信也科技杯 FINVOLUTION BIG DATA COMPETITION | 信也 FINVOLUTION

数智创新 声至未来

DEEP IN DIALECTS, FOR FUTURE WAVE

第八届信也科技杯算法大赛

THE 8TH FINVOLUTION DATA SCIENCE COMPETITION

DECEM

张毅　王晨跃　刘晓雅

目录

CONTENTS

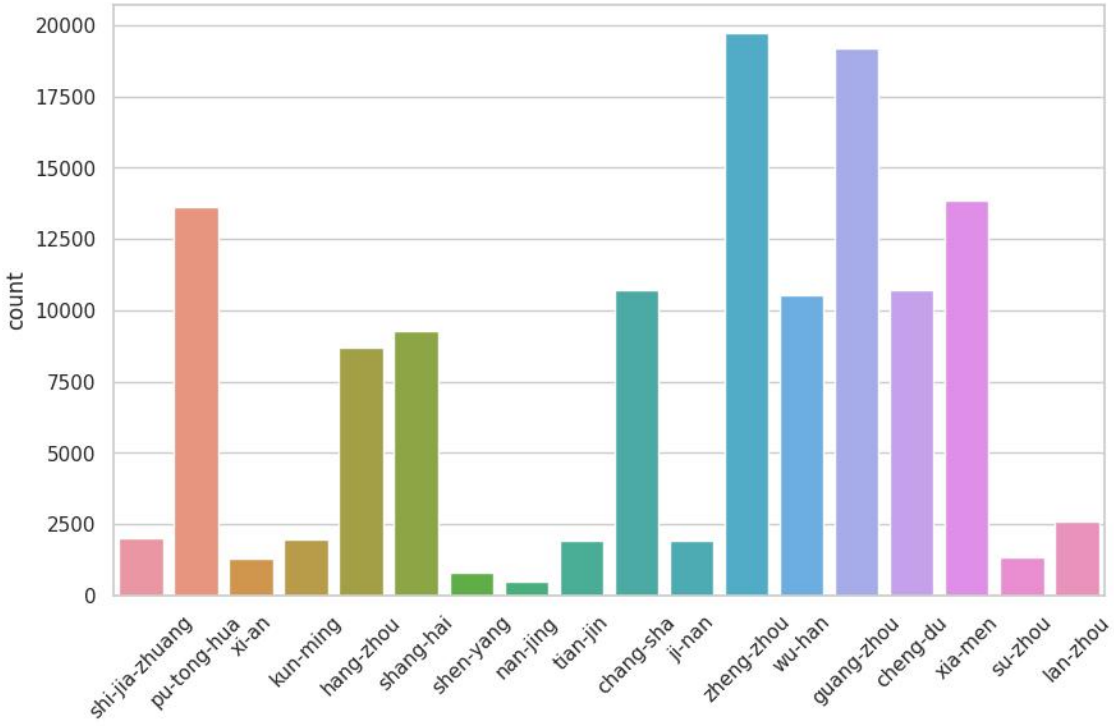
任务分析：赛题描述

- **赛题描述**

  在不断扩展的客户理解维度中，需要解决语音数据的ASR转写问题，但通用ASR模型无法正确处理方言。提出的问题是如何度量不同方言之间的距离，以便在核心方言ASR模型之间选择最接近的模型来处理未知方言。比赛旨在鼓励选手通过度量方言距离的方式开发算法来解决这个问题，有助于改进语音识别的能力，同时也可以应用于商业ASR接口的模型扩展。

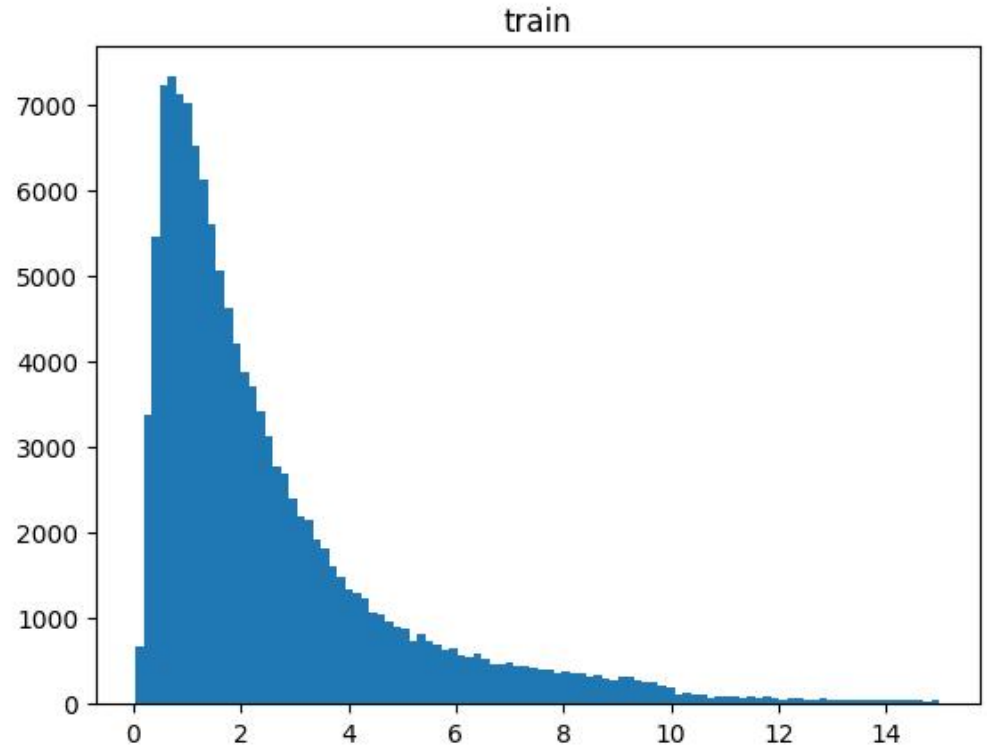
- **数据规模**
  - 训练数据：初赛和复赛总共给出了130541条语音
  - 方言种类：初赛和复赛各提供了9种方言，共18种方言

任务分析：可视化分析



不同方言类别的音频数量分布

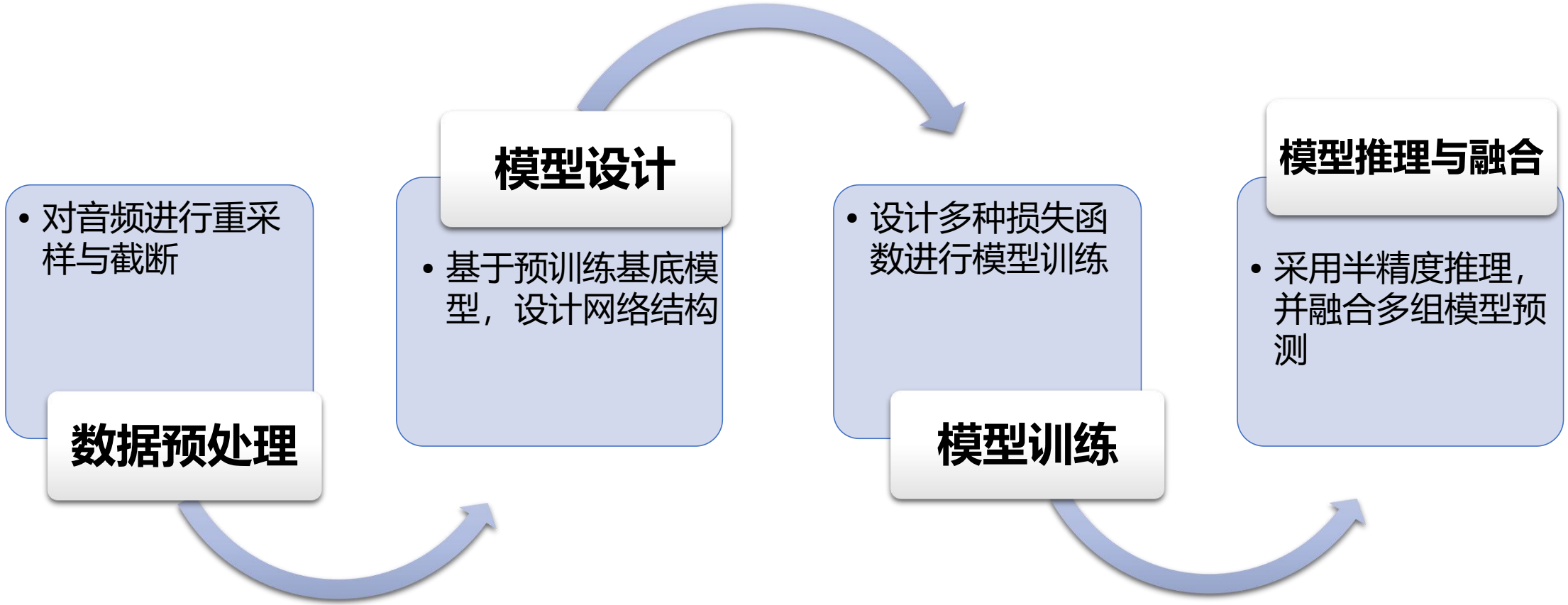
- 包括初赛和复赛两组方言
- 数量分布存在不平衡的情况



方言音频时长分布

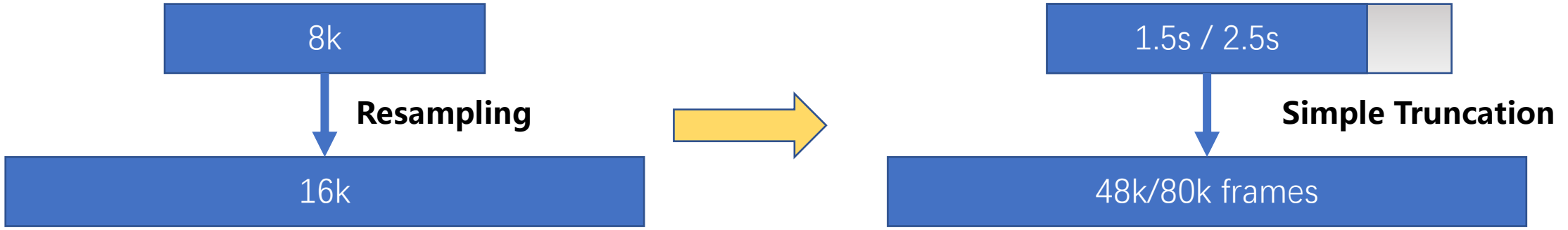
- 大部分音频时长为数秒
- 时长中位数约为3秒

技术实现


数据预处理
• 对音频进行重采样与截断
模型设计
• 基于预训练基底模型，设计网络结构
模型训练
• 设计多种损失函数进行模型训练
模型推理与融合
• 采用半精度推理，并融合多组模型预测

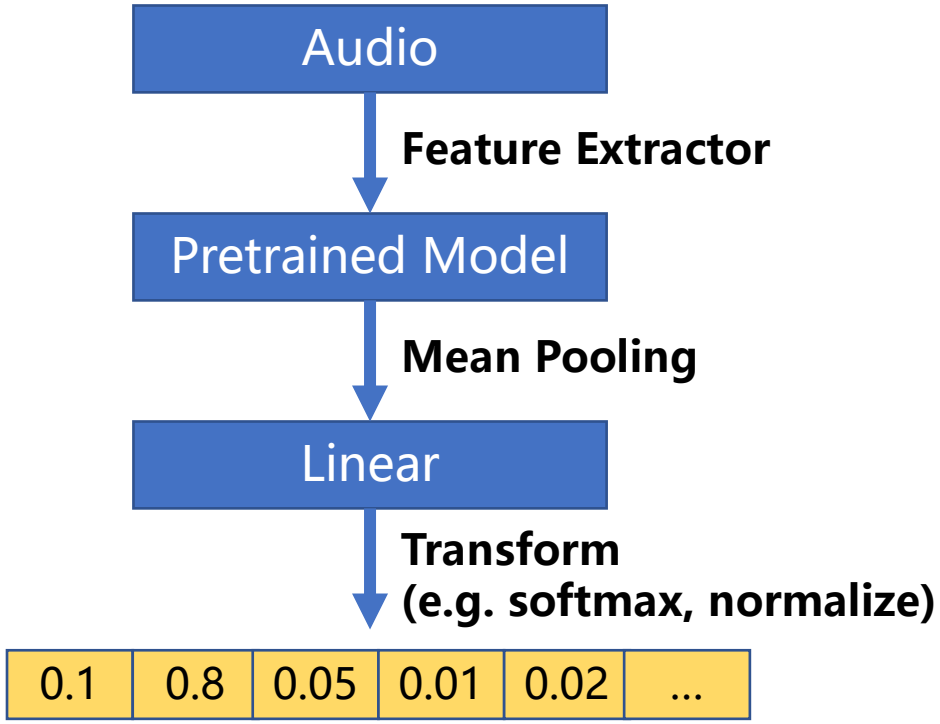
技术实现：数据预处理

- 音频重采样至16k，与预训练模型要求一致
- 截断音频，保留前48k/80k frames（对应原始音频时长为1.5s/2.5s）
- 划分训练数据与验证数据，用于模型训练与评估
  - 在最终提交的模型中，也包含了直接在所有音频数据上训练后的模型

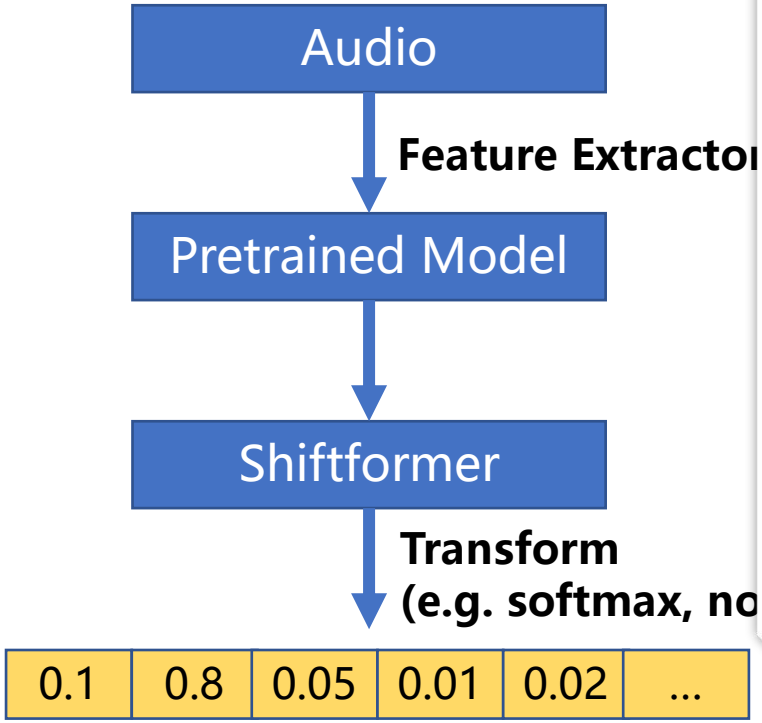


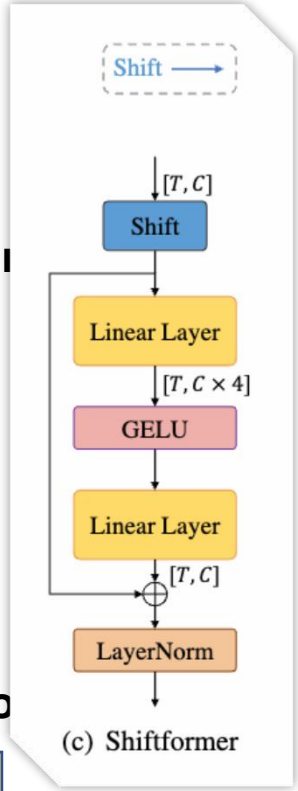
技术实现：模型设计

- 方案一：经典的基于预训练模型的网络结构



- 方案二：对预训练模型输出的向量经过Shiftformer进行通道混合





技术实现：模型设计

Wav2vec2.0：

- 语音自监督学习的先驱工作，在无标签数据上通过**端到端的方式**进行**联合训练**。

HuBERT：

- 先通过无监督训练得到的聚类模型将语音信号进行离散化得到 Hidden units 作为目标序列，再使用类似BERT的MLM自监督预训练方法使模型通过掩码后的语音信号去预测掩码位置的目标值。

最终选择HuBERT-large模型

MMS：

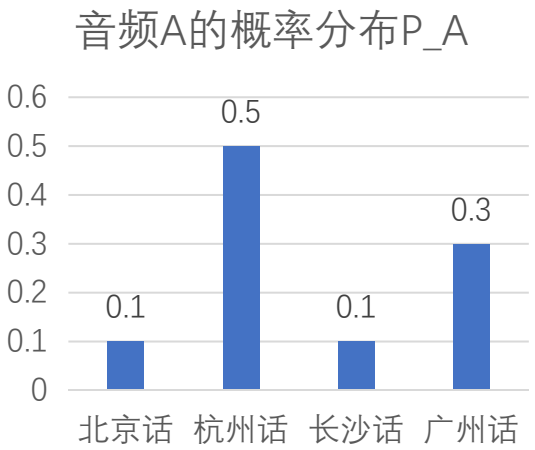
- 结合wav2vec2.0，使用 1400 多种语言的超过 50万小时的音频进行了预训练。

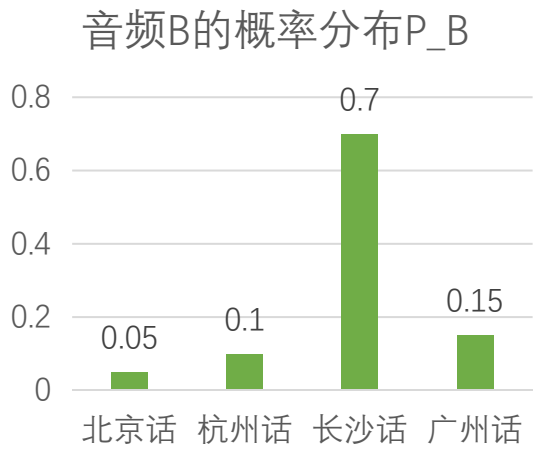
技术实现：模型训练

在初赛环节，我们通过基于交叉熵的分类任务训练模型，并计算音频对的方言距离

- 将模型输出的概率分布作为“相似度”，结合距离矩阵预测距离

$$Dist(A,B) = \sum_{d_1 \in D} \sum_{d_2 \in D} P_A^{d_1} P_B^{d_2} D_{d_1,d_2} = P_A^T D P_B$$





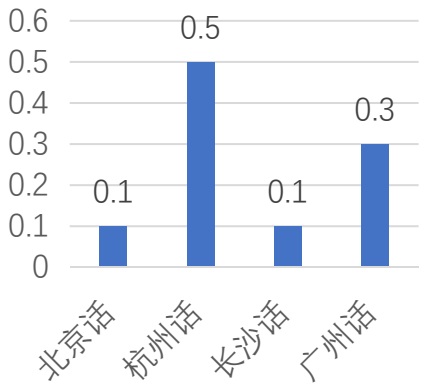
	北京	成都	郑州	武汉	广州	上海	杭州	厦门	长沙
北京	0	32	23	34	69	68	57	79	52
成都	32	0	38	25	66	61	50	77	44
郑州	23	38	0	40	69	71	62	79	56
武汉	34	25	40	0	66	63	54	77	34
广州	69	66	69	66	0	67	68	71	68
上海	68	61	71	63	67	0	41	78	64
杭州	57	50	62	54	68	41	0	76	57
厦门	79	77	79	77	71	78	76	0	77
长沙	52	44	56	34	68	64	57	77	0

- 解决方案简单高效
- 集内语种预测准确

- 训练与推理任务不一致
- 集外语种泛化能力差

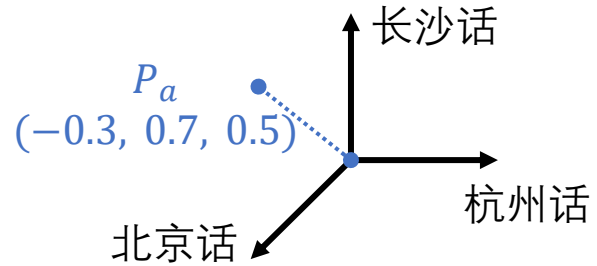
技术实现：模型训练

在复赛环节，我们通过基于MSE损失的回归任务训练模型

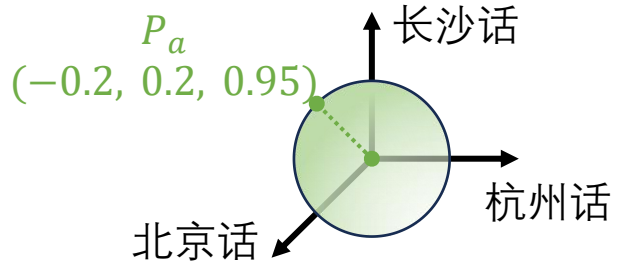


logits + softmax

	初赛	复赛
建模任务	分类	回归
损失函数	Cross Entropy	MSE
模型输出	类别概率分布 logits + softmax	特征空间坐标 logits (+normalization)
距离计算	$Dist(A,B) = P_A^T \, D \, P_B$	



logits



logits + normalization
投影到超球面

训练样本对的构建：计算一个batch内的所有样本对与其方言距离之间的损失

```
distance = torch.mm(torch.mm(logits, distance_matrix), logits.transpose(0, 1))
```

- 对于bs条样本，可高效构建bs*bs个音频对进行训练
- 每个epoch打乱样本顺序，得到几乎不重复的音频对
- 可通过控制采样权重调节方言类别不平衡的情况

技术实现：模型训练

从几何变换的角度重新审视距离计算公式...

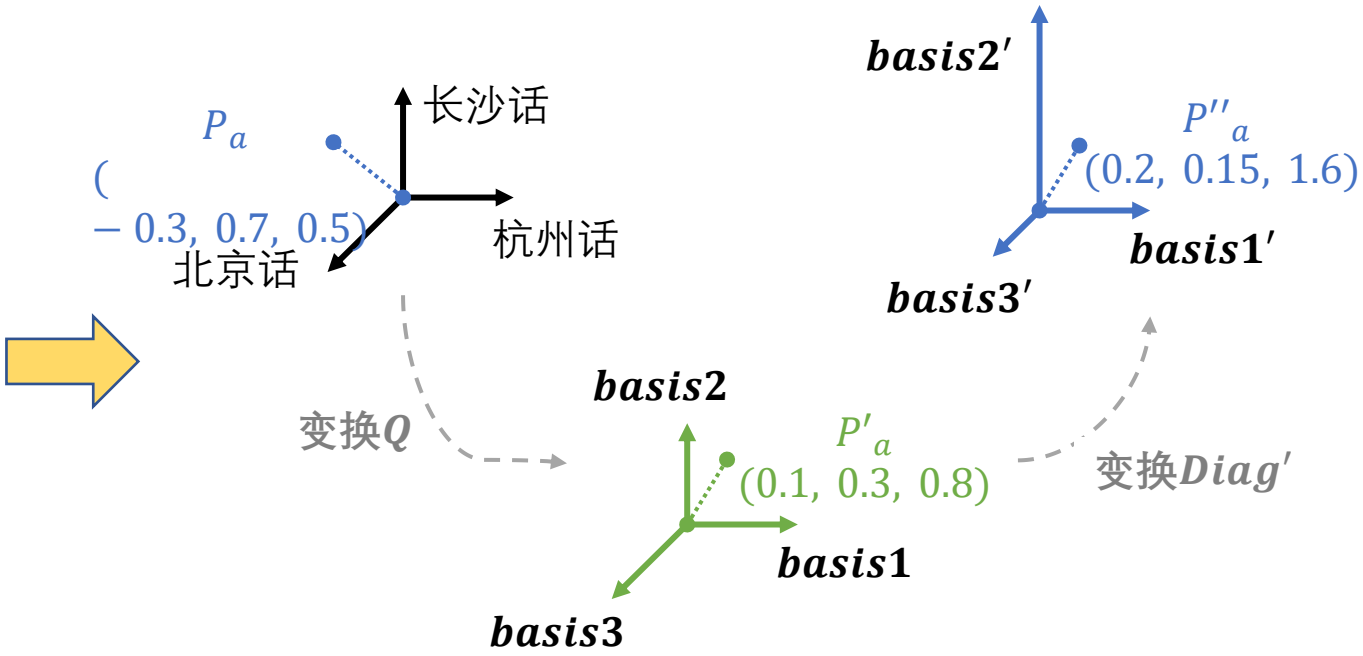
- 由于距离矩阵D 是对称方阵，因此可进行对角化

$$Dist(A,B) = P_A^T\ D\ P_B = P_A^T\ (Q^T\ Diag\ Q)\ P_B = (Diag'QP_A)^T\ (Diag'QP_B)$$

- 距离矩阵D对原特征空间的“坐标”进行了变换和伸缩，映射到新的特征空间
- 计算得到的距离是音频特征表示在**新的特征空间内的内积**

方言距离矩阵D

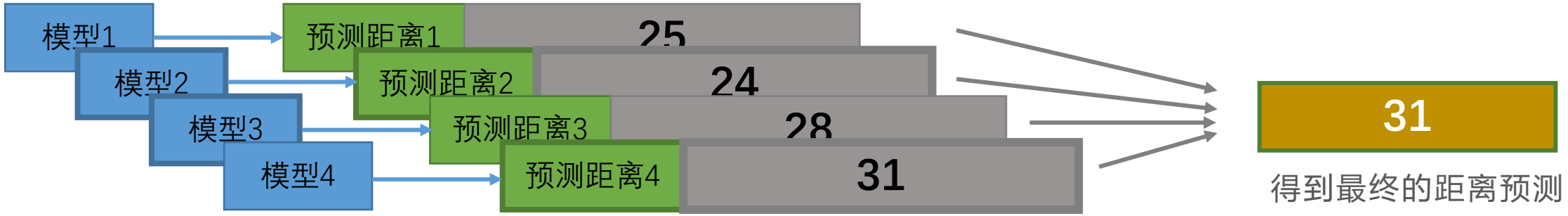
	北京	成都	郑州	武汉	广州	上海	杭州	厦门	长沙
北京	0	32	23	34	69	68	57	79	52
成都	32	0	38	25	66	61	50	77	44
郑州	23	38	0	40	69	71	62	79	56
武汉	34	25	40	0	66	63	54	77	34
广州	69	66	69	66	0	67	68	71	68
上海	68	61	71	63	67	0	41	78	64
杭州	57	50	62	54	68	41	0	76	57
厦门	79	77	79	77	71	78	76	0	77
长沙	52	44	56	34	68	64	57	77	0



技术实现：模型推理与融合

最大化利用决赛容器资源，提升模型预测效果

- 推理时采用半精度，加快推理速度并降低显存占用，同时预测结果与线上分数并未显著劣化
- 观察到预测值与实际值偏小的情况，采用取max的方法得到各组模型预测的最大值作为最终结果



多组模型推理与融合，
得到最终的距离预测

不同的融合方法：算术mean、几何mean、调和mean、**max**

实现效果

pretrained model	model type	loss type	n_frames	online
wav2vec2_base	simple	CE	3 (1.5s)	15.959569
wav2vec2_base	simple	Norm + MSE	3 (1.5s)	15.198462
wav2vec2_large	simple	CE	3 (1.5s)	15.85332
Hubert_large	shiftformer	Norm + MSE	3 (1.5s)	14.703963
Hubert_large	simple	Norm + MSE	3 (1.5s)	13.884777
Hubert_large	simple	Norm + MSE	5 (2.5s)	13.66821

pretrained model	model type	loss type	n_frames	
Hubert_large	simple	MSE	5 (2.5s)	All data
Hubert_large	simple	Norm + MSE	5 (2.5s)	All data
Hubert_large	simple	MSE	3 (1.5s)	All data
Hubert_large	simple	Norm + MSE	5 (2.5s)	Training data
			Merge by mean	**12.406032**
			Merge by max	**11.483201**

创新观点

创新点

- 模型设计：采用简单易用的模型网络结构，探索并比较了不同先进预训练模型的任务效果。
- 模型训练：提出基于方言距离矩阵变换的内积距离计算方式，高效可控地构造训练样本对。
- 模型推理：采用半精度推理与模型融合方法，提高推理效率及预测准确性。

展望

- 改进度量方法：进一步研究和改进方言距离的度量方法，以提高模型效果及泛化能力。
- 调优超参数：积累音频与回归任务的调参经验与技巧，炼出更好的模型。
- 数据扩充：探索数据扩充技术，特别是在数据有限的方言中，以进一步提高模型性能。

信也科技杯 FINVOLUTION BIG DATA COMPETITION | 信也 FINVOLUTION

THANK YOU