国际人工智能会议
AAAI 2021 论文北京预讲会

Inferring Emotion from Large-scale Internet Voice Data: A Semi-supervised Curriculum Augmentation based Deep Learning Approach

Presentor: Suping Zhou

Suping Zhou, Jia Jia, Zhiyong Wu, Zhihan Yang, Yanfeng Wang, Wei Chen, Fanbo Meng, Shuo Huang, Jialie Shen, Xiaochuan Wang

Department of Computer Science and Technology

Tsinghua University, China

清華大學 Tsinghua University

Index

- Introduction
- Related works
- Ideas & contributions
- Methodology
- Experiments
- Conclusion

清華大學
Tsinghua University

1 Introduction

Inferring Emotion from Large-scale Internet Voice Data
Introduction

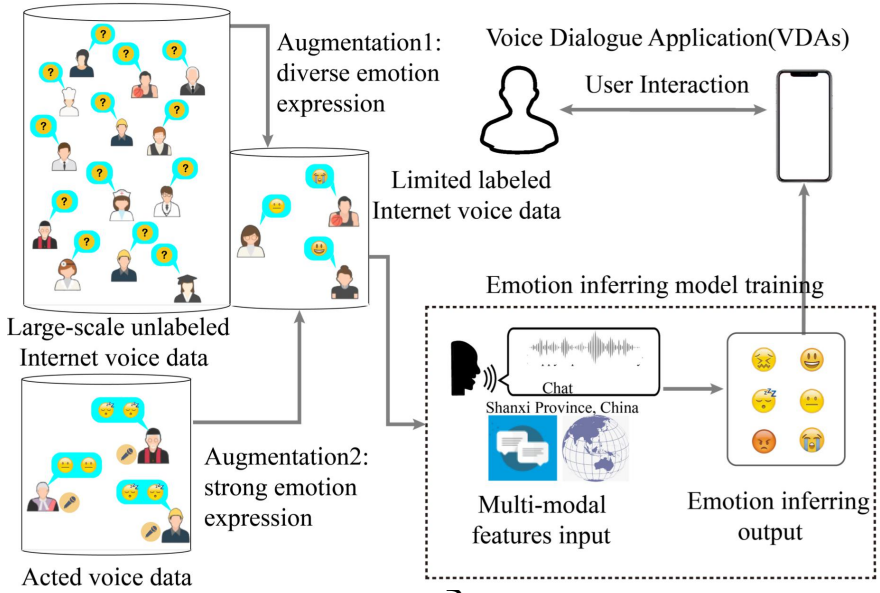
Tsinghua University

- Internet voice data
  - Internet voice data are users' speech queries from the Voice Dialogue Applications(VDAs), such as Siri, Sogou Voice Assistant
  - tremendous amounts of VDA users bring in diverse emotion expressions
  - weak and unbalanced emotion expressions.

- acted voice data
  - Traditionally, researches on speech emotion recognition are based on acted voice datasets, which have limited speakers but strong and clear emotion expressions.

Inferring Emotion from Large-scale Internet Voice Data
Introduction

- Inspired by the contrast between the large scale internet voice data from VDAs and the acted emotional voice dataset, can we use one's strengths to compensate for the other's weakness?



清華大學
Tsinghua University

2 Related works

Tsinghua University

- Inferring voice emotion
  - In terms of emotion analysis for voice, previous works have focused primarily on extracting effective features and utilizing diverse types of learning methods. (Neumann and Vu 2019)

  (Freitag et al. 2017)(Zhang et al. 2019)

  - However, all these researches mainly focused on inferring emotions from acted corpora data, few have been done to address the problem for real-world large-scale internet voice data with weak emotion expressions and tremendous uncertain speakers.
  - It is potential to transfer the emphasis on emotion recognition in the wild and assist this work through the augmentation of acted corpus.

Tsinghua University

- Curriculum learning
  - Curriculum learning is training strategy to learn from simple to complex and proved to achieve great improvements in generalization and speed of convergence. (Bengio et al. 2009)
  - It is natural to apply curriculum learning to emotion recognition since we learn to perceive emotions gradually from infantry to adulthood.
  - Previous work on speech emotion recognition has utilize curriculum learning to solve the problem of Crowd-sourced Labels and achieve improvements (Lotfian and Busso 2019).

清華大學 Tsinghua University

- Semi-supervised learning
  - Autoencoders have always been a common way to make better use of unlabeled data in speech emotion recognition.(Deng et al. 2017)(Jia et al. 2018). Some Generative and Adversarial Networks(Semi-VAE (Zhou et al. 2018b), DCGAN(Chang and Scherer 2017), ADDoG(Gideon, McInnis, and Provost 2019) are also utilized to make improvements.

  - (Berthelot et al. 2019) propose a hybrid method named Mix- match which combines several ideas and components from the current dominant paradigms for SSL.

清華大學
Tsinghua University

3 Ideas & contributions

Tsinghua University

Challenges

1)how to effectively leverage acted voice dataset with strong and clear emotion expressions to enhance internet voice data?

2)how to utilize large-scale unlabeled data with diverse user emotion expressions to augment few labeled data.

Tsinghua University

We proposed A Semi-supervised Curriculum Augmentation based Deep Learning Approach

1、 Curriculum learning based epoch-wise training strategy

2、 Multi-path Mix-match Multimodal Deep Neural Network(MMMD)

清華大學
Tsinghua University

4 Methodology

Workflow of our framework

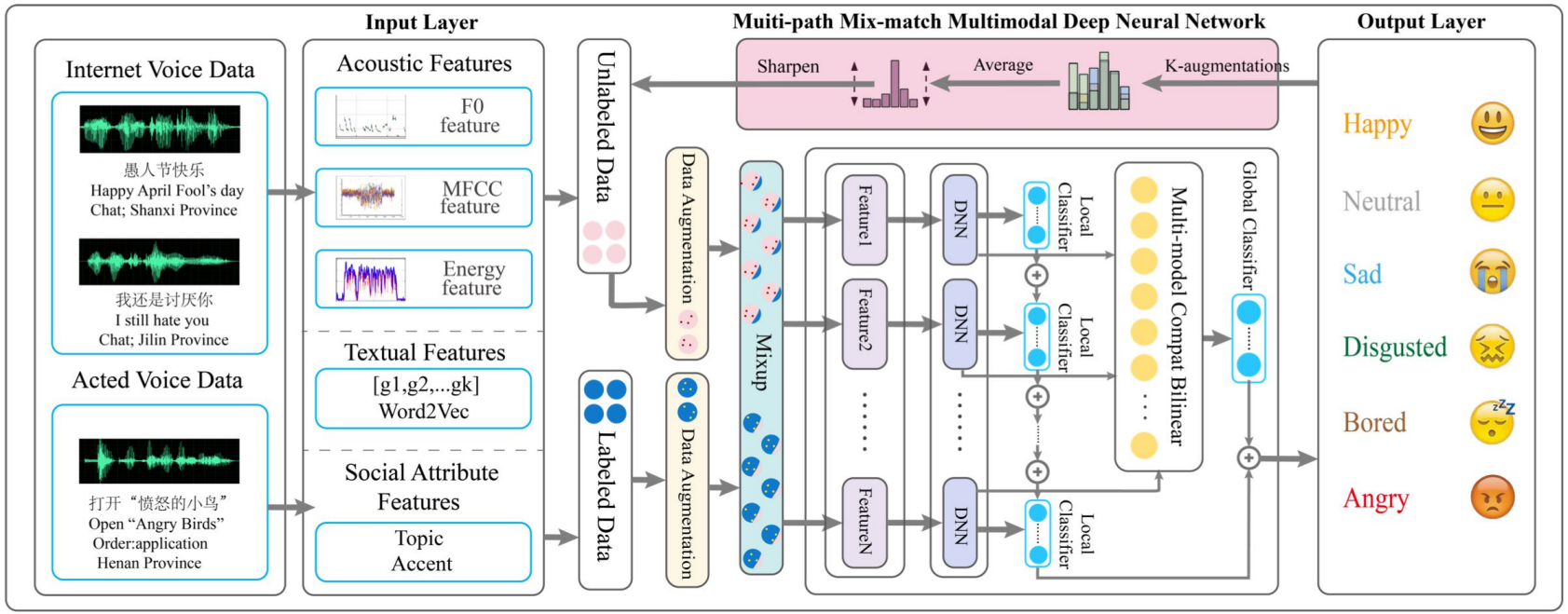
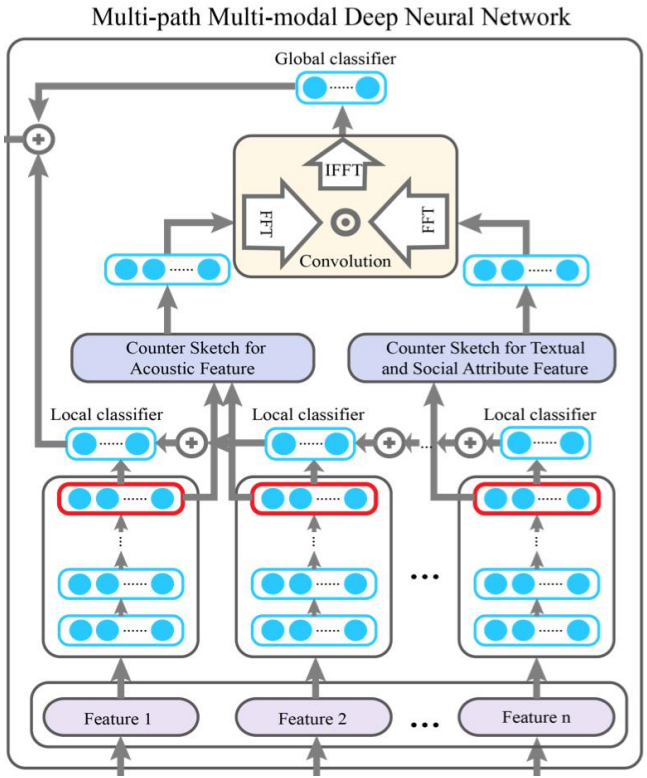


Figure 2: The workflow of our framework.

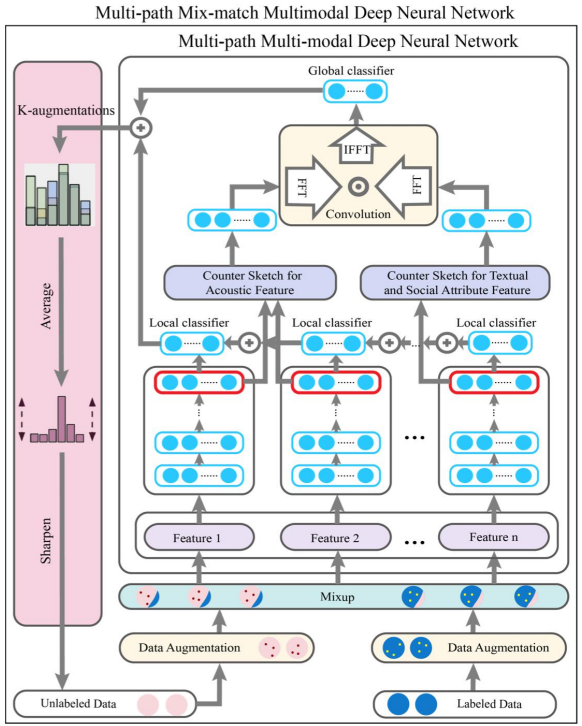
Multi-path Multimodal Deep Neural Network(MMD)



- Multi-path solution
- Multimodal Compact Bilinear pooling
- to model the complex intra-modality relationship which balances both the independencies and dependencies of multi-modal features.

Semi-supervised MMMD

Multi-path Mix-match Multimodal Deep Neural Network(MMMD)



1、Augmentation

Gaussian noise

2、Mixup

$$\lambda \sim Beta(0.75, 0.75)$$
$$\lambda' = max(\lambda, 1-\lambda)$$
$$x' = \lambda' x_1 + (1-\lambda')x_2$$
$$p' = \lambda' p_1 + (1-\lambda')p_2$$

3、Entropy minimization

$$Sharpen(p, T) = p^{\frac{1}{T}} / \sum_{j=1}^{L} p_j^{\frac{1}{T}}$$

MixMatch(Berthelot et al. 20

Epoch-wise Training Strategy

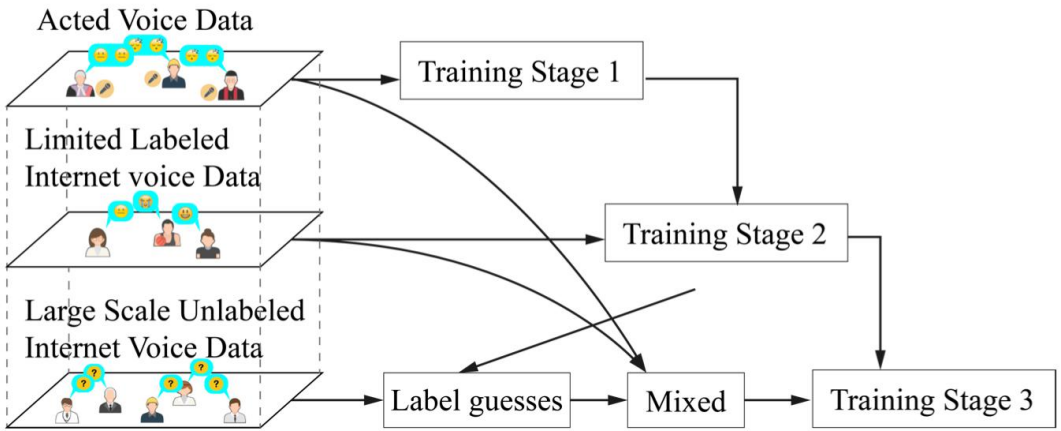


Figure 4: Learning Strategy in an epoch.

$$X' = Augment(X)$$

$$V_t^{U\prime} = Augment(V_t^U)$$

$$W = Shuffle(Concatenate(X', V_t^{U\prime}))$$

$$X'' = MixUp(X', W_1)$$

$$V_t^{U\prime\prime} = MixUp(V_t^{U\prime}, W_2)$$

$$\mathcal{L}_{Ve} = \frac{1}{|V_e|} \sum_{x,p \in V_e} H(p, P(y|x;\theta))$$

$$\mathcal{L}_{V_t^L} = \frac{1}{|V_t^L|} \sum_{x,p \in V_t^L} H(p, P(y|x;\theta))$$

$$\mathcal{L}_X = \frac{1}{|X''|} \sum_{x,p \in X''} H(p, P(y|x;\theta))$$

$$\mathcal{L}_U = \frac{1}{L|V_t^{U\prime\prime}|} \sum_{u,q \in V_t^{U\prime\prime}} ||q - P(y|u;\theta)||_2^2$$

$$\mathcal{L} = \mathcal{L}_{Ve} + \mathcal{L}_{V_t^L} + \mathcal{L}_{\mathcal{X}} + \lambda_U \mathcal{L}_U$$

5 Experiment

Dataset1

Experiment

- Internet voice Dataset.（Chinese）
  - a corpus of voice data from Sogou Voice Assistant recorded in 2013 (SVAD13)
  - Speech information , speech-to-text information, social attributes
  - 50,000 unlabeled utterances, 2946 manually labeled utterances
  - Emotion category:
    - Neutral: 49.3%, Happiness: 16.5%, Disgust: 11.0%, Boredom: 8.7%, Anger: 9.8% and Sadness: 4.6%.
  - 5- fold cross validation

Dataset2

Experiment

Tsinghua University

- Acted voice Dataset.（Chinese）
  - Speech information , speech-to-text information, social attributes
  - Only when two annotators and the volunteer who read the utterance have same opinion about the emotion labeling, the utterance and its label will be adopted.
  - 2397labeled utterances
  - Emotion category:
    - *Neutral:* 14.0%, Happiness: 23.8%, Disgust: 17.4%, Anger: 22.6% and Sadness: 22.2%. 5- fold cross validation

Dataset3

Experiment

Tsinghua University

- Public dataset **IEMOCAP** （English）
  - Speech information and text information
  - Emotion category:

Emotion	Happy	Anger	Sad	Neutral	Total
Utterances	1636	1103	1084	1708	5531
Proportion(%)	29.6	19.9	19.6	30.9	-

  - 10-fold leave-one-speaker-out(LOSO) cross-validation

Feature Extraction

Experiment

Tsinghua University

- Acoustic feature
  - openSMILE toolkit
  - 1,582 statistic acoustic features
  - the INTERSPEECH 2010 Paralinguistic Challenge

Low level Descriptors (LLDs)
PCM loudness
MFCC [0-14]
log Mel Frequency Band [0-7]
Line Spectral Pairs (LSP) Frequency [0-7]
F0 by sub-harmonic summation
F0 Envelope
Voicing probability
Jitter local
Jitter difference of difference of periods (DDP)
Shimmer local

Table 2. 38-dimensional frame-level acoustic features

Statistics Functions
Position maximum/minimum
Arithmetic mean, standard deviation
Linear regression coefficients 1/2
Linear regression error quadratic/absolute
Quartile 1/2/3
Quartile range 2-1/3-2/3-1
Percemtile 1/99
Percemtile range 99-1
Up-level time 75/99

Table 3. 21 kinds of statistics functions applied on LLDs

Tsinghua University

Textual feature

- Chinese text
  - Thulac Tool for word segmentation
  - word embeddings is learned with word2vec
  - 31.2 million chinese word as the training corpora
- English text
  - public 300 dimensional vectors
  - Trained on 100 billion words From Google News(Mikolov et al. 2013)
- extract 4200-dimensional utterance-level textual features according to the statistic functions (mean,max) over the above LLDs

Tsinghua University

Social feature

- we define 7 query topic types {Chat, Consultation, Joke, Entertainment, Operation, Search and Other} as type features and user query locations as the accent features.

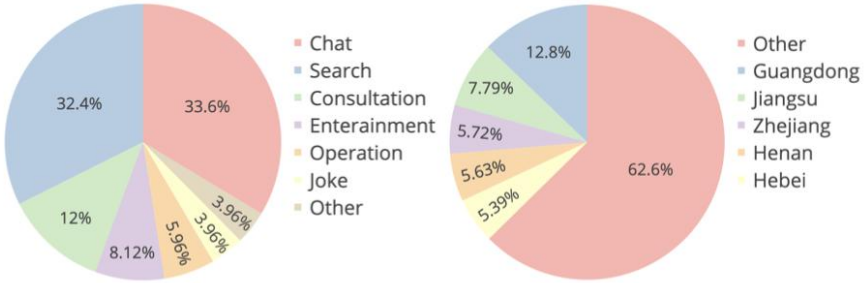


Figure 5: (a) The Topics of User Queries. (b)Top 5 User Locations.

Performance of epoch-wise MMMD

Experiment

Table 1: The F1-Measure of inferring emotion in different classification models.

	Method	Neutral	Sadness	Disgust	Anger	Happiness	Boredom	Average
F1-Measure	DNN	0.7072	0.3319	0.2198	0.3755	0.4197	0.2064	0.3768
	SAE	0.7045	0.3971	0.2278	0.3746	0.3945	0.206	0.3841
	MixMatch	0.7079	0.3554	0.2446	0.368	0.4407	0.2366	0.3922
	MMMD-w/o-avd	0.7097	0.3663	0.2104	0.4082	0.4608	0.2104	0.3972
	MMMD-w-avd	0.6936	0.3473	0.2496	0.4135	0.4662	0.2496	0.4028
	epoch-wise MMMD	0.6874	0.3976	0.2511	0.4115	0.4618	0.2772	**0.4144**

DNN: Learning a Deep neural network(Ren et al. 2014b) merely in labeled internet voice data.

SAE: Learning a DNN with labeled and unlabeled internet voice data pre-trained with Stacked Autoencoder(SAE)(Vincent et al. 2010).

Mixmatch: Learning a DNN with labeled and unlabeled internet voice data augmented with Mixmatch(Berthelot et al. 2019).

MMMD without acted voice data(MMMD-w/o-avd): Training labeled and unlabeled internet voice data with our proposed MMMD.

MMMD with acted voice data(MMMD-w-avd): Learning MMMD with acted voice Data, labeled and unlabeled internet voice data. The data samples are trained without curriculum and in random turn.

Our proposed epoch-wise-MMMD: Learning MMMD with acted voice Data, labeled and unlabeled internet voice data in epoch-wise strategy.

1、Semi-supervised MMMD：MMMD improves the F1 by 3.08% comparing to baseline SAE relatively.

2、Epoch-wise Training Strategy：The epoch-wise-MMMD with a epoch-wise learning strategy to leverage AVD improves the F1 by 4.12% relatively.

Performance of MMD

Experiment

Tsinghua University

Comparison to the state-of-art method on public dataset IEMOCAP

	Method	A(%)		T(%)		A+T(%)	
		UA	WA	UA	WA	UA	WA
RNN	[ICASSP, 2017]	58.8	**63.5**	-	-	-	-
MDNN	[AAAI, 2018]	62.7	61.8	**66.9**	65.8	76.7	75.2
AE-ACNN	[ICASSP, 2019a]	59.54	-	-	-	-	-
CNN-LSTM	[ICASSP, 2019b]	53.23	53.43	59.40	59.63	65.9	64.97
Attention-GMU	[ACL, 2019]	59.76	-	-	-	71.69	-
MMD	Our Method	**63.7**	62.2	66.06	**66.37**	**77.0**	**76.6**

Supervised MMD

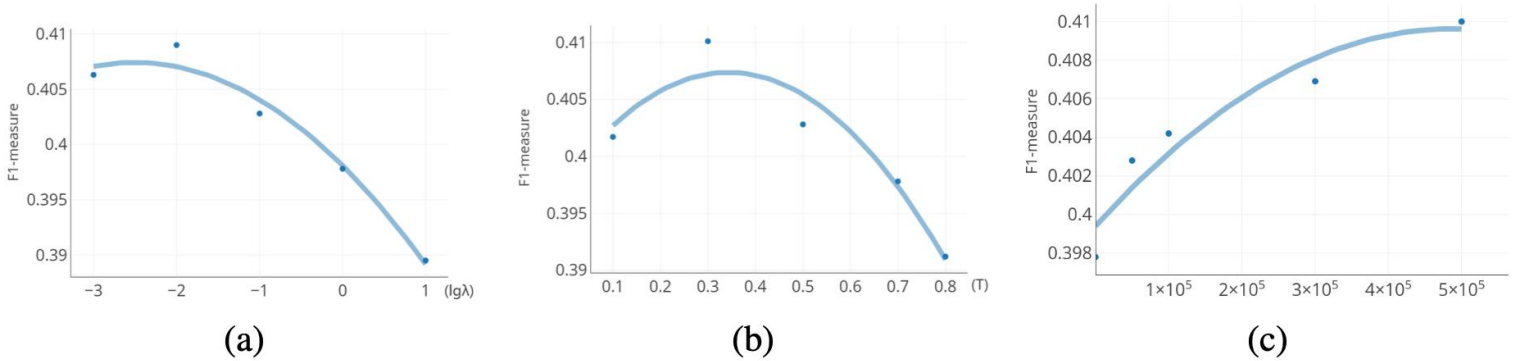
unweighted accuracy (UA)

weighted accuracy (WA)

comparing the performance ‘feature A+T’, our proposed method outperforms all the state-of- the-art baseline methods. Especially, for the UA of the ‘feature A+T’, +11.1% compared with [ICASSP, 2019b] using CNN-LSTM and +5.3% compared with [ACL, 2019] using Attention-GMU.

Parameter and Data scalability Analysis.

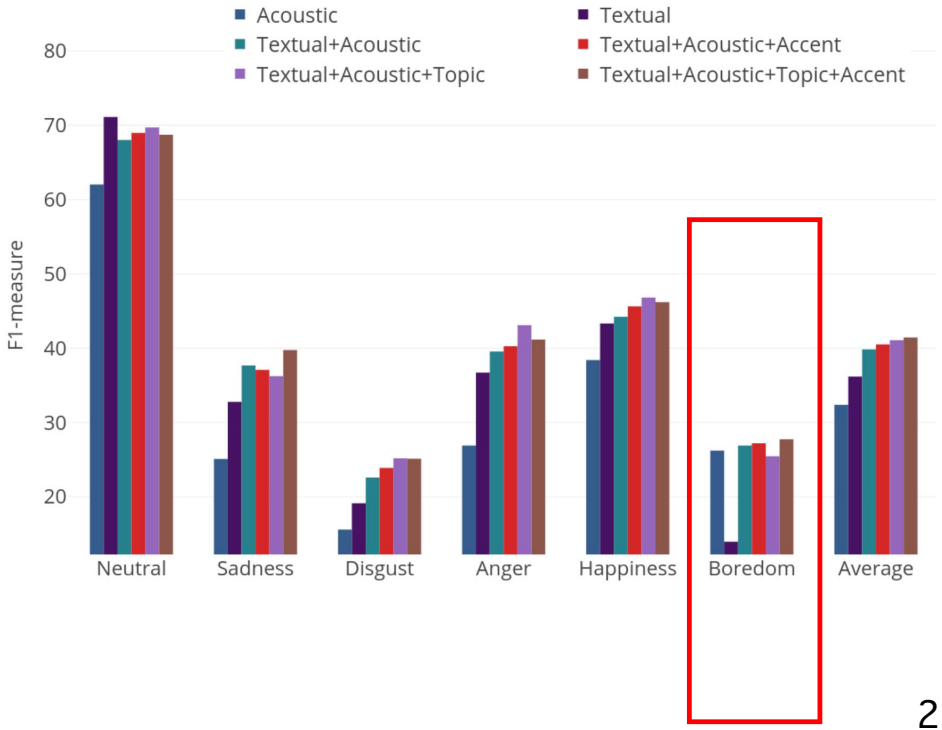
Experiment



(a)Effects of *Lamda*. (b)Effects of *T*. (c) Performance with different amount of unlabeled data.

Analysis

Experiment



- the textual information can contribute more to the emotion recognition in the real world VDAs.

- utilize all modalities simultaneously can be more effective to infer emotional utterances.

清華大學
Tsinghua University

6 Conclusion

Conclusion

Inferring Emotion from Large-scale Internet Voice Data

Tsinghua University

- We design a curriculum learning based epoch-wise training strategy to effectively utilize the strong and clear emotion from acted corpus to enhance internet voice data
- We propose a Multi-path Mixmatch multimodal deep learning method (MMMD) to utilize large-scale unlabeled data to augment few labeled data.
- Our approach turns out to be effective in real-world speech emotion inferring, which can provide more intelligent response in real-world VDA applications.

国际人工智能会议
AAAI 2021 论文北京预讲会

THANKS

2020.12.19