MOTIVATION:
- Video understanding is a challenging task for numerous applications and research.
- This project addresses the problem of multi-label video classification and temporal localizations for user-generated videos.

INPUTS:
- YouTube-8M frame-level features dataset and segment-rated dataset.

APPROACH:
- Video-level
  - visual and audio features aggregation with NetVLAD.
  - Mixture-of-Experts for final classification
- Segment-level
  - Transfer learning based on video-level model.
  - Context-ignore and context-aware combined model.

RESULTS:
- Video-level model achieves 85% global average precision. Segment-level model achieves 82% mean average precision.

DATA
- YouTube-8M dataset released by Google. Millions of YouTube videos, with machine-generated annotations from a diverse vocabulary of 3,800+ visual entities
- YouTube-8M Segments Dataset: which includes human verified labels at the 5-second segment level

METHODS
- Video-level model: NetVLAD layer for features aggregation, MoE for the final classification
- Segment-level model: Transfer learning based on video-level model, Context-ignore and context-aware combined model.

CONCLUSION
- Classifier with NetVLAD aggregation and Mixture-of-Experts achieves gAP of 85% in large-scale video classification.
- Transfer learning with context-aware and context-ignore combined model achieves mAP of 82% in temporal localization.

SUMMARY:
- Incorporate temporal features in video classification as the current algorithm is focused on static features. E.g. combine NetVLAD, RNN and MoE.
- Reduce model size. Current video-level model is 3.72G, and segment-level model is 10G.