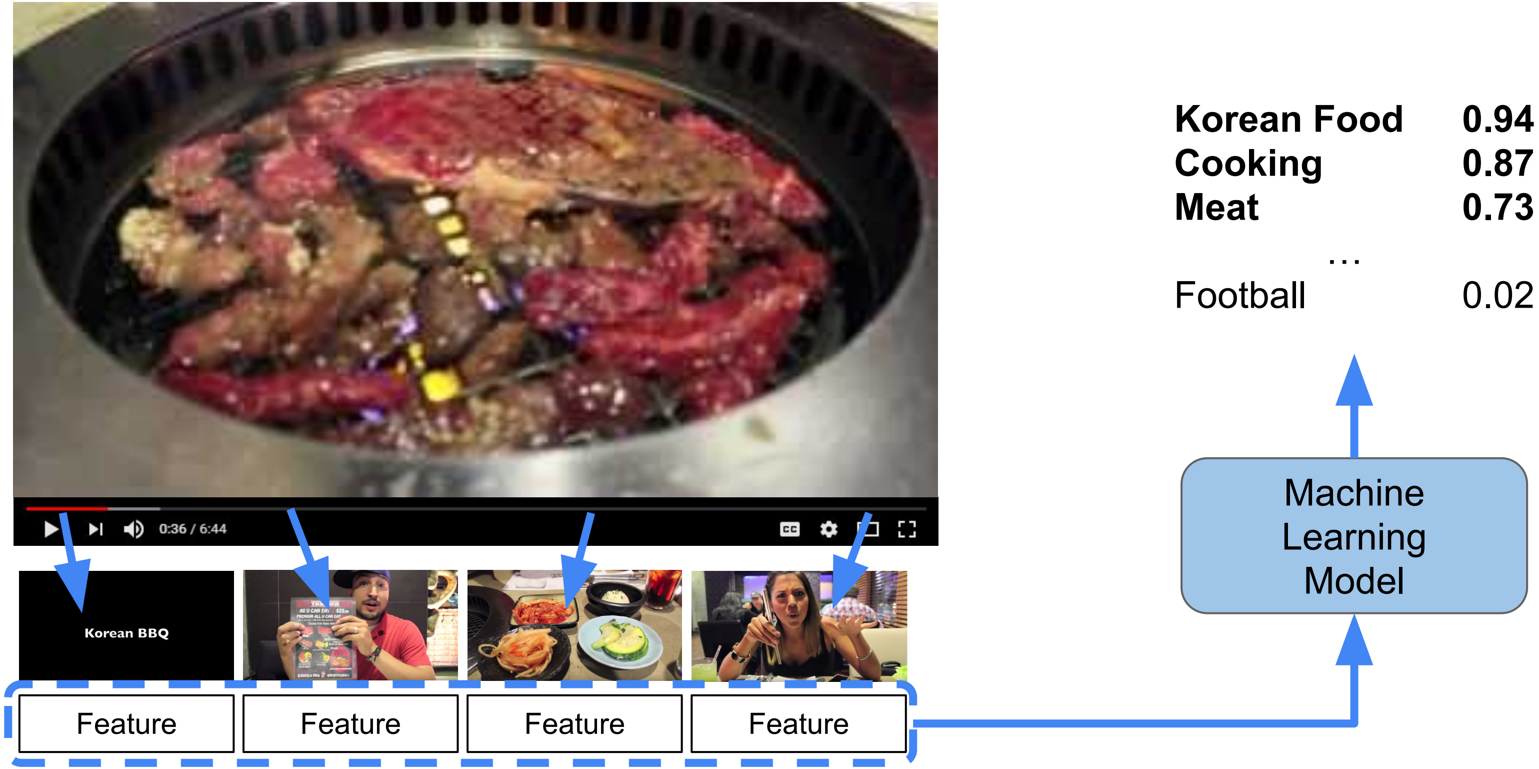
Filter

Guitar

Entities

2018 YouTube-8M Challenge

YouTube-8M Classification Challenge Task



Feature

Feature

Feature

Feature

Machine Learning Model

Korean Food	**0.94**
Cooking	**0.87**
Meat	**0.73**
…	
Football	0.02

YouTube-8M Classification Challenge Task

- Input:
  - A sequence of frame-level audio-visual features, extracted at 1 fps
  - Each video has [120, 300] frame-level features
  - Visual Inception-V3 bottleneck features extracted from pixels (PCA-ed to **1024D**)
  - Audio Resnet-ish bottleneck features extracted from audio spectrograms (**128D**)
- Target:
  - Video topics from a 3,862 Knowledge Graph entity vocabulary
  - The target topics cover the **main themes** in the video (vs. object detection, scene parsing, etc.)
  - Each video has 3.0 ground truth labels on average
- New in 2018: **Model size must be < 1GB**.
- Goal: Predict target video topics from the sequence of frame-level features

Last Year’s Leaderboard

Rank	Team Name	Best Performance (GAP)		# models in ensemble
		Single model	Ensembled	
1	**WILLOW**	**0.8300**	**0.8496**	25
2	**monkeytyping**	0.8179	0.8458	74
3	**offline**	0.8275	0.8454	57
4	**FDT**	0.8178	0.8419	38
5	**You8M**	0.8225	0.8418	33
6	**Rankyou**	0.8246	0.8408	22
7	**Yeti**	0.8254	0.8396	21
8	**SNUVL X SKT**	0.8200	0.8389	22
9	**LanzanRamen**	—	0.8372	—
10	**Samartian**	0.8139	0.8366	36

Scores in GAP; higher values are better.

Gray scores mean that it’s not published, but we got to know it by contacting them.

Approaches Overview

- Temporal aggregation
  - (Variants of) NetVLAD: most widely used.
  - LSTM/GRU
  - Attention model
- Architecture
  - WILLOW architecture (2017 Winner): most widely used.
  - ResNet

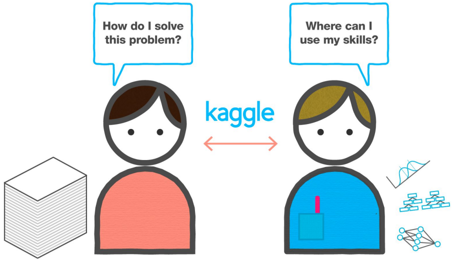
Approaches Overview

- Ensembles
  - Top performers are still taking advantage of ensembling.
  - # of models decreased: mostly around 3 - 6 models.
    - Heaviest ensemble model combined 115 models.
- Distillation
  - Most top performers distilled from larger, ensembled teacher model.
- Quantization
  - float16 is used instead of full 4 bytes float.

Kaggle Overview and Community View of Competition

Walter Reade (inversion@kaggle)

Kaggle Background



- Well-Known for Machine Learning Contests
    - Connect talent to business
    - Diverse methods of approaching the problem
    - Find upper limit of signal in the data

- Rapidly Becoming the Place To Do Data Science Projects
    - Find and upload high-quality datasets
    - Build models in the cloud (Kernels)
    - Connect with the Community (world's largest)
    - Faster Data Science Education

"No one beginning a new data project should start from a blinking cursor"

YT8M: A Unique Competition

- Large dataset
    - 1.7 TB was the largest dataset on Kaggle when 1st competition launched
    - (TSA Passenger Screening took 1st place with ~6 TB)
- Strong baseline starter code to help level the playing field
    - Runs on Google Cloud ML Engine
    - TensorFlow
- Google Cloud Credits
    - Free GCP credit ($300 x 200) provided by Kaggle
- Strong and high quality participants

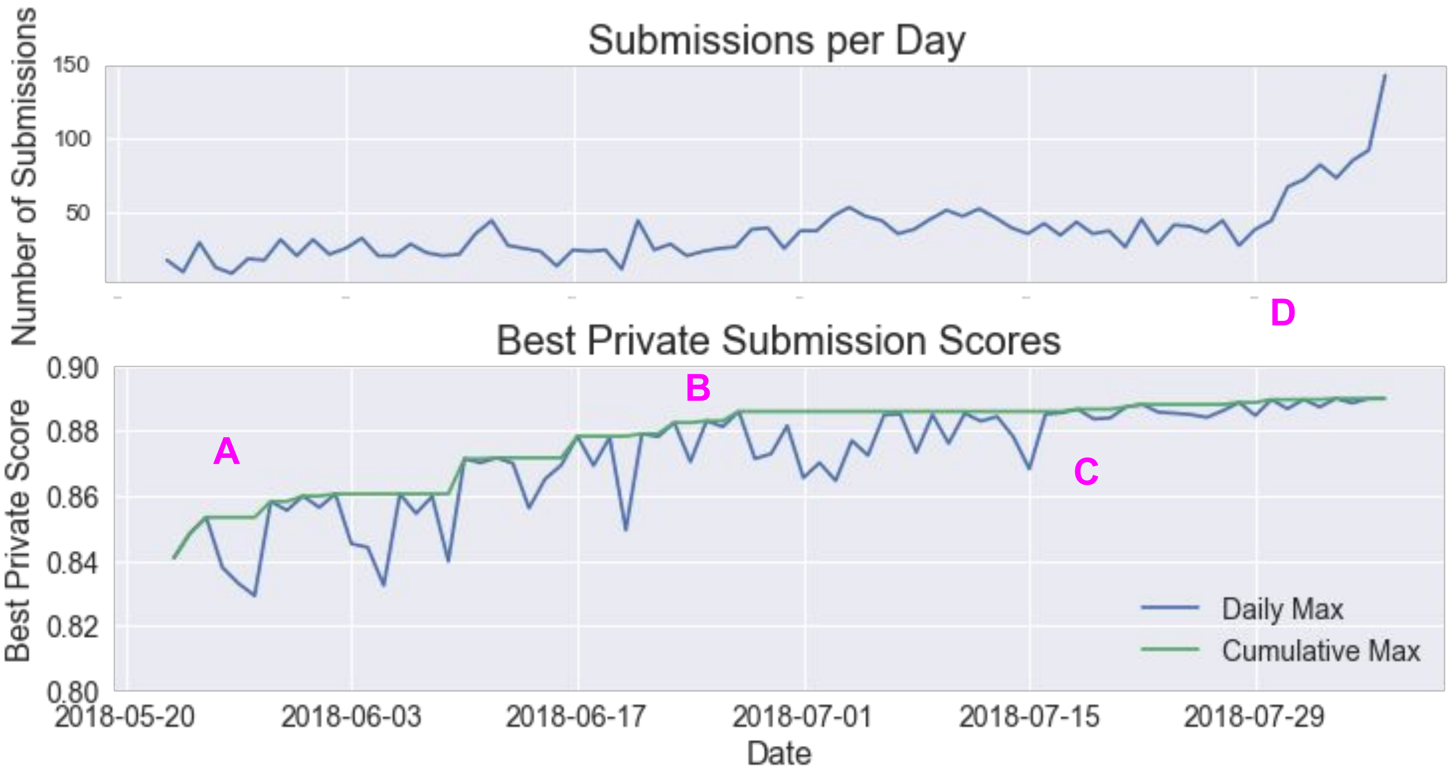
Where were the participants from?

- 394 teams
- 531 competitors
  - 106 - First Kaggle competition
  - 61 - Also participated in 1st competition
- Participants from 40+ countries

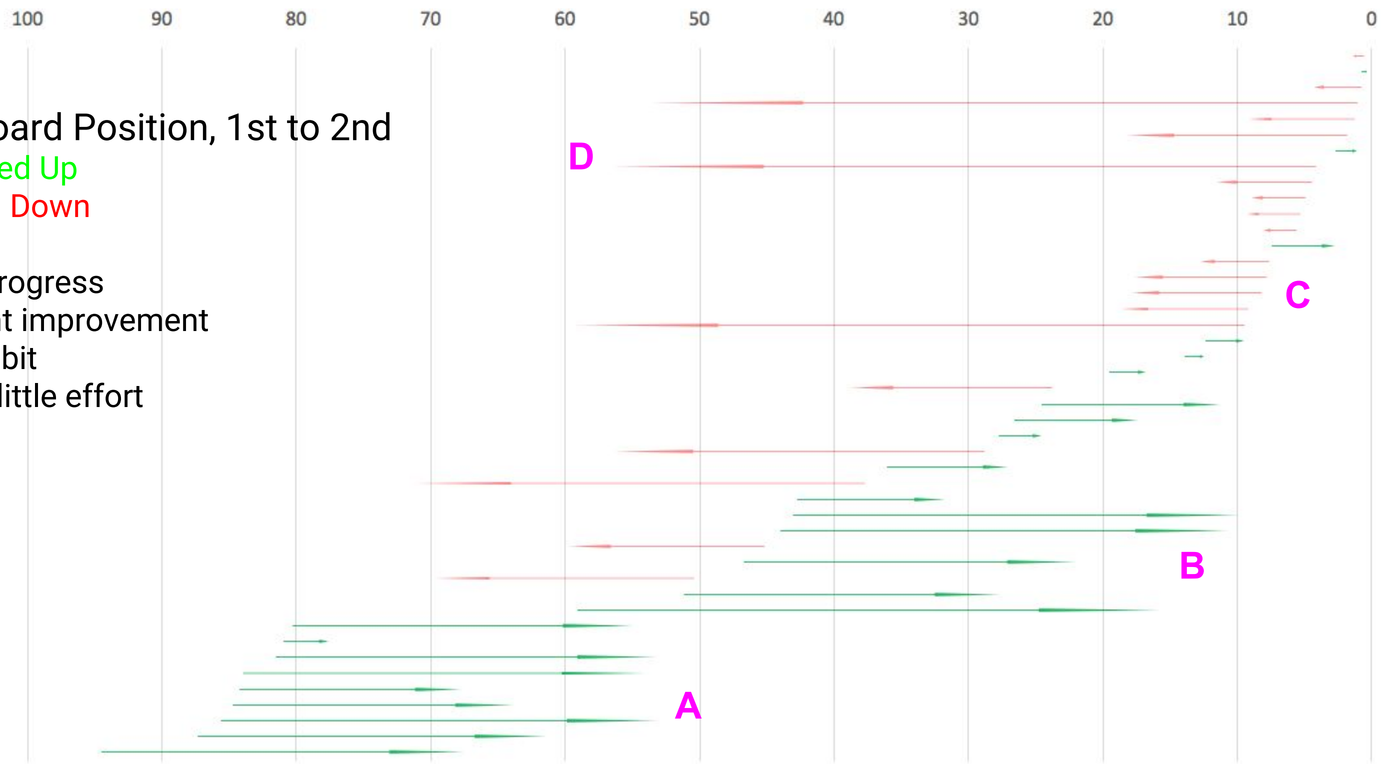
- Total of 3,805 submissions
  - Relatively low ~10 subs/team
  - Median Competition ~15

Country	#Competitors
US	136
CN	69
IN	56
RU	30
KR	25
JP	19
FR	15
CA	15
GB	14
TW	10
SG	9
HK	9
BY	8
UA	8
DE	7
PL	6
AU	5
GR	4

Competition Progression



How did returning competitors do?



% Leaderboard Position, 1st to 2nd

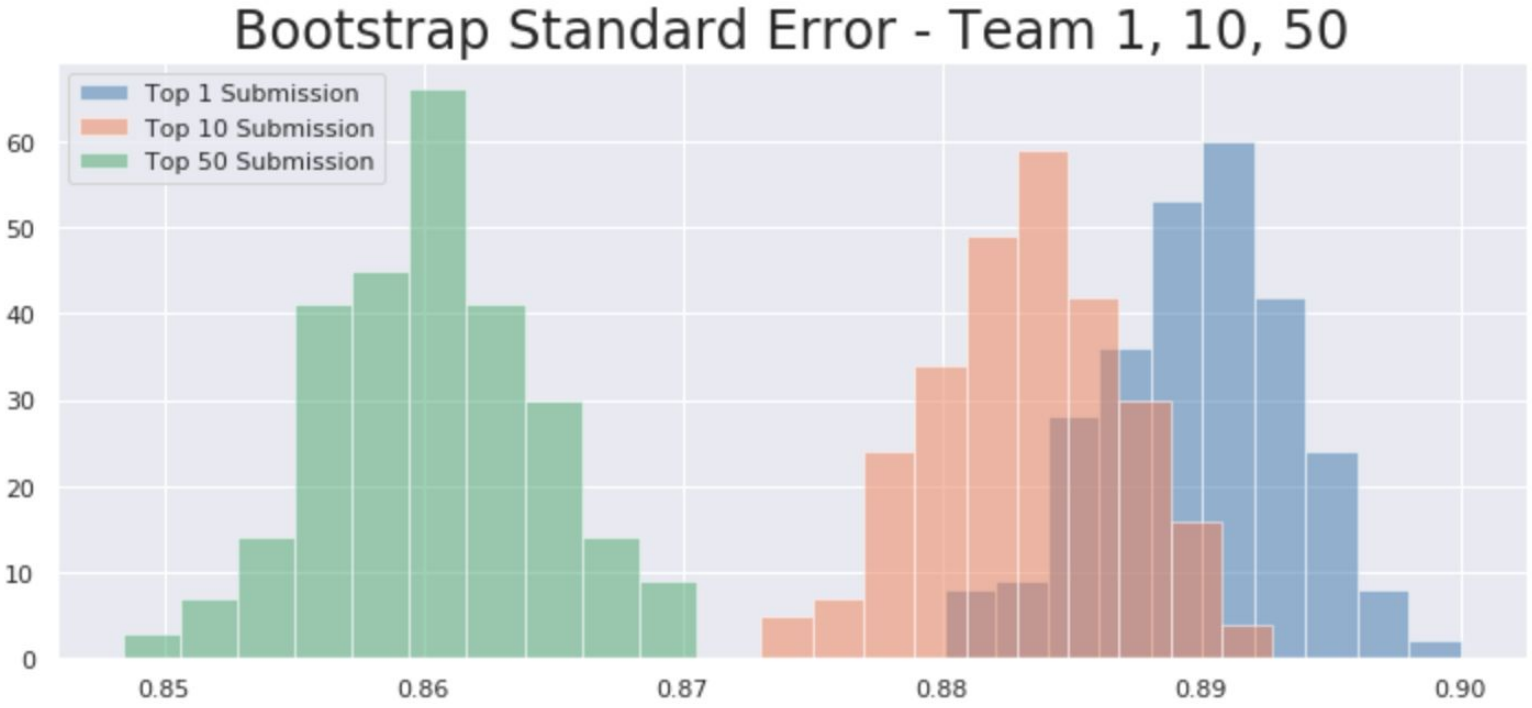
Green = Moved Up

Red = Moved Down

A - Modest progress
B - Significant improvement
C - Slipped a bit
D - Probably little effort

Out of Top 10 Teams, 1, 4, and 8 had returning competitors!

Separation Between Models



Google AI

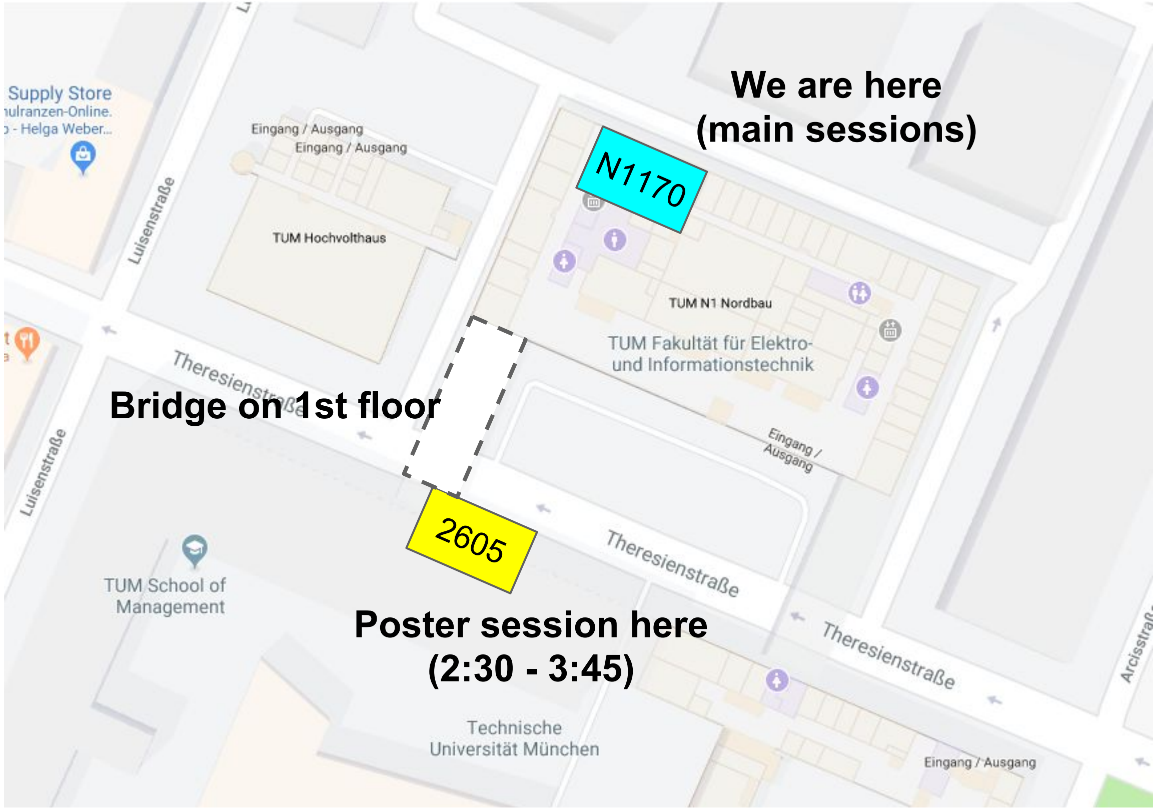
Agenda (Morning)

Time	**Content**	**Presenter**
9:00 - 9:05	Opening Remarks	Paul Natsev
9:05 - 9:30	**Overview of 2018 YouTube-8M Dataset & Challenge**	Joonseok Lee, Walter Reade
Session 1		
9:30 - 10:00	**Invited Talk 1**: Human action recognition and the Kinetics dataset	Andrew Zisserman
10:00 - 10:30	**Invited Talk 2**: Segmental Spatio-Temporal Networks for Discovering the Language of Surgery	Rene Vidal
10:30 - 10:45	*Coffee Break*	
Session 2		
10:45 - 12:00	**Oral Session 1** ● Building a Size Constrained Predictive Model for Video Classification ● Temporal Attention Mechanism with Conditional Inference for Large-Scale Multi-Label Video Classification ● Label Denoising with Large Ensembles of Heterogeneous Neural Networks ● NeXtVLAD: An Efficient Neural Network to Aggregate Frame-level Features for Large-scale Video Classification ● Non-local NetVLAD Encoding for Video Classification	● Next top GB model (#1) ● KANU (#5) ● Samsung AI Moscow (#2) ● PhoenixLin (#3) ● YT8M-T (#4)
12:00 - 1:00	*Lunch on your own*	

Agenda (Afternoon)

Time	Content	Presenter
Session 3		
1:00 - 1:30	**Invited Talk 3**: Learning video representations for physical interactions and language-based retrieval	Josef Sivic
1:30 - 2:00	**Invited Talk 4**: Towards Video Understanding at Scale	Manohar Paluri
2:00 - 2:30	**Context-Gated DBoF Models for YouTube-8M**	Paul Natsev
2:30 - 3:45	**Poster Session**	Participants
3:45 - 4:00	*Coffee Break*	
Session 4		
4:00 - 4:45	**Oral Session 2** ● Learnable Pooling Methods for Video Classification ● Training compact deep learning models for video classification using circulant matrices ● Axon AI's Solution to the 2nd YouTube-8M Video Understanding Challenge	● Deep Topology ● Alexandre Araujo (#36) ● Axon AI (#17)
4:45 - 5:00	Closing and Award Ceremony	Paul Natsev

Poster Session Location



Room **2605** (+upstairs)
in Building 6 (Theresianum)
across the street.

Please set up your poster at the designated board **during/after** lunch hour.

Google AI

Thank you for your attention.