
# The Monkeytyping Solution to YouTube-8M Video Understanding Challenge

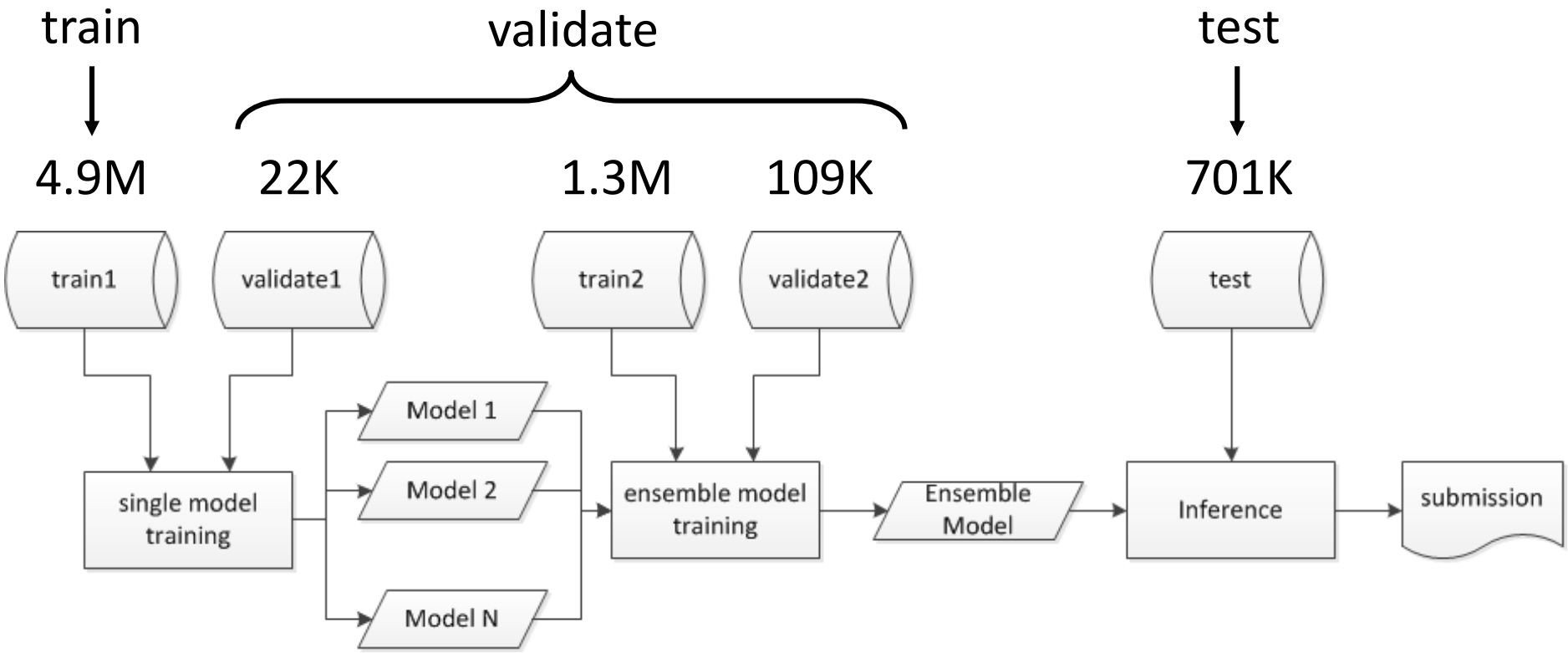
**Heda Wang**
whd.thu@gmail.com

Teng Zhang
zhangteng1887@gmail.com

Multimedia Signal and Intelligent Information Processing Laboratory
Department of Electronic Engineering
Tsinghua University

2017/07/26

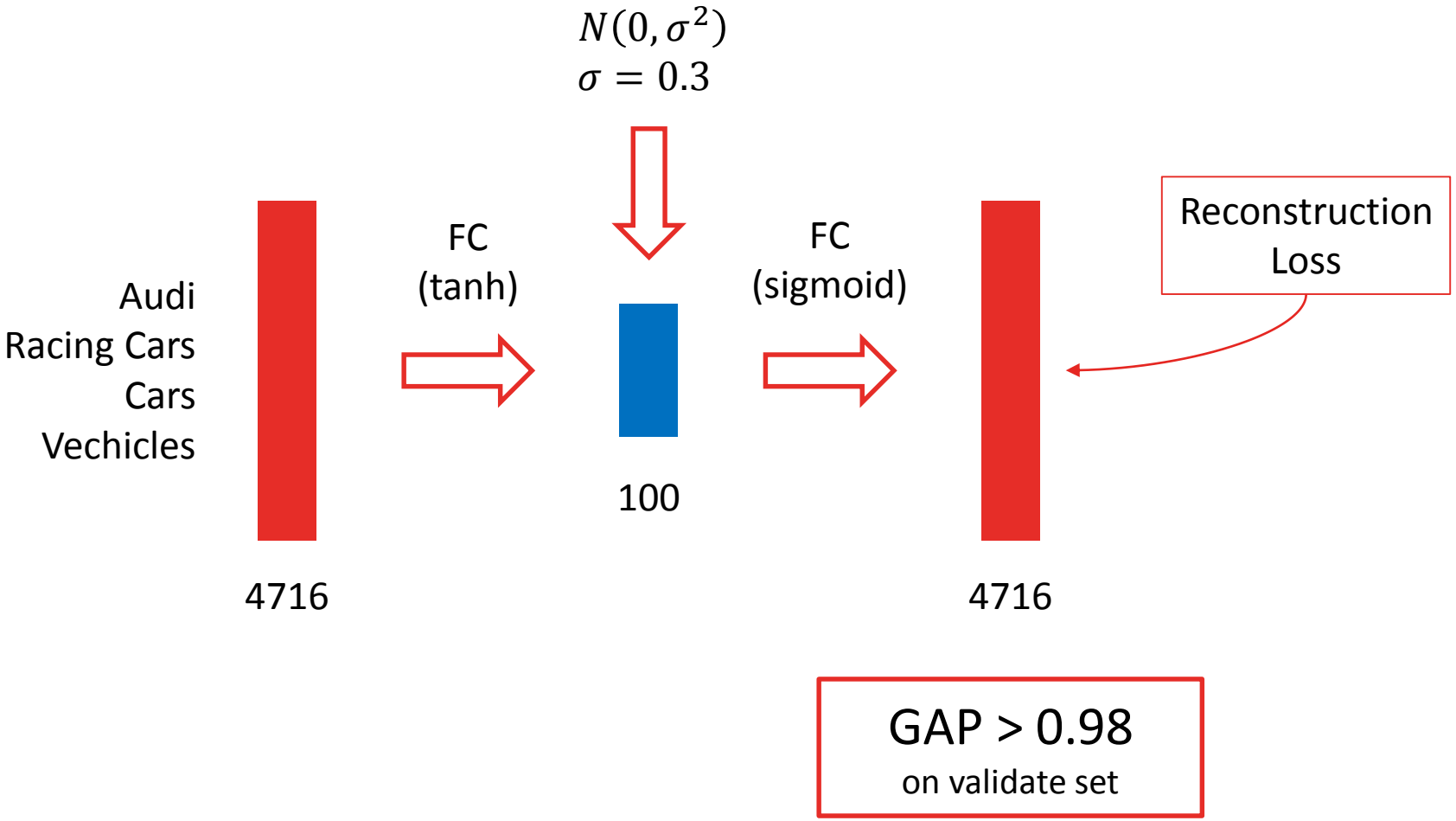
# The framework




4.9M -> 6.3M, single model GAP@20 +0.4%

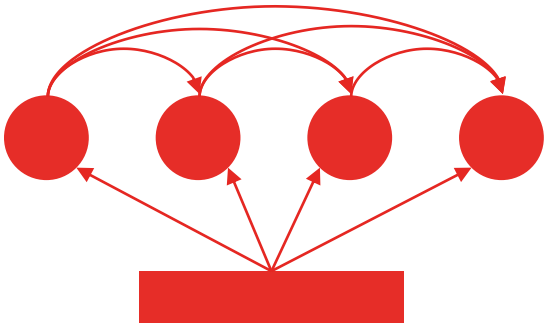
Linear stacking -> attention stacking, ensemble GAP@20 +0.1%

# Labels are correlated

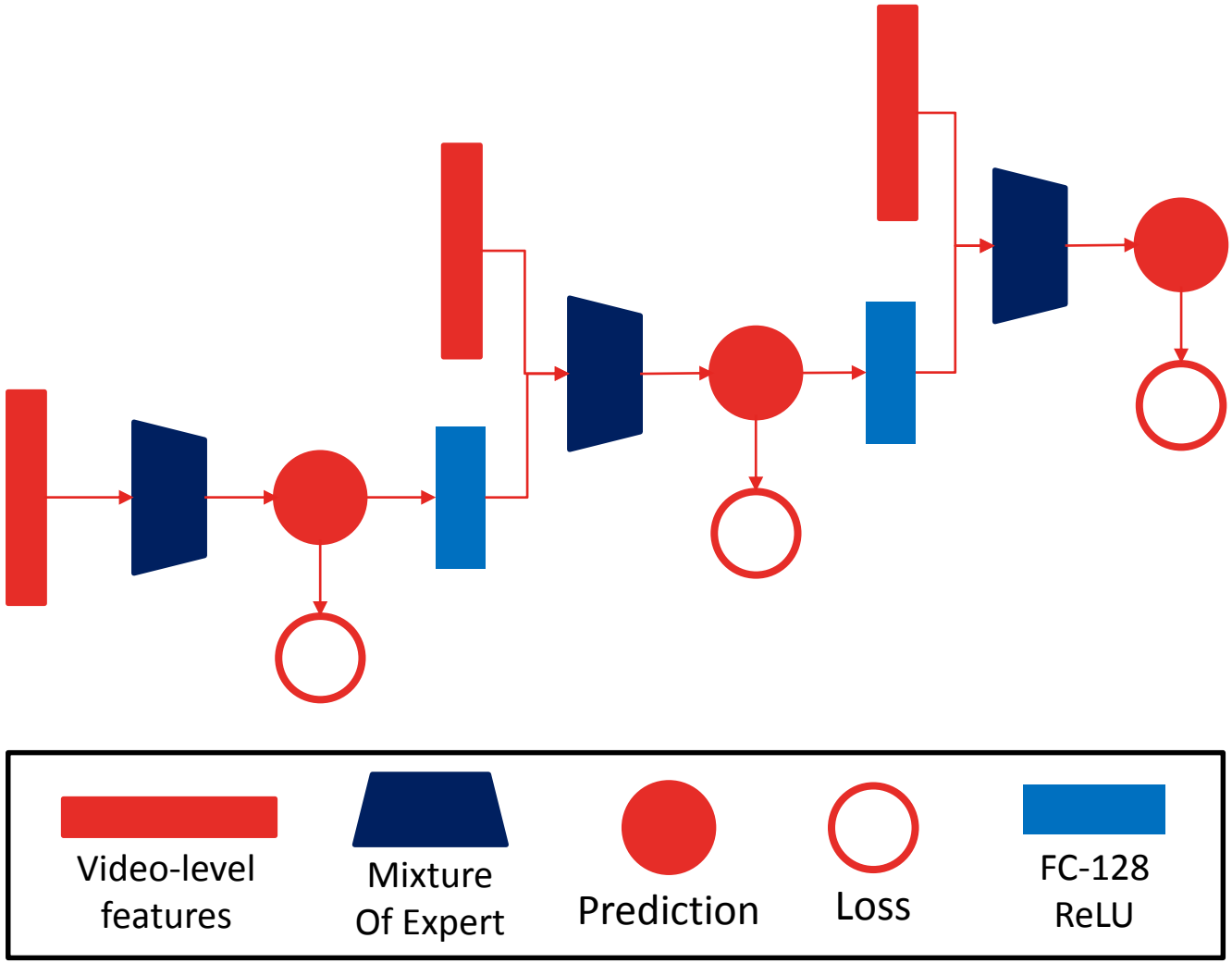


# Existing approaches for multi-label classification

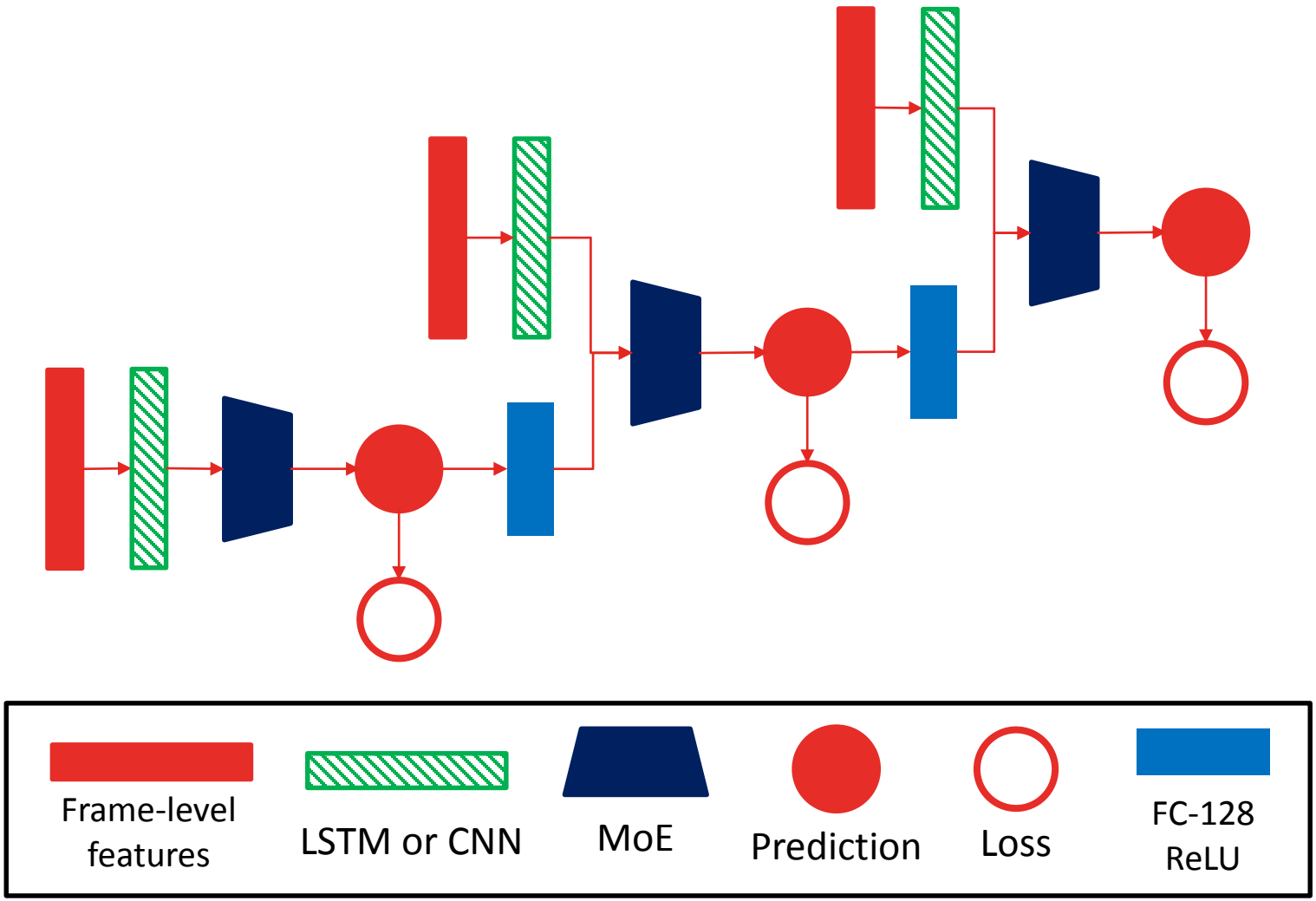
- Probabilistic Graphic Models
  - $P(L_1, L_2, \ldots, L_n|X)$
  - Typically n < 100
- (Ensemble of) Classifier Chains
  - Sequentially training and testing
  - Typically n < 200
  - Need to train a lot of classifiers



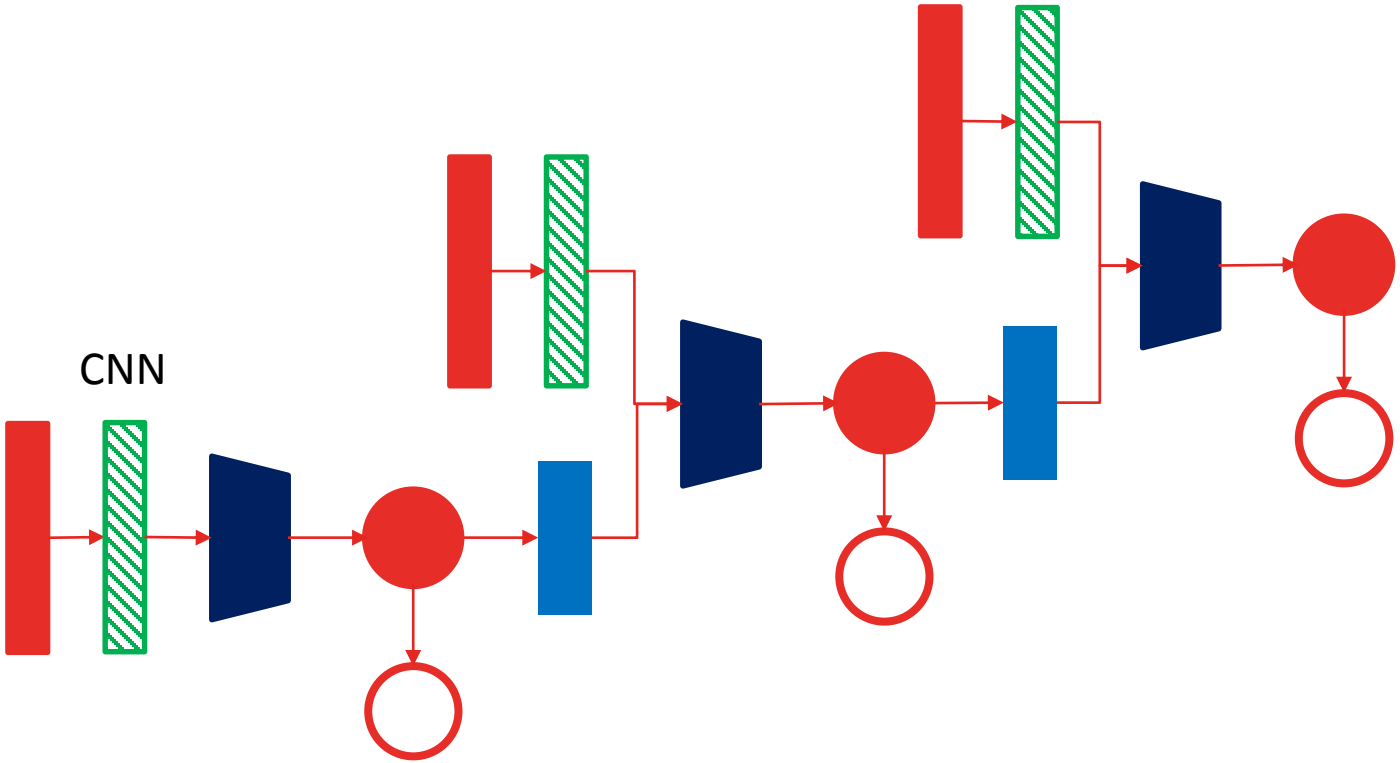
# Explicitly model label correlation by Chaining

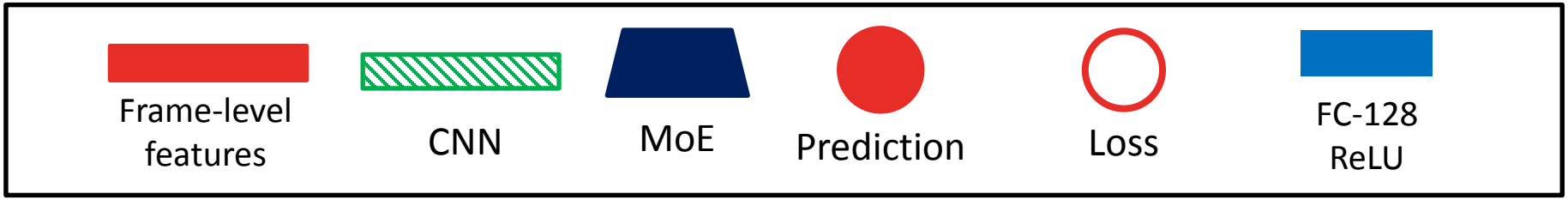


# Explicitly model label correlation by Chaining



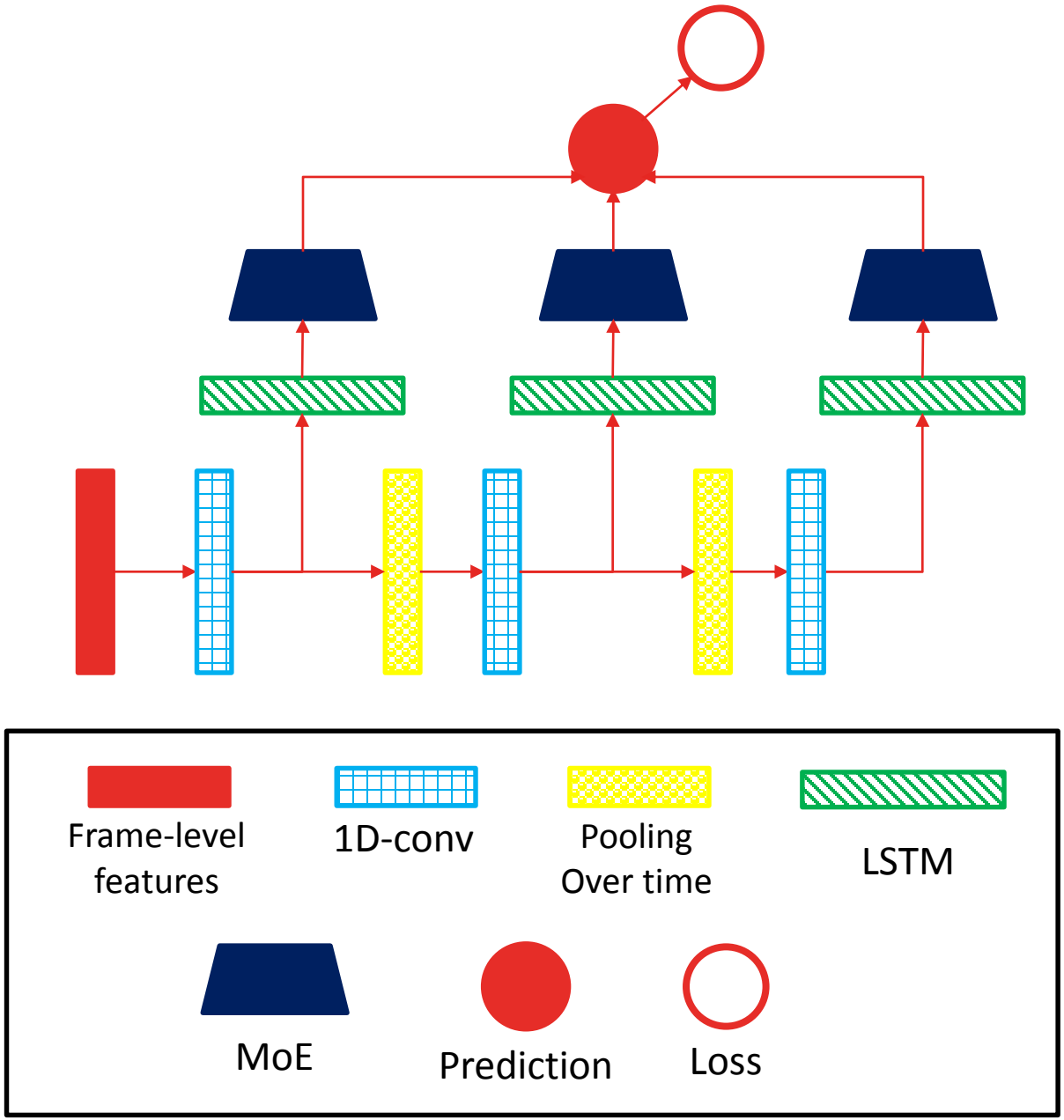
# Explicitly model label correlation by Chaining



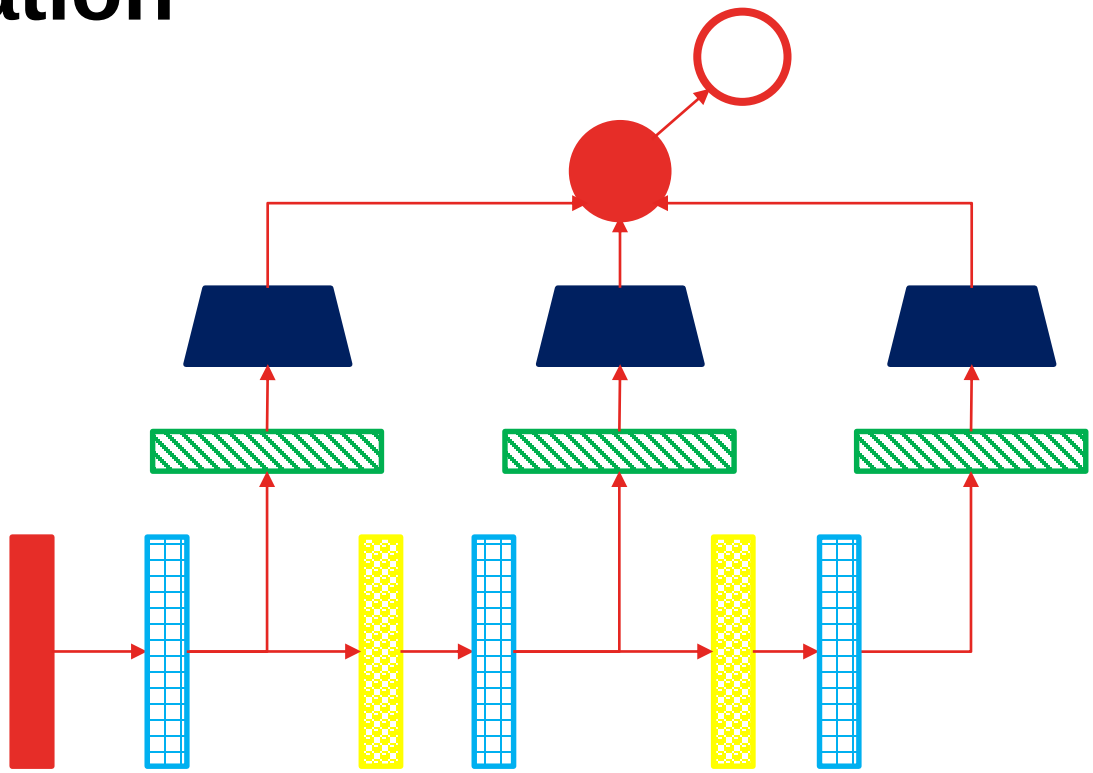



# Explicitly model label correlation by Chaining

| Model | | Parameters | Chaining |
|---|---|---|---|
| Video-level MoE | Original | #mixture=16 | 0.7965 |
| | Chaining | #stage=8, #mixture=2 | **0.8106** |
| 1D-CNN | Original | (1,2,3,3)x512 | 0.7904 |
| | Chaining | #stage=4, (1,2,3,3)x128 | **0.8179** |
| LSTM | Original | #mixture=8 | 0.8131 |
| | Chaining | #stage=2, #mixture=4 | **0.8172** |

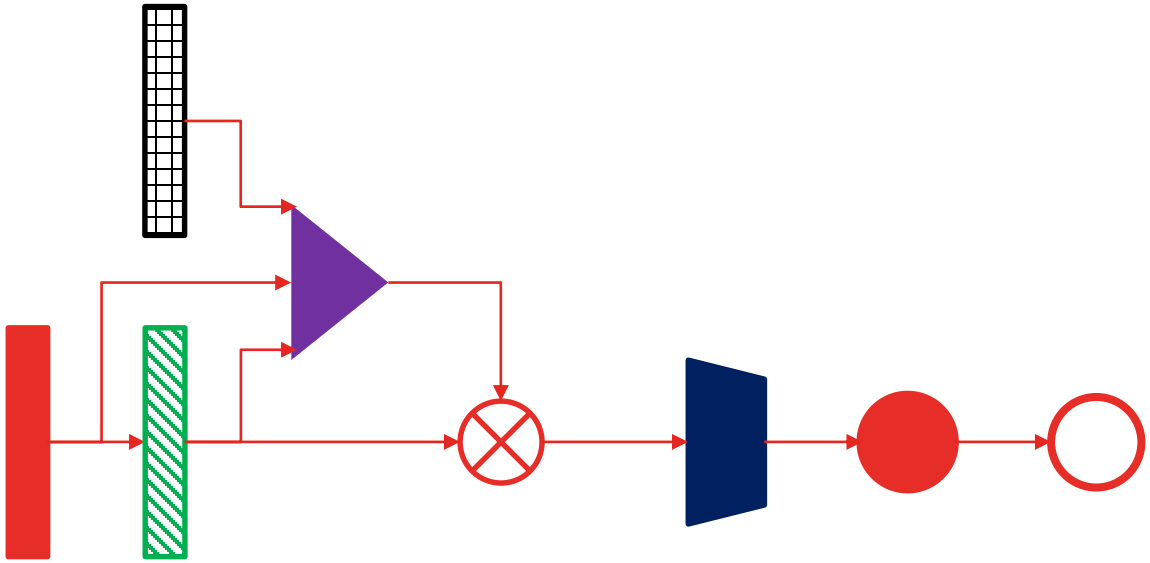

Frame-level features
1D-conv
Pooling Over time
LSTM
MoE
Prediction
Loss

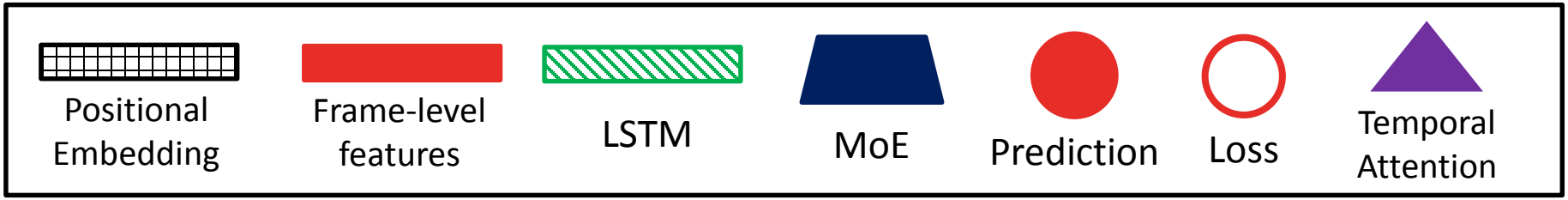
# Modeling temporal multi-scale information



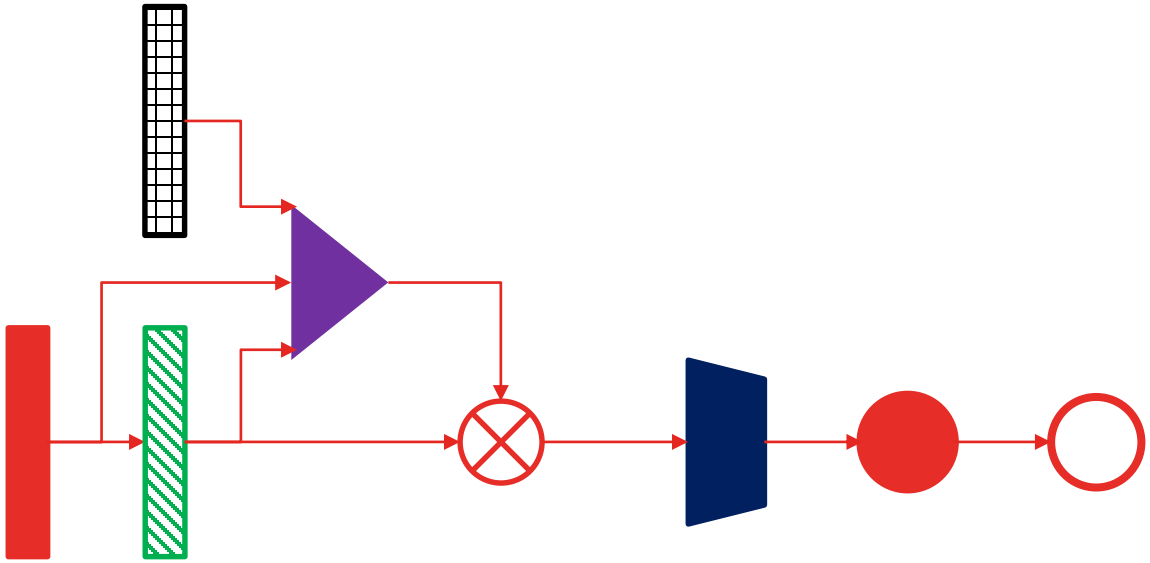
| Network type | GAP@20 |
|---|---|
| Vanilla LSTM | 0.8131 |
| Multi-Scale CNN-LSTM | 0.8204 |

# Attention pooling for saliency detection





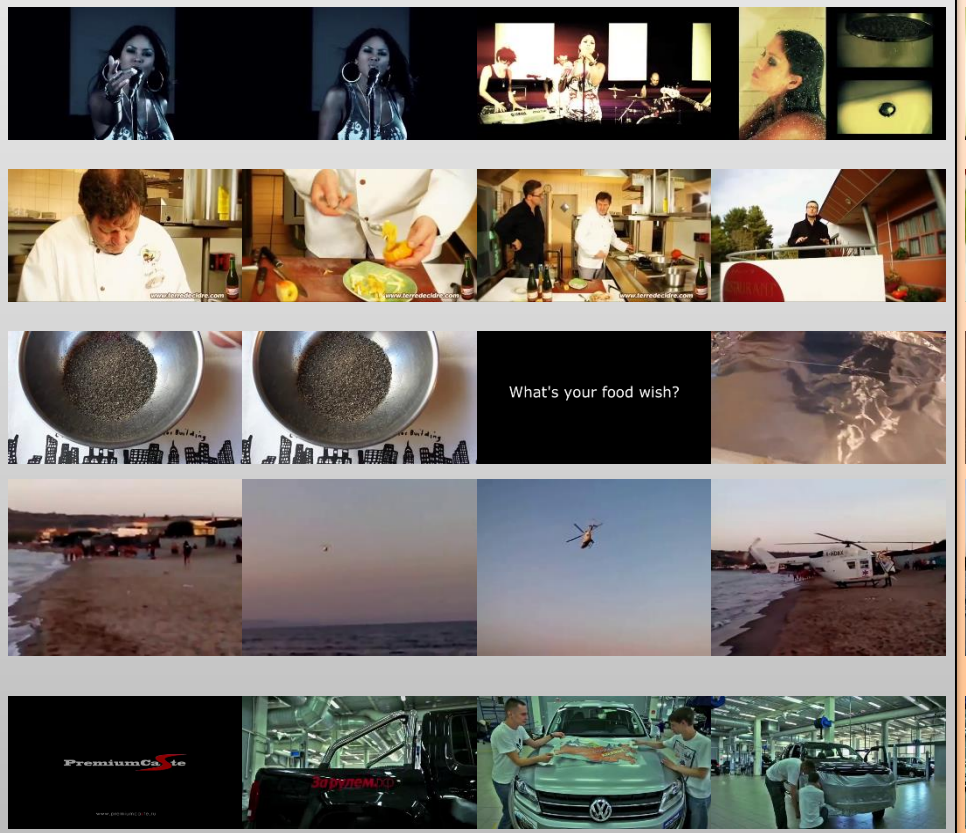
# Attention pooling for saliency detection



| Network type | GAP@20 |
|---|---|
| Vanilla LSTM | 0.8131 |
| Attention LSTM | 0.8157 |
| Positional-embedded Attention LSTM | 0.8169 |

# Attention pooling for saliency detection

## Frames with low attention value




## Frames with high attention value

# The roadmap

| Ensembles | GAP@20 (Private LeaderBoard) |
|---|---|
| Ensemble of 27 single models<br>Includes 7 chaining models, 5 multi-scale models, 5 attention-pooling models, and 10 lstm models | 0.8425 |
| + 11 bagging & boosting models | 0.8435 |
| + 8 distillation models | 0.8437 |
| + 28 cascade models | 0.8453 |
| Attention Weighted Stacking | 0.8459 |

# Summary

- Multi-label video classification
  - Address multi-label problem with chaining
  - Model multi-scale temporal information
  - Select salient frames with attention pooling-over-time

# Summary

- Multi-label video classification
  - Address multi-label problem with chaining
  - Model multi-scale temporal information
  - Select salient frames with attention pooling-over-time

- More details
  - And bagging, boosting, distillation, cascade, stacking, etc.
  - Please refer to our paper
  - Paper: https://arxiv.org/abs/1706.05150
  - Code: https://github.com/wangheda/youtube-8m

- Thank you