Video Emotion Recognition with Transferred Deep Feature Encodings

Baohan Xu[1], Yanwei Fu[12],Yu-Gang Jiang[1], Boyang Li[2] and Leonid Sigal[2]

1. **Fudan University, Shanghai, China**
2. **Disney Research**

ICMR 2016,, USA, June 2016 New York

Introduction

Motivation

YouTube vimeo Dailymotion twitch ...

300 hours of video uploaded to YouTube every minute

Fear

Sadness

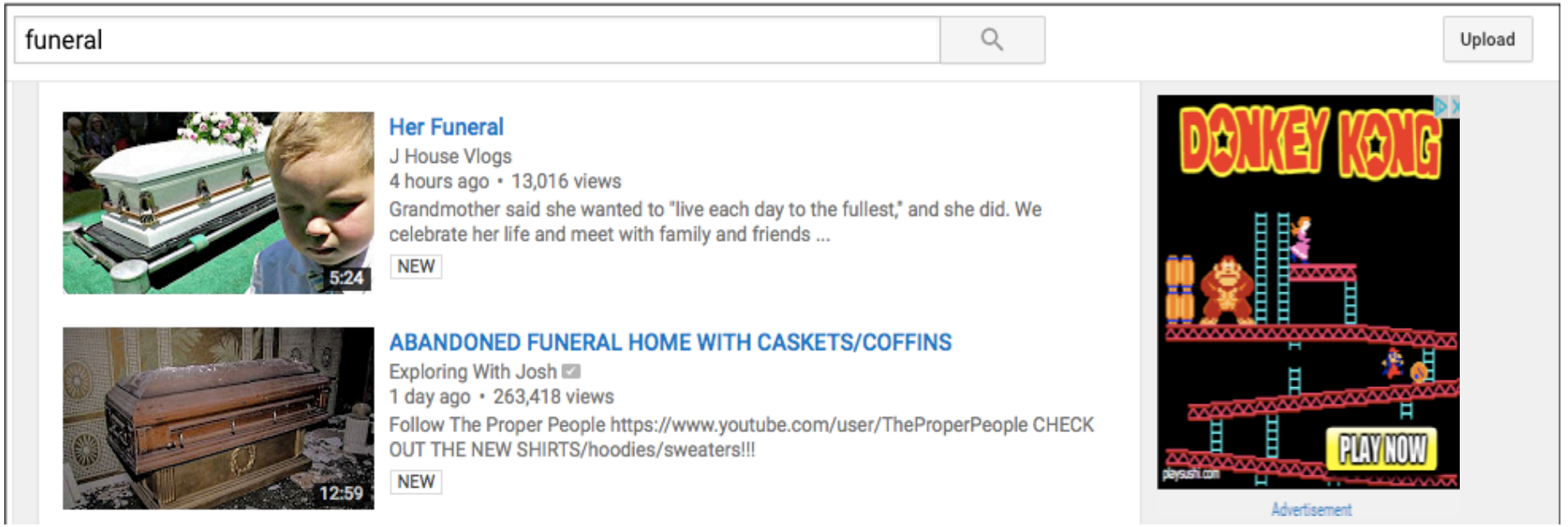
Joy

Anger

Disgust

Applications

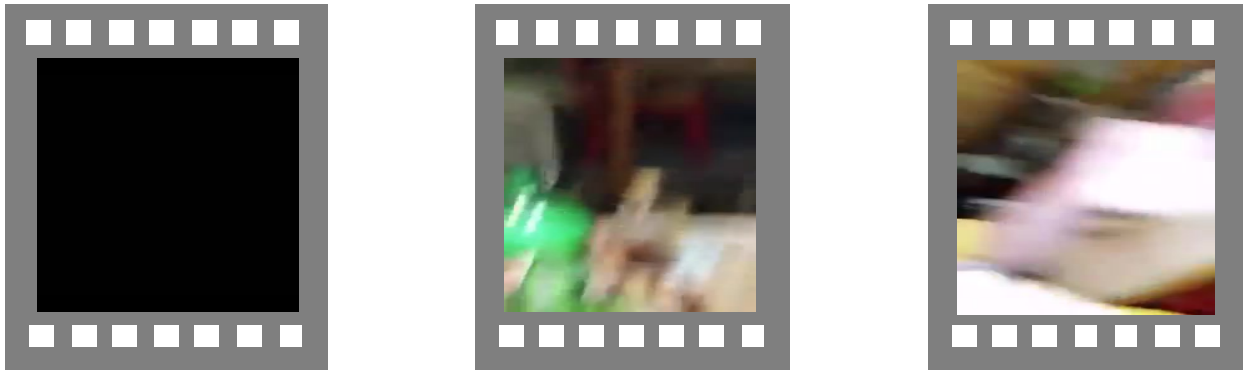
- Web Video Search
- Video Recommendation System
- Avoid Inappropriate Advertisement



Challenges

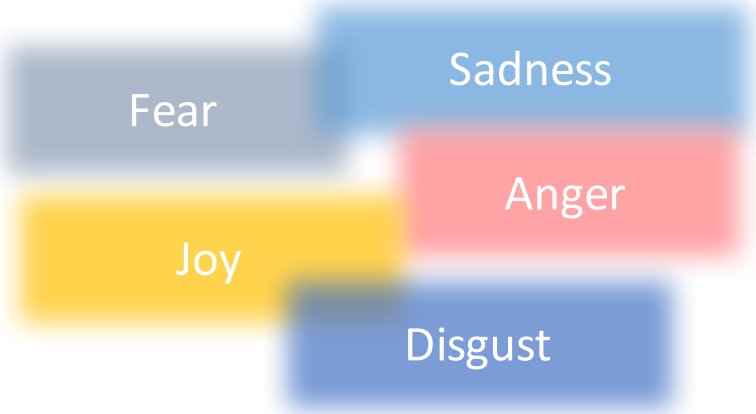
- Sparsely expressed in videos
  - Only 10% frames directly relate to dominant emotion

- Diverse content and variable quality



Knowledge Transfer

- Emotions are not atomic, indivisible, and categorical entities (Barrett 2006; Barrett 2011; Cunningham, Dunfield, & Stillman 2013)
- Difficult to label all emotions / affects: remorse, love, nostalgia, suspense, etc.



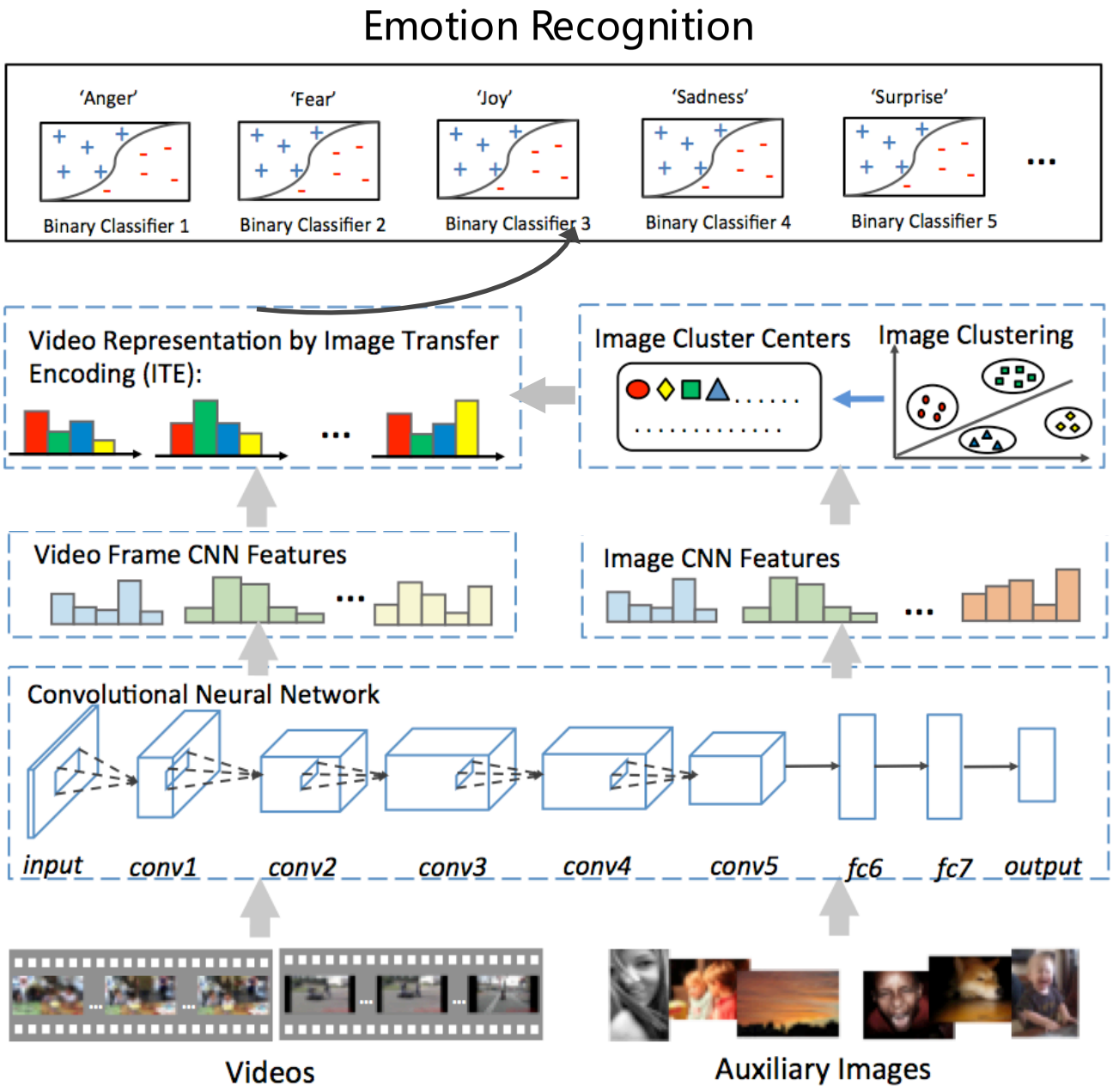
- Can we recognize emotions that are **not** in our training set?
- **Yes**, but we need to use knowledge transferred from external sources

Contributions

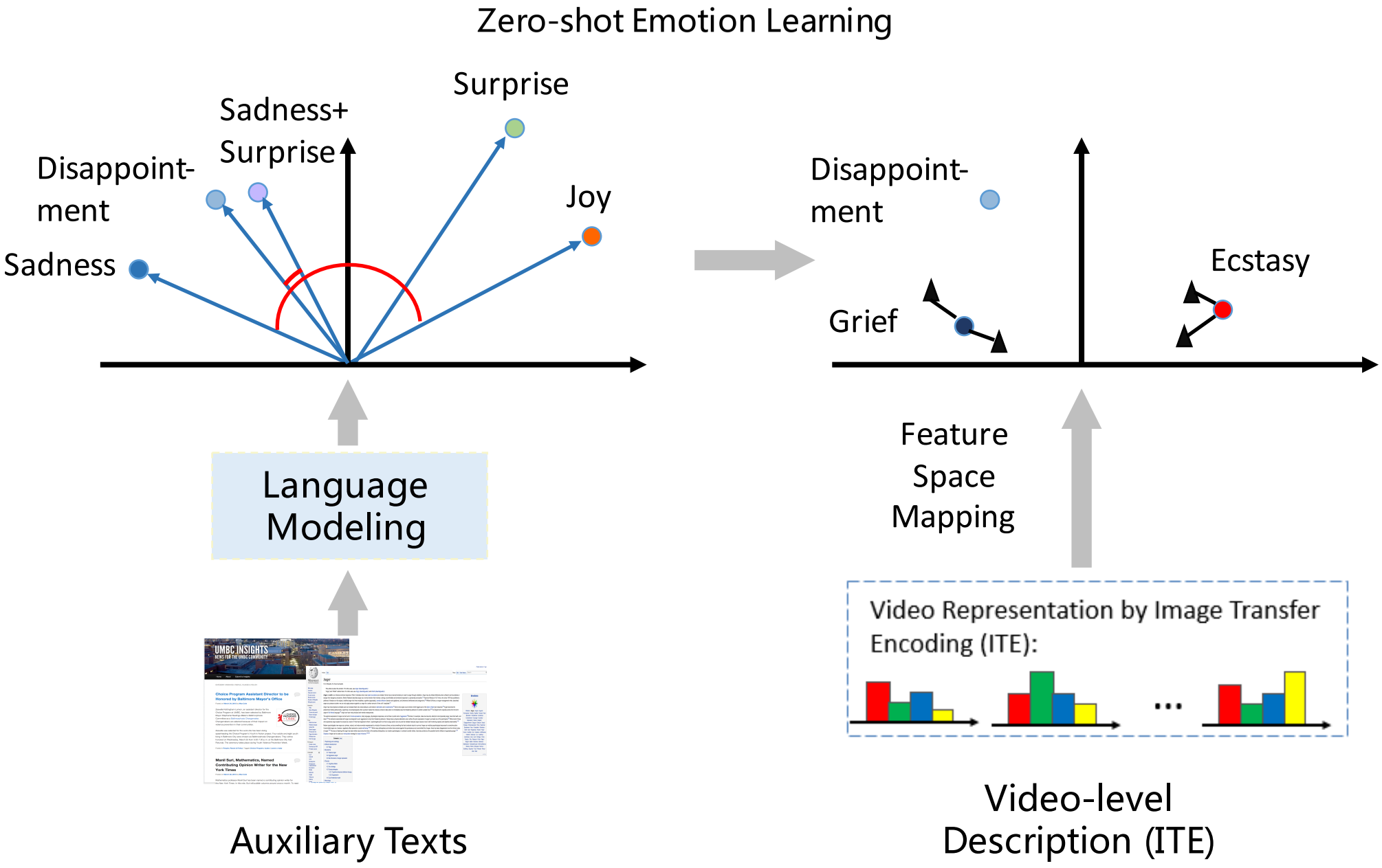
- Propose a novel auxiliary Image Transfer Encoding (ITE) process
- Investigate the effectiveness of features from different CNN architectures and layers
- Explore the complementarity of deep features with the existing visual and audio hand-crafted features

Approach

Framework

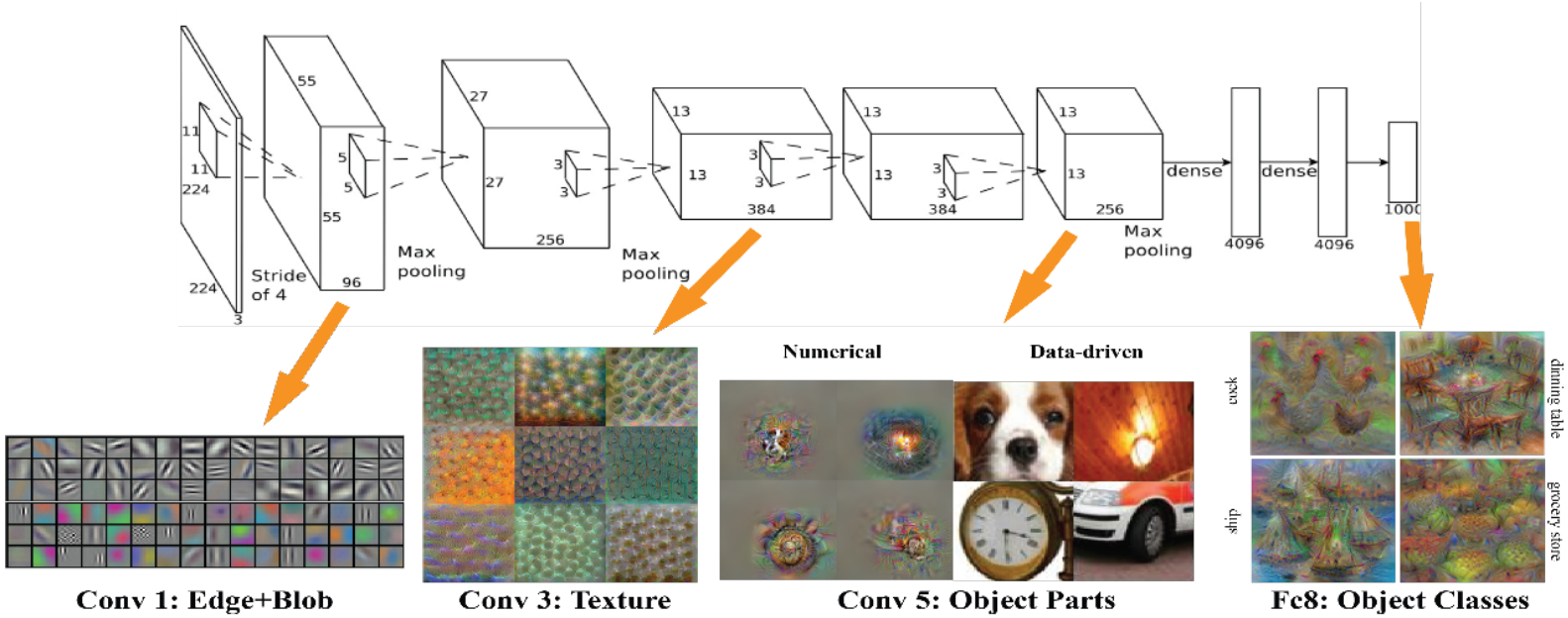

Emotion Recognition
'Anger'
'Fear'
'Joy'
'Sadness'
'Surprise'
Binary Classifier 1
Binary Classifier 2
Binary Classifier 3
Binary Classifier 4
Binary Classifier 5
Video Representation by Image Transfer Encoding (ITE):
Image Cluster Centers
Image Clustering
Video Frame CNN Features
Image CNN Features
Convolutional Neural Network
input
conv1
conv2
conv3
conv4
conv5
fc6
fc7
output
Videos
Auxiliary Images

Framework


Zero-shot Emotion Learning
Surprise
Sadness+
Surprise
Disappoint-
ment
Joy
Sadness
Disappoint-
ment
Ecstasy
Grief
Feature
Space
Mapping
Language
Modeling
Video Representation by Image Transfer
Encoding (ITE):
Auxiliary Texts
Video-level
Description (ITE)

Deep Neural Architectures

- Different deep architectures : AlexNet, VGG-16/VGG-19, GoogLeNet
- Layer-wise features of deep architecture



General → Specific

Deep Neural Architectures

- Concatenation of different layer features

- Complementarity of CNN with hand-crafted features

Visual

DenseSIFT

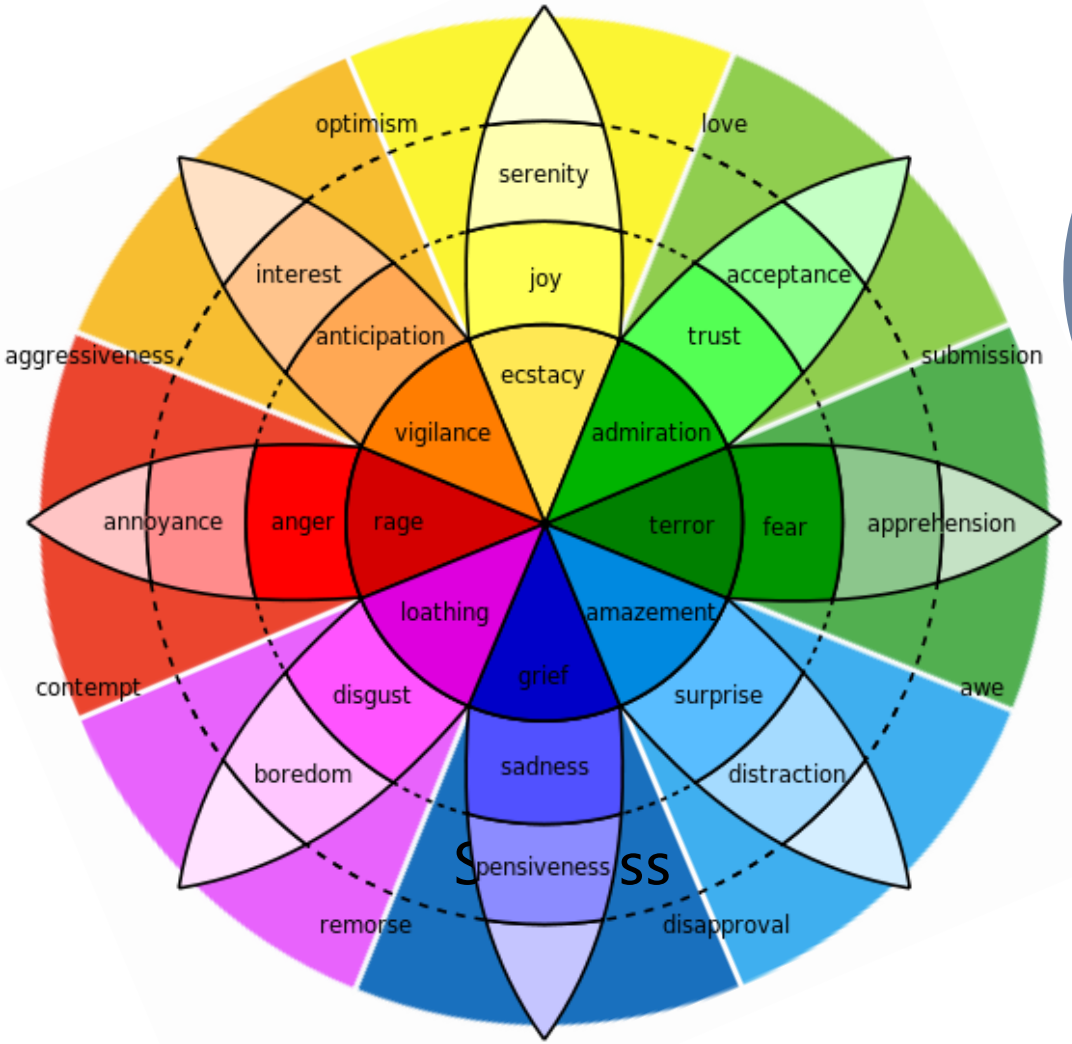
Audio

MFCC

Experiments

Our Datasets

The YouTube Dataset



Plutchik's Wheel

Dataset

The Ekman-6 Dataset

Category	Num
Anger	225
Disgust	239
Fear	287
Joy	305
Sadness	221
Surprise	360
Ave. duration	**112s**

Dataset

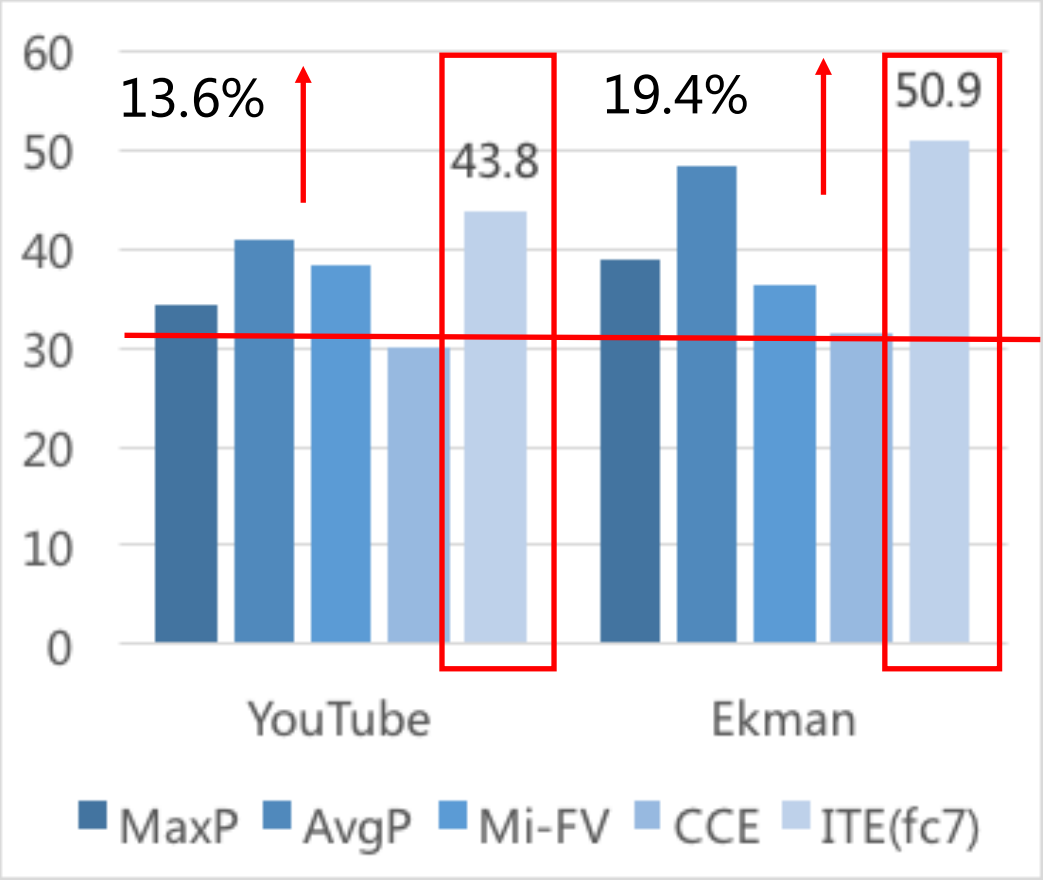
- Auxiliary images
  - **110K** images of Adjective-Noun Pairs (ANPs) *(Borth, 2013)*

Adorable Smile

Crying baby

- Auxiliary texts
  - **7** billion words
  - Strict definition about emotion related words

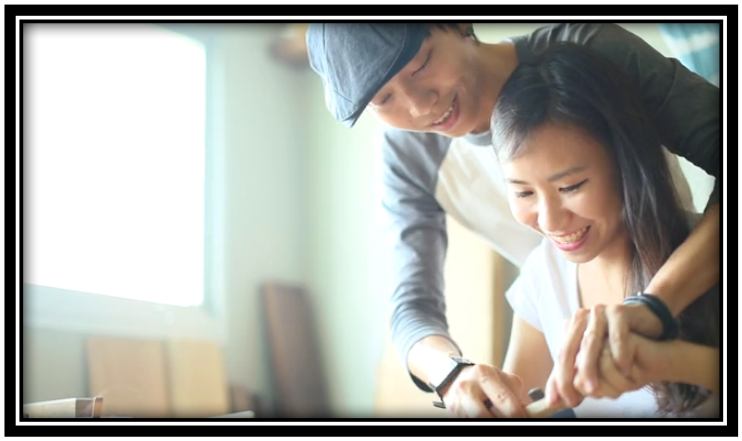
Supervised Emotion Recognition



- Compared with state-of-art multi-instance learning method
- ITE outperforms the four methods on both datasets
- Auxiliary image dataset create better video-level feature representations

Supervised Emotion Recognition

Success Case



Joy
(Happy couple)

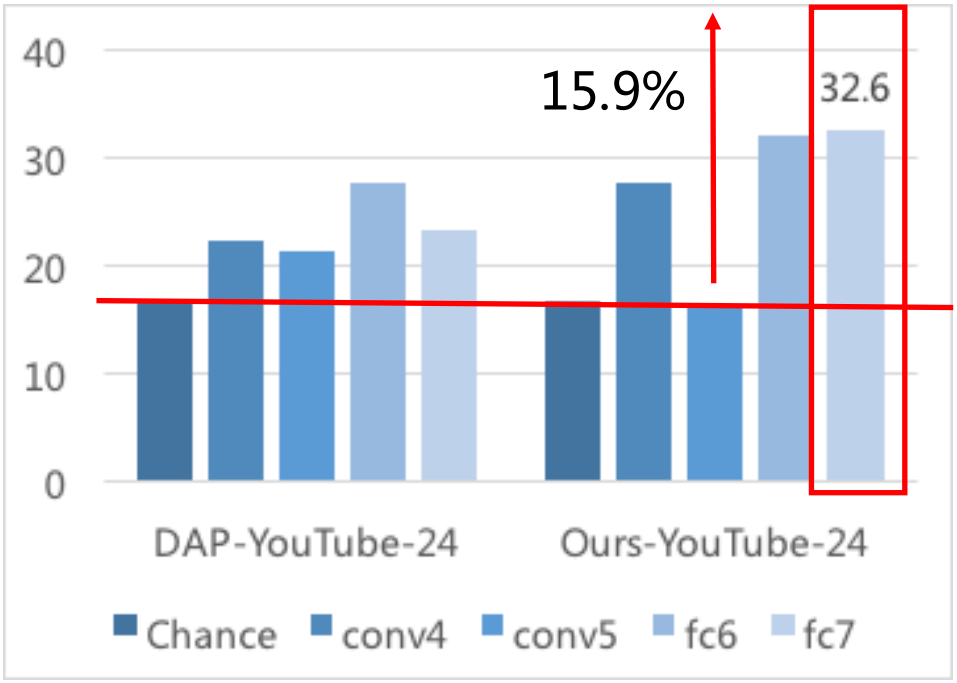
Bright
Smile

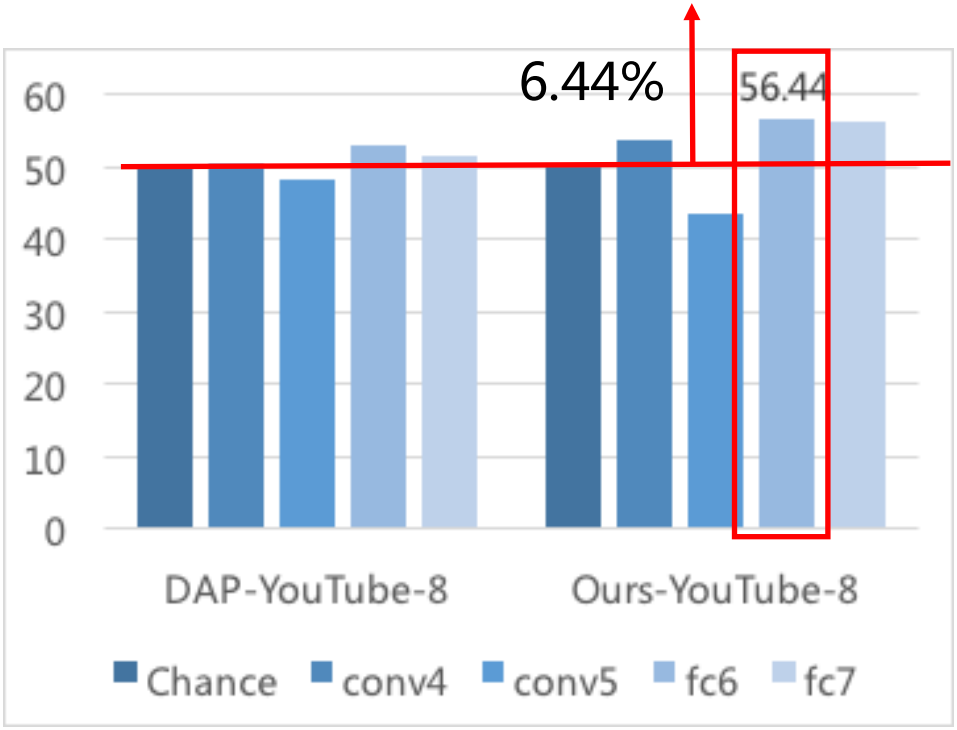
Failure Case

Joy
(Sadness memorial)

Bright
Flowers

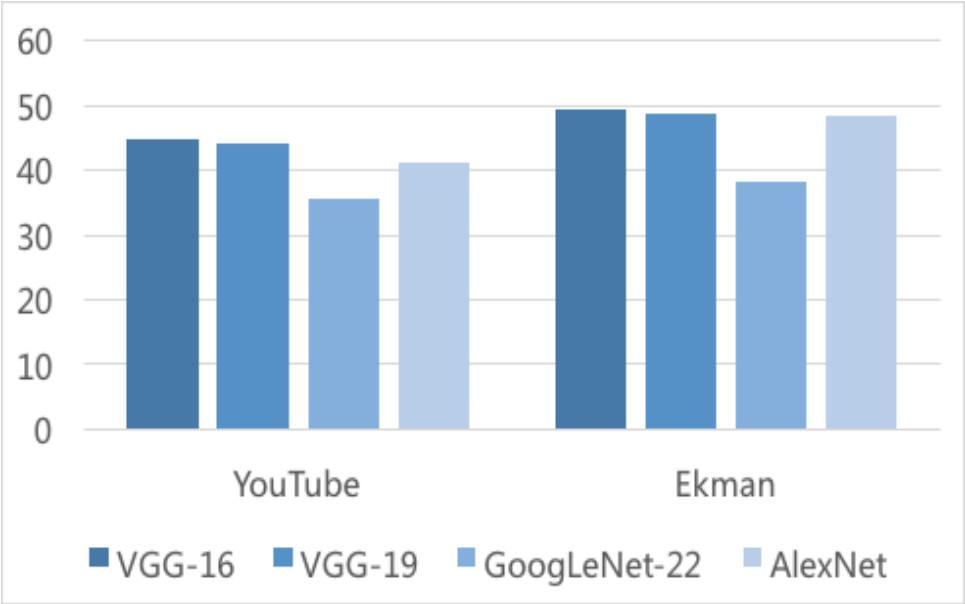
Zero-shot Emotion Recognition

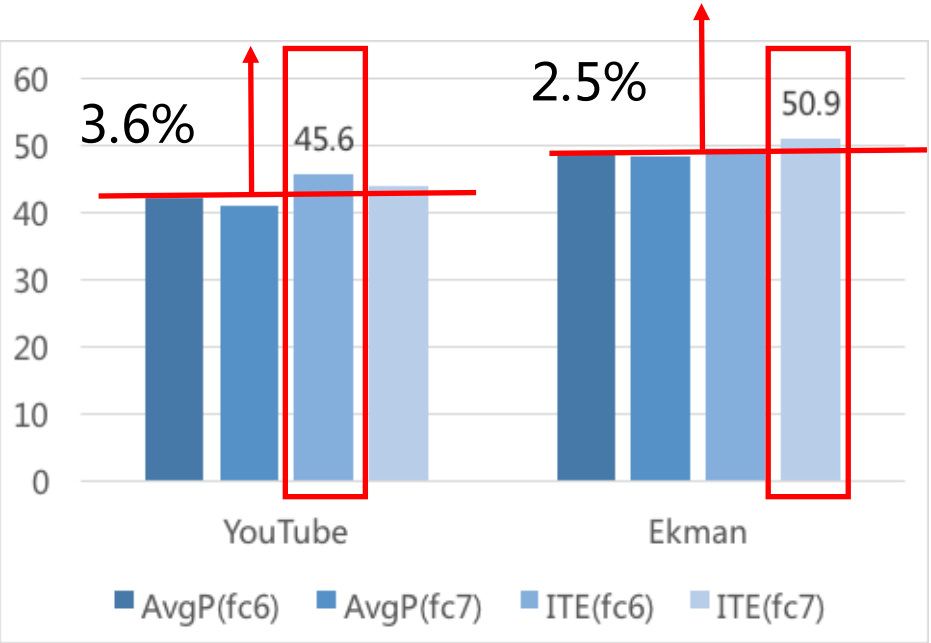




- Features of fully connected layers (fc6 and fc7) are generally more favorable for zero-shot emotion recognition than those of convolutional layers.
- Finer-grained variant set of auxiliary emotions can enable better zero-shot learning.

Validate Deep Architecture

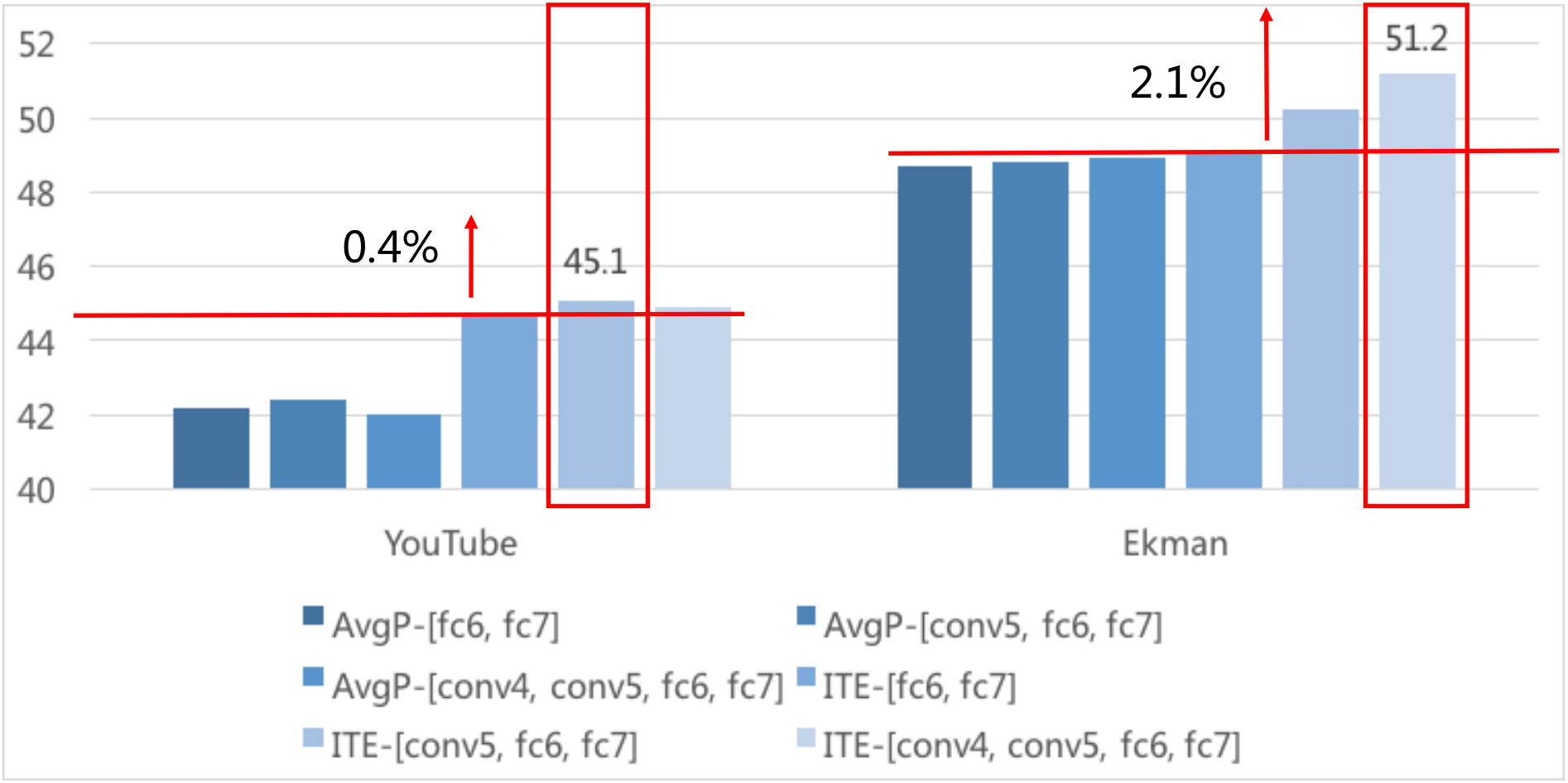




- Different deep architecture
- Layer-wise features of deep architecture

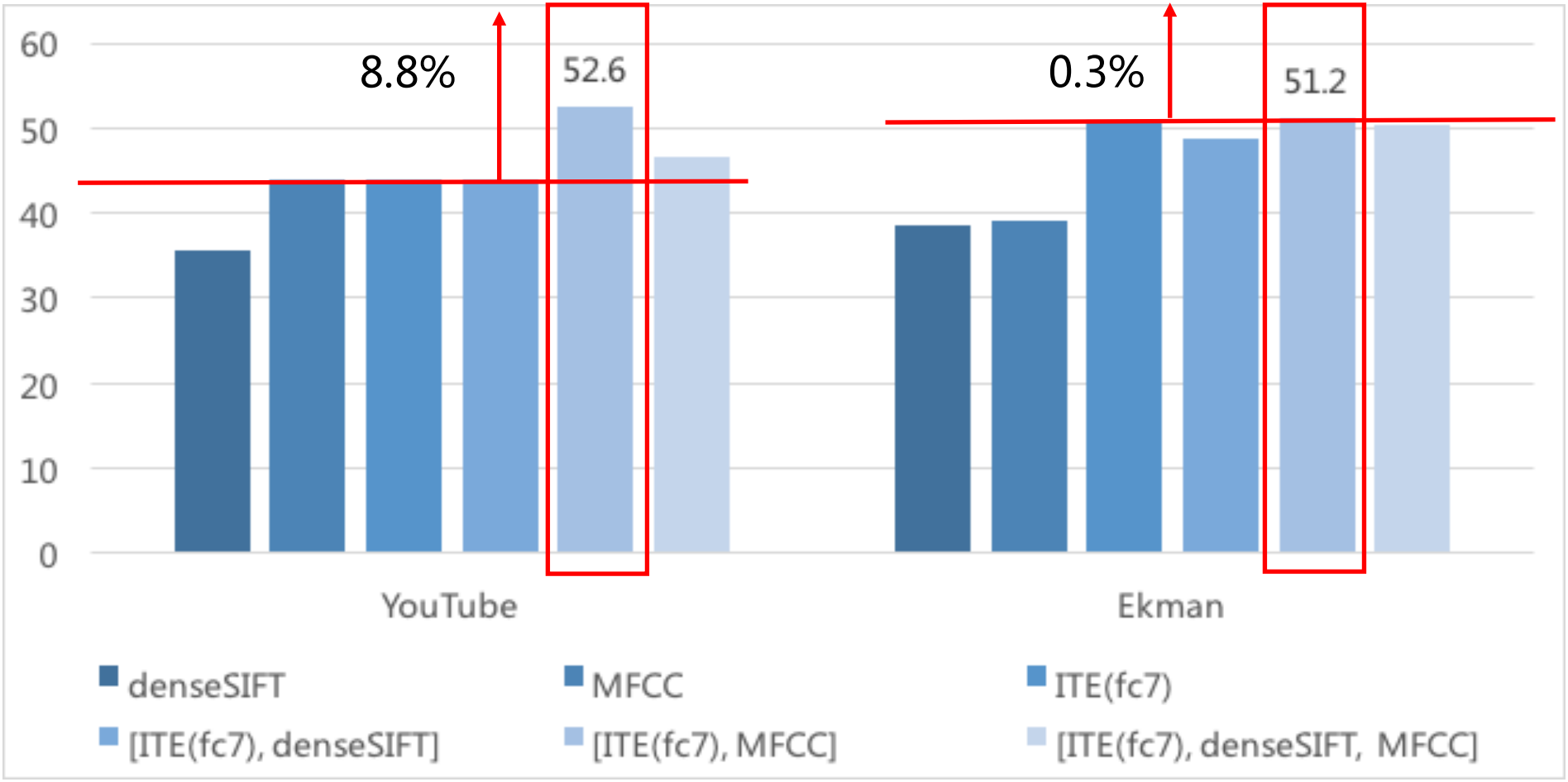
Validate Deep Architecture

- Feature Complementarity



Validate Deep Architecture

- Feature Complementarity



Conclusions

Conclusions

- What is more useful for emotion recognition?
  - Supervised emotion recognition
    - Auxiliary image information
    - VGG / AlexNet
    - Fully connected layers
    - Audio Features
  - Zero-shot emotion recognition
    - Auxiliary text information
    - More variant set emotions

Q & A

References

- *P. Ekman. Universals and cultural differences in facial expressions of emotion. Nebrasak Symposium on Motivation, 19:207–284, 1972.*
- *K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman. Return of the devil in the details: Delving deep into convolutional nets. In BMVC, 2014.*
- *P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. LeCun. Overfeat: Integrated recognition, localization and detection using convolutional networks. In ICLR, 2014.*
- *Q. You, J. Luo, H. Jin, and J. Yang. Robust image sentiment analysis using progressively trained and domain transferred deep networks. In AAAI, 2015.*
- *S. Andrews, I. Tsochantaridis, and T. Hofmann. Support vector machines for multiple-instance learning. In NIPS, 2003.*
- *Z.-H. Zhou and M.-L. Zhang. Solving multi-instance problems with classifier ensemble based on constructive clustering. Knowledge and Information Systems, 11(2):155–170, 2007.*
- *Y.-G. Jiang, B. Xu, and X. Xue. Predicting emotions in user-generated videos. In AAAI, 2014.*
- *D. Borth, R. Ji, T. Chen, T. M. Breuel, and S.-F. Chang. Large-scale visual sentiment ontology and detectors using adjective noun pairs. In ACM MM, 2013.*