

Video Emotion Recognition with Transferred Deep Feature Encodings

Baohan Xu¹, Yanwei Fu^{1,2}, Yu-Gang Jiang¹, Boyang Li² and Leonid Sigal²

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



復旦大學
FUDAN UNIVERSITY



Disney Research Pittsburgh

Introduction

Motivation

YouTube vimeo Dailymotion twitch ...

300 hours of video uploaded to YouTube every minute



Fear

Joy

Sadness

Anger

Disgust



Applications

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

A screenshot of a YouTube search results page for the query "funeral". The search bar at the top contains the word "funeral". Below the search bar, there are two video thumbnails. The first video thumbnail on the left shows a stack of ornate wooden caskets. The title of this video is "ABANDONED FUNERAL HOME WITH CASKETS/COFFINS" and it is uploaded by "Exploring With Josh" (verified). It has 1 day ago views and 263,418 views. The second video thumbnail on the right shows a person's face looking down at a casket. The title of this video is "Her Funeral" and it is uploaded by "J House Vlogs". It was posted 4 hours ago with 13,016 views. A description for this video reads: "Grandmother said she wanted to "live each day to the fullest", and she did. We celebrate her life and meet with family and friends ..." A "NEW" badge is visible next to the video. At the bottom of the screen, there is a horizontal advertisement for a game.

funeral

Search

Upload

ABANDONED FUNERAL HOME WITH CASKETS/COFFINS

Exploring With Josh

1 day ago • 263,418 views

Follow The Proper People <https://www.youtube.com/user/TheProperPeople> CHECK OUT THE NEW SHIRTS/hoodies/sweaters!!!

12:59

NEW

Her Funeral

J House Vlogs

4 hours ago • 13,016 views

Grandmother said she wanted to "live each day to the fullest", and she did. We celebrate her life and meet with family and friends ...

NEW

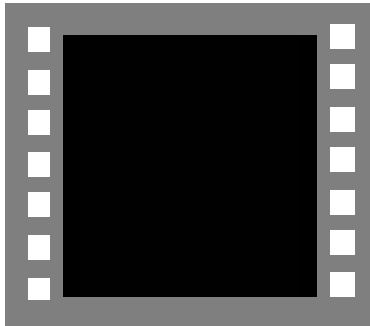
PLAY NOW

Advertisement

DONKEY KONG

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

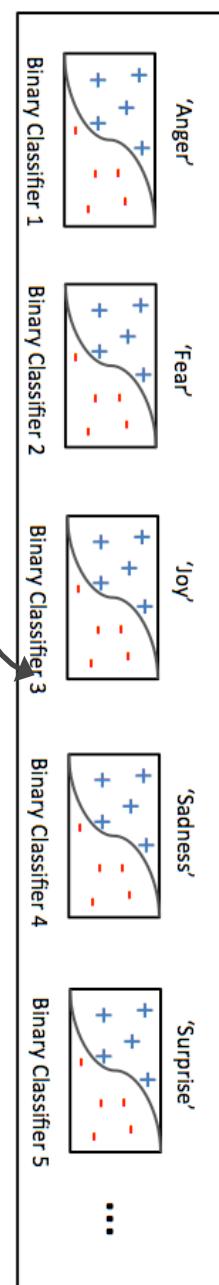
Contributions

- 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

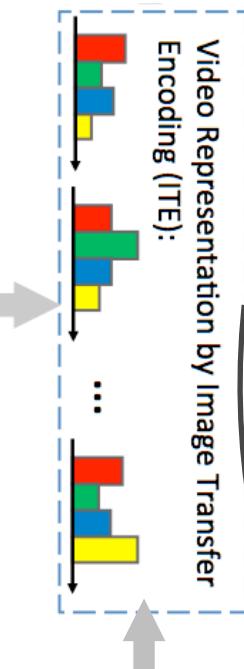
Approach

Framework

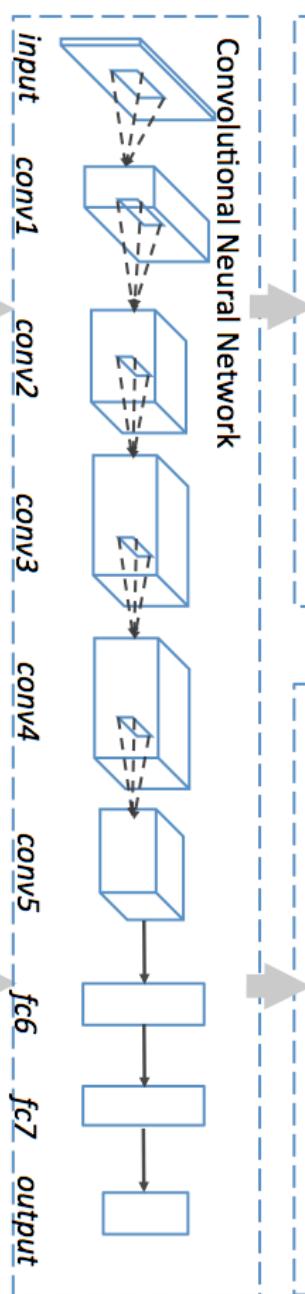
Emotion Recognition



Video Representation by Image Transfer
Encoding (ITE):



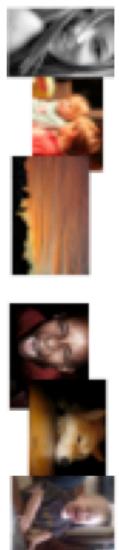
Convolutional Neural Network



Videos

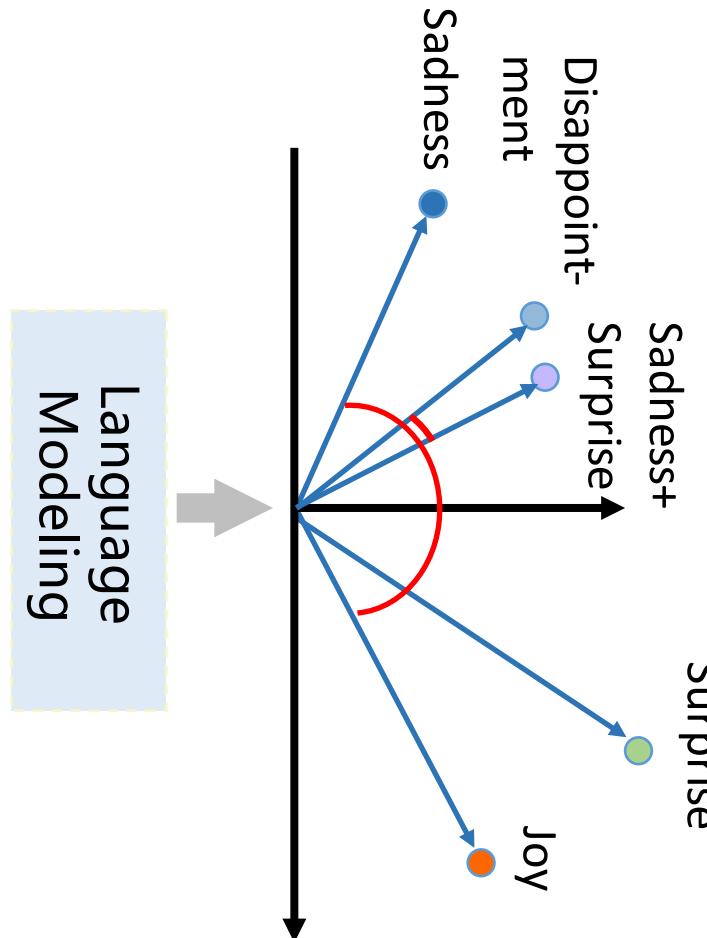


Auxiliary Images



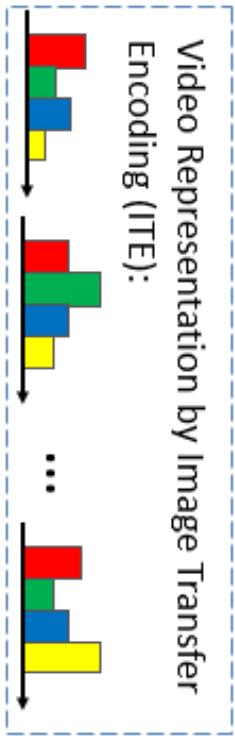
Framework

Zero-shot Emotion Learning



Feature Space Mapping

Video Representation by Image Transfer Encoding (ITE):



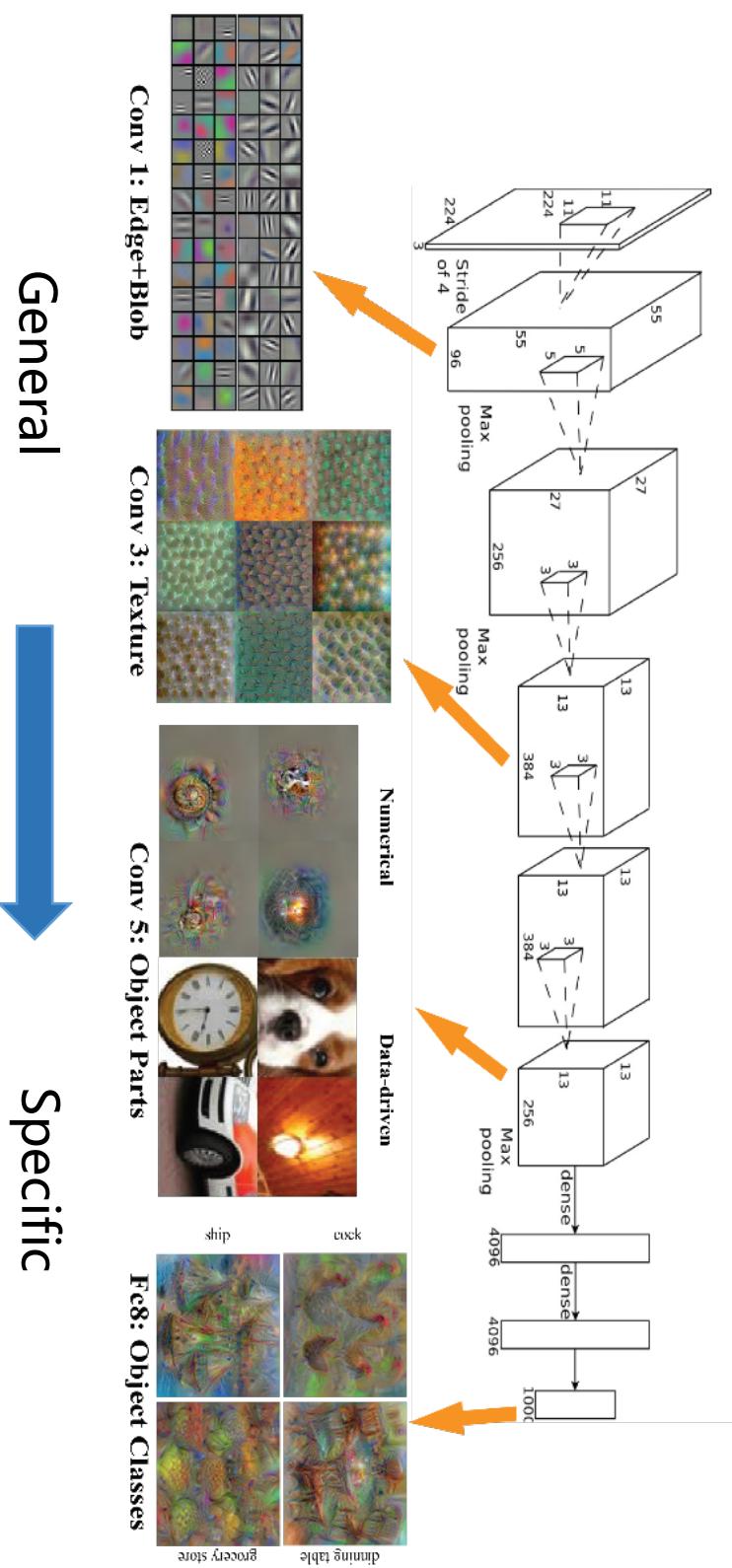
Auxiliary Texts

Video-level Description (ITE)

Deep Neural Architectures

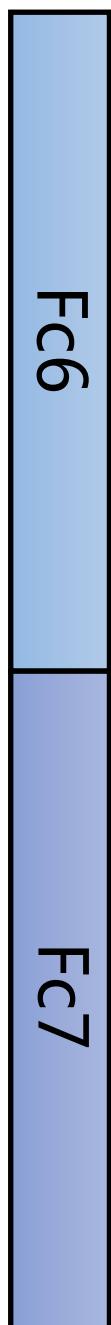
- Different deep architectures : AlexNet, VGG-16/VGG-19, GoogleNet

- Layer-wise features of deep architecture



Deep Neural Architectures

- Concatenation of different layer features



- Complementarity of CNN with hand-crafted features



Visual

DenseSIFT



Audio

MFCC

Fc7

DenseSIFT

MFCC

Experiments

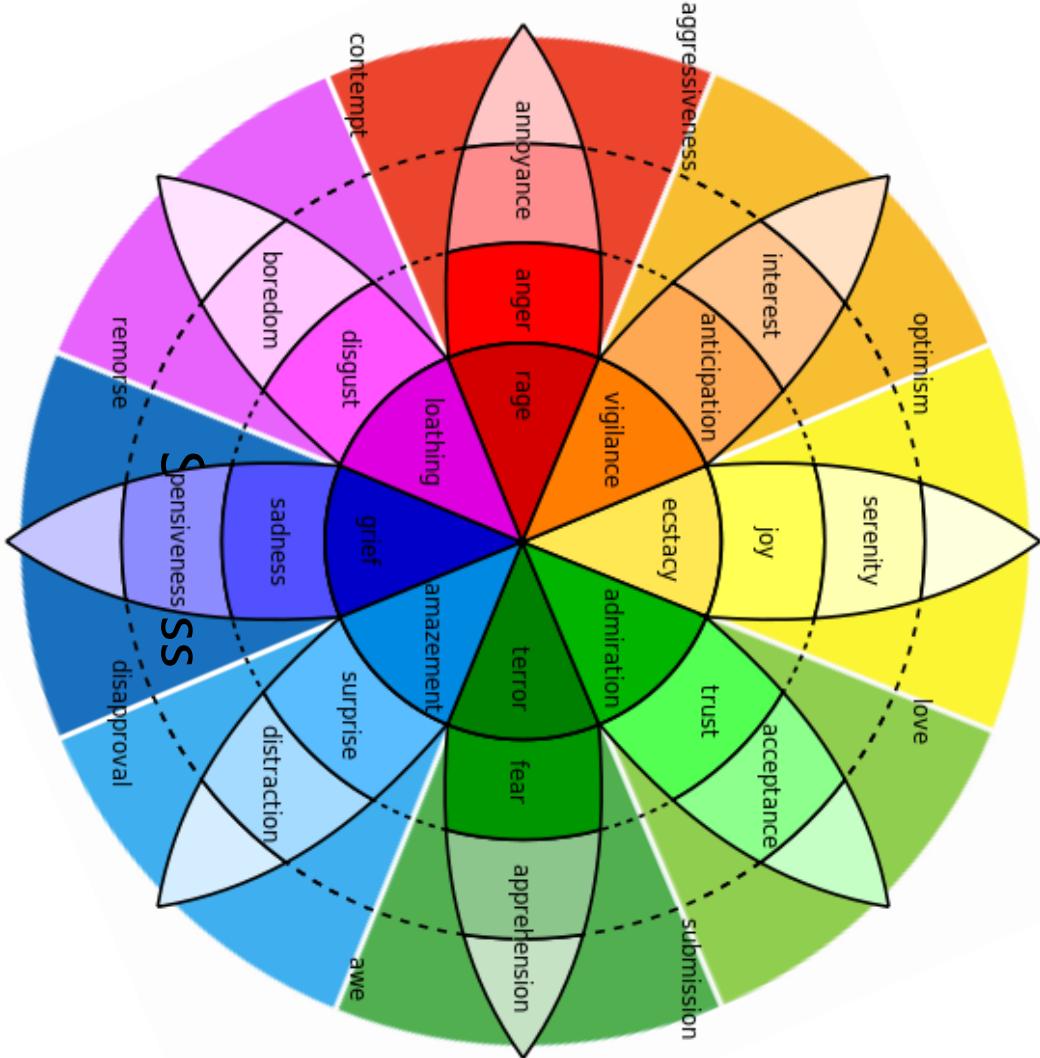
Our Datasets

The YouTube Dataset

24
fine-grained
emotions

8
emotions

1101
videos

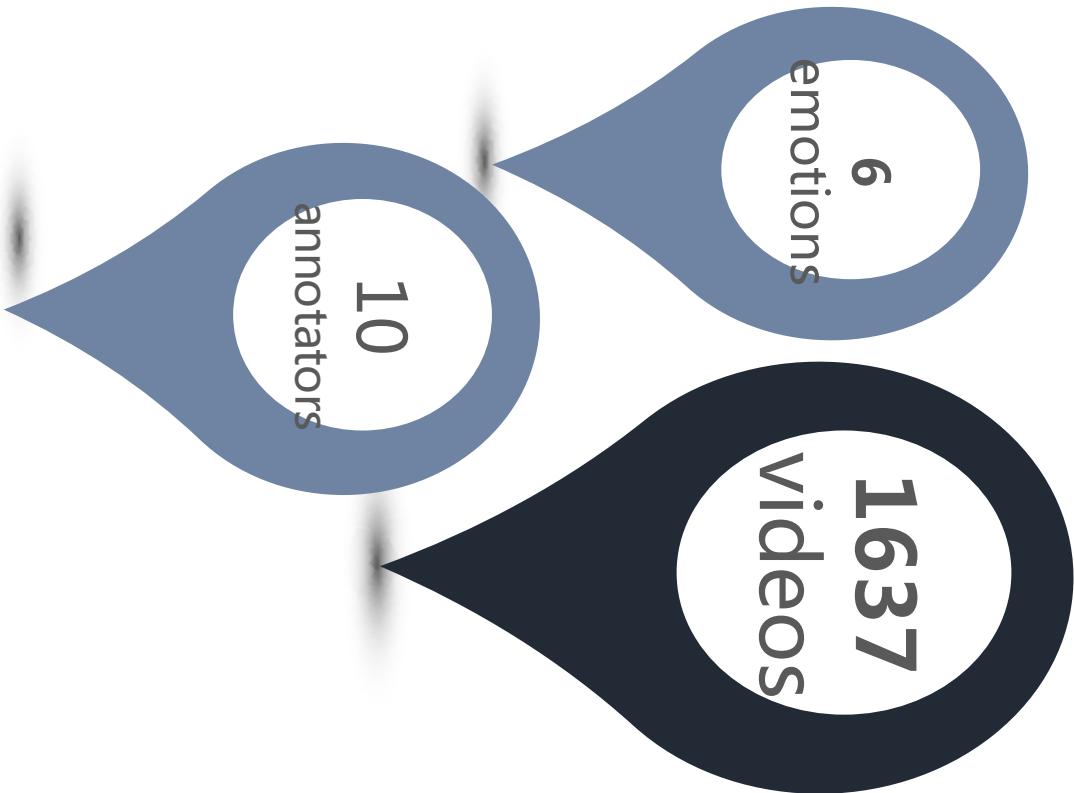


Plutchik's Wheel

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

WIKIPEDIA

The Free Encyclopedia

Article Talk

Anger

From Wikipedia, the free encyclopedia

Main page
Contents
Featured content
Current events
Random article
Donate to Wikipedia
Wikipedia store

For other uses, see *Anger* (disambiguation).

"*Wrath*" redirects here. For other uses, see *wrath* (disambiguation).

Anger or *wrath* is an intense emotional response. It is a normative response to a perceived provocation.^[1] Often, it indicates when learned tenderness to react to anger through retaliation. Anger is often a learned response to dangerous situations. Haymond Novaco of UC Irvine, who

WIKIPEDIA

The Free Encyclopedia

Article Talk

Sadness

From Wikipedia, the free encyclopedia

(Redirected from *Sad*)

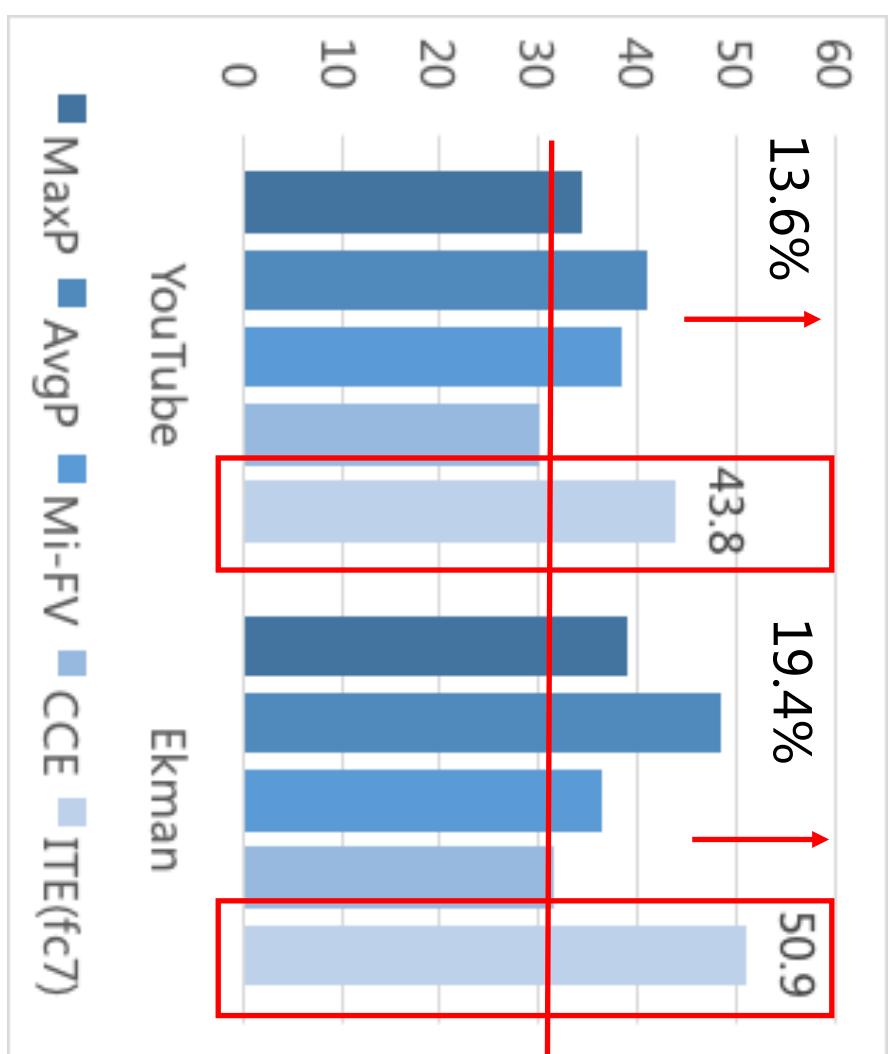
For the video game, see *Sadness* (video game).

"*Sad*" redirects here. For other uses, see *Sad* (disambiguation).

Sadness is an emotional pain associated with, or characterized by, feelings of悲哀 (disappointment and sorrow). An individual experiencing sadness may feel lonely, hopeless, or pessimistic. Others. An example of severe sadness is depression.

Main page
Contents
Featured content
Current events
Random article
Donate to Wikipedia
Wikipedia store

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



Failure Case



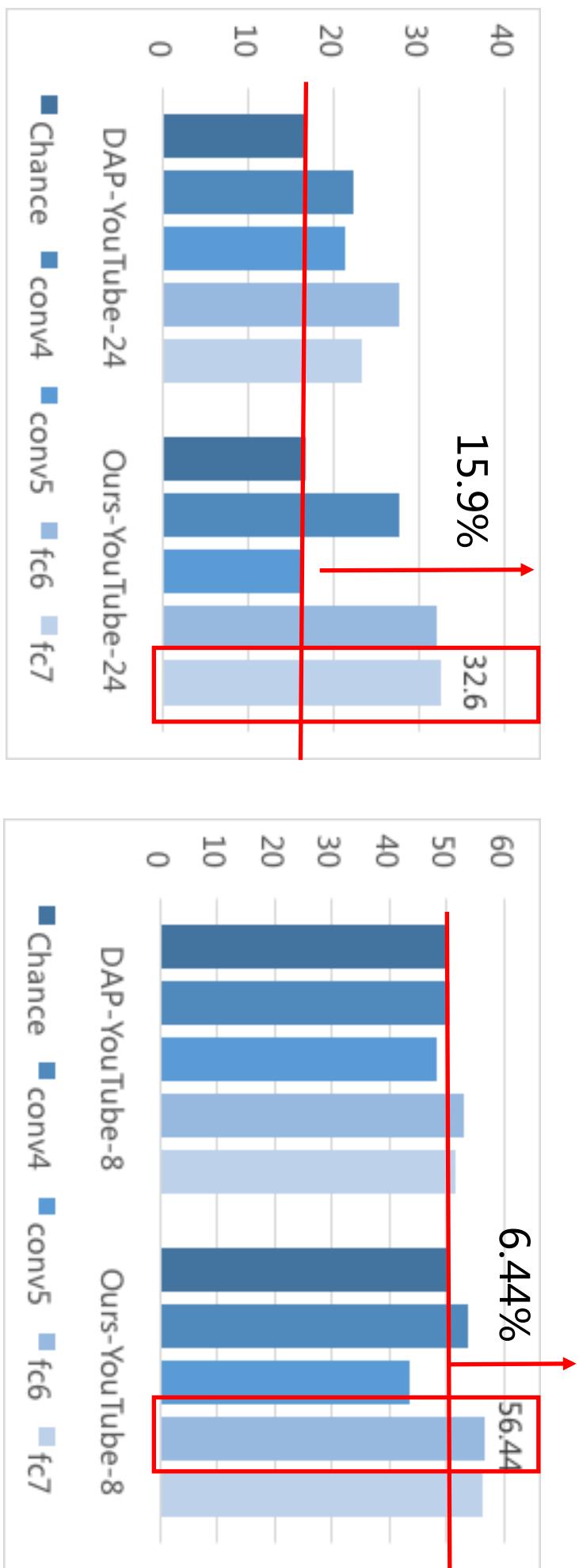
Joy
(Happy couple)

Joy
(Sadness memorial)

Bright
Smile

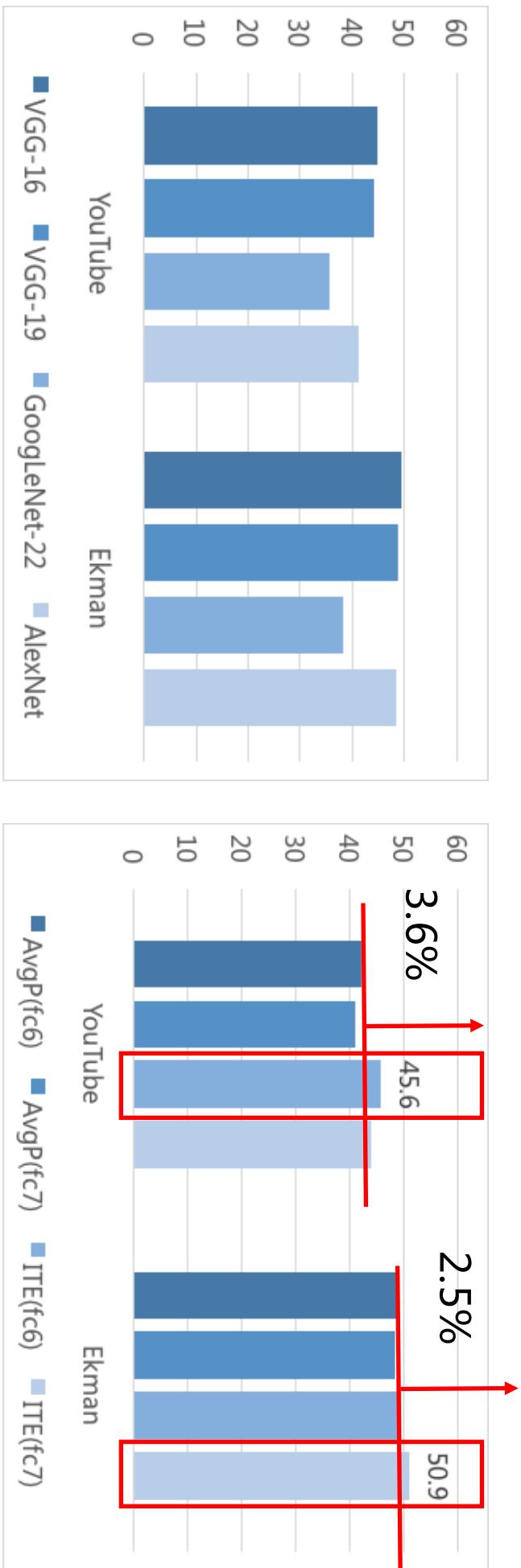
Bright
Flowers

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.

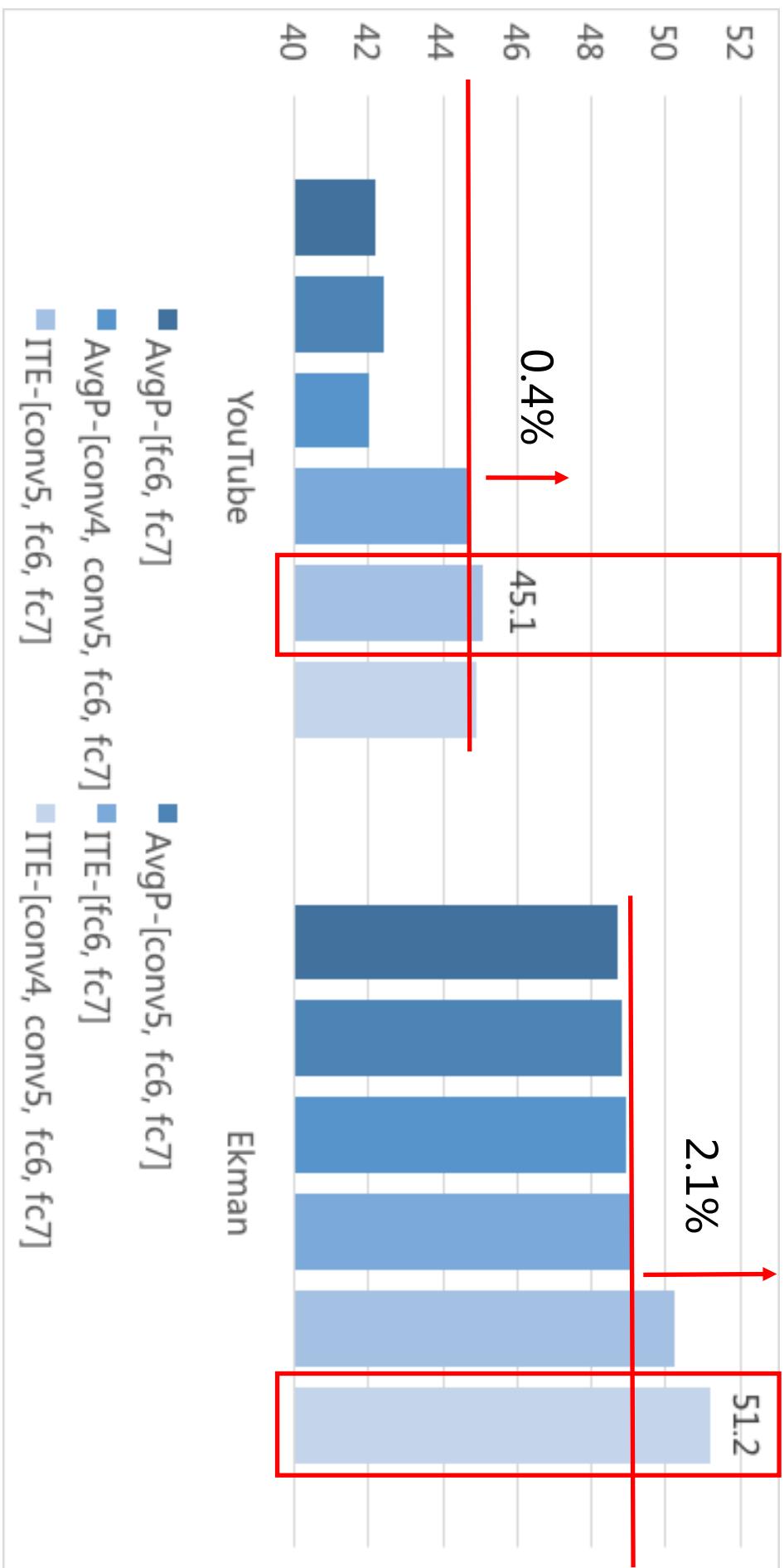
Validate Deep Architecture



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

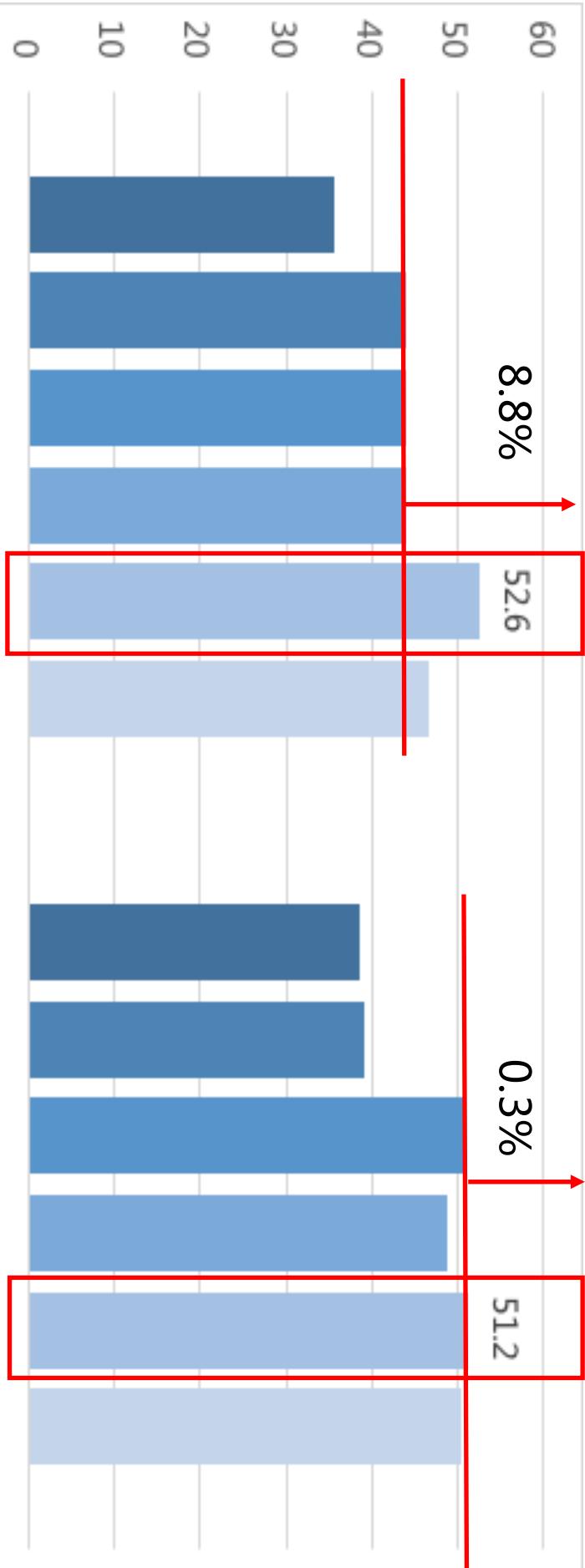
Validate Deep Architecture

- 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

THANKS

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. Tsachantaridis, 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.