语音识别：从入门到精通

第九讲：区分性训练和LF-MMI

主讲人 张彬彬

西北工业大学

binbzha@gmail.com

背景知识回顾

- 最大似然估计Maximum Likelihood(ML)
- HMM
  - 前向后向算法
  - Viterbi算法
- 统计语言模型n-gram
- 解码
  - WFST
  - 1-Best
  - N-Best
  - Lattice

参考资料（论文/讲义/博客）

- *2007 - Gales, Young - The application of hidden Markov Models in speech recognition*
- *2013 - Veselý et al. - Sequence-discriminative training of deep neural networks*
- *2016 - Povey et al. - Purely sequence-trained neural networks for ASR based on lattice-free MMI*
- *2016 - Xiong et al. - Achieving Human Parity in Conversational Speech Recognition*
- *2018 - Hadian et al. - End-to-end speech recognition using lattice-free MMI*
- *2019 - Peter Bell - Lattice-free MMI(lecture)*
- *2020 – Chao Yang - Sequence-discriminative training of DNNs笔记(blog)*

内容提要

- 区分性训练(Discriminative Training)
- LF-MMI(Lattice Free MMI)
- Kaldi chain model
- 作业

高级话题 高阶内容

最大似然训练

$$P(W|O) = \frac{P(O|W)P(W)}{P(O)}$$

- 语言模型$P(W)$
- 声学模型$P(O|W)$
- 最大似然声学模型训练

$$\theta_{ML} = \arg\max_{\theta} P_{\theta}(O|W)$$

ML: Maximum Likelihood

区分性训练

$$\mathrm{P}(W|O)=\frac{P(O|W)P(W)}{P(O)}$$

- 语言模型$P(W)$
- 声学模型$P(O|W)$
- 我们可不可以直接最大化$\mathrm{P}(W|O)$？ ==> 基于MMI的区分性声学模型训练

$$\theta_{MMI}=\arg\max_{\theta} P_{\theta}(W|O)$$

MMI: Maximum Mutual Information

ML和MMI

$$\theta_{\mathrm{ML}} = \arg\max_{\theta} \sum_{u} \log P_{\theta}(O_u|W_u)$$

$$\theta_{\mathrm{MM}I} = \arg\max_{\theta} \sum_{u} \log P_{\theta}(W_u|O_u)$$

$$= \arg\max_{\theta} \sum_{u} \log \frac{P_{\theta}(O_u|W_u)P(W_u)}{P(O_u)}$$

$$= \arg\max_{\theta} \sum_{u} \log \frac{P_{\theta}(O_u|W_u)P(W_u)}{\sum_W P_{\theta}(O_u|W)\, P(W)}$$

ML仅考虑最大化正确路径（标注）概率

MMI（思考）：

- 如何优化该式？
- 分母是个有限集合吗？

MMI

$$\theta_{MMI} = \arg\max_{\theta} \sum_{u} \log \frac{P_\theta(O_u|W_u)P(W_u)}{\sum_W P_\theta(O_u|W)\,P(W)}$$

- 如何优化该式，这是个分式，所以？
  - 增大分子(Numerator)
  - 减小分母(Denominator)
- 声学模型$P_\theta(O|W)$
  - GMM(均值、方差）
  - DNN（网络参数）

ML/MMI in HMM with DNN

音频语音与语言处理研究组 SLP Audio, Speech and Language Processing Group NPU

ML

$$\mathcal{F}_{CE} = -\sum_{u=1}^{U}\sum_{t=1}^{T_u} \log y_{ut}(s_{ut}),$$

$$\frac{\partial \mathcal{F}_{CE}}{\partial a_{ut}(s)} = -\frac{\partial \log y_{ut}(s_{ut})}{\partial a_{ut}(s)} = y_{ut}(s) - \delta_{s;s_{ut}},$$

MMI(将W展开成HMM state sequence S)

$$\mathcal{F}_{MMI} = \sum_{u} \log \frac{p(\mathbf{O}_u|S_u)^{\kappa} P(W_u)}{\sum_W p(\mathbf{O}_u|S)^{\kappa} P(W)},$$

$$\frac{\partial \mathcal{F}_{MMI}}{\partial a_{ut}(s)} = \sum_{r} \frac{\partial \mathcal{F}_{MMI}}{\partial \log p(\mathbf{o}_{ut}|r)} \frac{\partial \log p(\mathbf{o}_{ut}|r)}{\partial a_{ut}(s)},$$
$$= \kappa(\delta_{s;s_{ut}} - \gamma_{ut}^{DEN}(s)).$$

思考：目标函数与梯度的关系？

2013 - Veselý et al. - Sequence-discriminative training of deep neural networks（上文公式来源）

2020 – Chao Yang - Sequence-discriminative training of DNNs笔记(对推导感兴趣的同学请参考这篇）

MMI

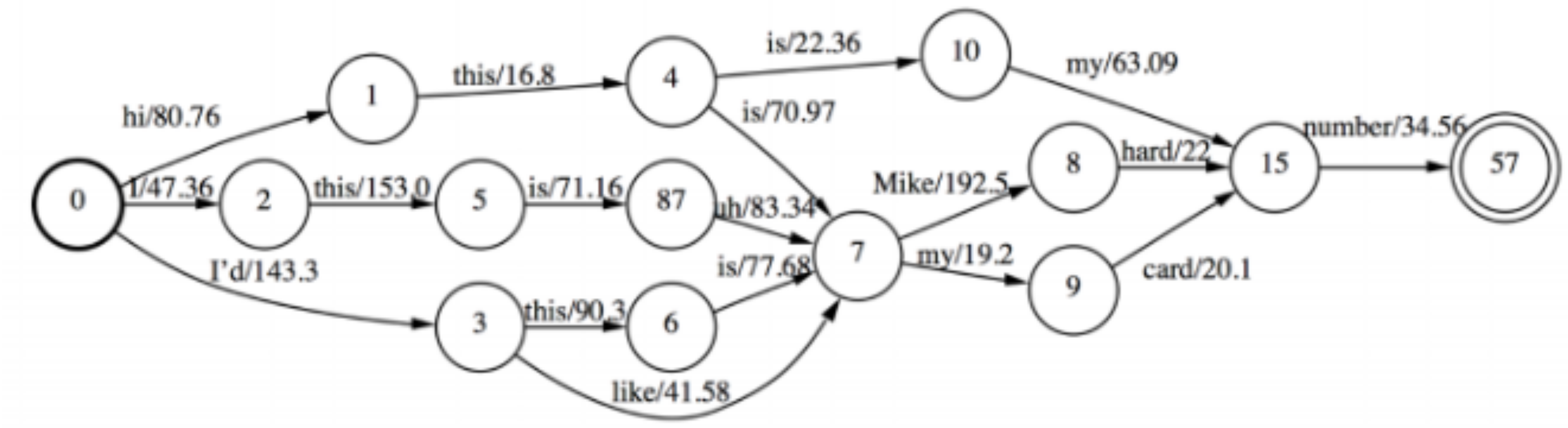
MMI(将W展开成HMM state sequence S)

$$\mathcal{F}_{MMI} = \sum_u \log \frac{p(\mathbf{O}_u|S_u)^\kappa P(W_u)}{\sum_W p(\mathbf{O}_u|S)^\kappa P(W)},$$

$$\frac{\partial \mathcal{F}_{MMI}}{\partial a_{ut}(s)} = \sum_r \frac{\partial \mathcal{F}_{MMI}}{\partial \log p(\mathbf{o}_{ut}|r)} \frac{\partial \log p(\mathbf{o}_{ut}|r)}{\partial a_{ut}(s)},$$
$$= \kappa(\delta_{s;s_{ut}} - \gamma_{ut}^{DEN}(s)).$$

- 似乎一切都很顺利，甚至在没有穷举W的情况下，我们将其梯度都计算了出来？
- But，为了计算$\gamma_{ut}^{DEN}(s)$，必须要给出W的所有可能？
  - 怎么办？
  - 怎么办？
  - W应该是有限空间，可枚举的。

Lattice based MMI



- 利用原语音解码生成的Lattice来近似所有的W的可能
  - 概率低的序列在解码阶段会被及时裁剪掉
  - 如何在Lattice上计算$\gamma_{ut}^{DEN}(s)$：前向后向算法
- Tricks:
  - Wider lattice，弱语言模型(uni-gram/bi-gram)

区分性训练其他准则

- MPE/sMBR

$$\mathcal{F}_{MBR} = \sum_{u} \frac{\sum_{W} p(\mathbf{O}_u|S)^{\kappa} P(W) A(W, W_u)}{\sum_{W'} p(\mathbf{O}_u|S)^{\kappa} P(W')},$$

- MCE
- bMMI
- …

MPE: Minimum Phone Error
sMBR: state- level Minimum Bayes Risk
MCE: Minimum Classification Error
bMMI: boosted MMI

Lattice based区分性训练流程

数据准备 → 特征提取 → 单音素训练 → 三音素训练 → 对齐 → DNN训练 → 解码生成Lattice → 区分性训练 → 解码

- Lattice生成需要解码，代价很高，一般只在DNN模型的基础上一次生成，模型训练中不重新生成Lattice。
- 优点：我们的识别率越来越好
- 缺点： 我们的流程越来越长，系统越来越复杂

Lattice based区分性训练实验

Table 3: Results (% WER) of the DNNs trained on the full 300 hour training set using different criteria.

System	Hub5 '00 SWB	Hub5 '00 CHE	Hub5 '00 Total	Hub5 '01 SWB	Hub5 '01 SWB2P3	Hub5 '01 SWB-Cell	Hub5 '01 Total
GMM BMMI	18.6	33.0	25.8	18.9	24.5	30.1	24.6
DNN CE	14.2	25.7	20.0	14.5	19.0	25.3	19.8
DNN MMI	12.9	24.6	18.8	13.3	17.8	23.7	18.4
DNN sMBR	12.6	24.1	18.4	13.0	17.7	22.9	18.0
DNN MPE	12.9	24.1	18.5	13.2	17.7	23.4	18.2
DNN BMMI	12.9	24.5	18.7	13.2	17.8	23.5	18.3

相对于CE，通常会有5%～15%WERR

2013 - Veselý et al. - Sequence-discriminative training of deep neural networks

MMI

MMI(将W展开成HMM state sequence S)

$$\mathcal{F}_{MMI} = \sum_u \log \frac{p(\mathbf{O}_u|S_u)^\kappa P(W_u)}{\sum_W p(\mathbf{O}_u|S)^\kappa P(W)},$$

$$\frac{\partial \mathcal{F}_{MMI}}{\partial a_{ut}(s)} = \sum_r \frac{\partial \mathcal{F}_{MMI}}{\partial \log p(\mathbf{o}_{ut}|r)} \frac{\partial \log p(\mathbf{o}_{ut}|r)}{\partial a_{ut}(s)},$$
$$= \kappa(\delta_{s;s_{ut}} - \gamma_{ut}^{DEN}(s)).$$

- 似乎一切都很顺利，甚至在没有穷举W的情况下，我们将其梯度都计算了出来？
- But，为了计算$\gamma_{ut}^{DEN}(s)$，必须要给出W的所有可能？
  - Lattice
  - 表示W，还有别的办法吗？比如说统计？

Lattice Free MMI

$$\mathcal{F}_{MMI} = \sum_u \log \frac{p(\mathbf{O}_u|S_u)^\kappa P(W_u)}{\sum_W p(\mathbf{O}_u|S)^\kappa P(W)},$$

- 如何表示分母W的所有可能？统计n-gram
  - Word?
  - Phone?
  - State?
- Lattice free MMI
  - 由训练数据训练Phone/State的n-gram, and no back-off
  - WFST Compose成State level的FST
  - FST + AM score + 前向后向算法计算$\gamma_{ut}^{DEN}(s)$
- Lattice free MMI 训练流程

数据准备 → 特征提取 → 单音素训练 → 三音素训练 → 对齐 → DNN训练 → Lattice Free MMI → 解码

Lattice Free MMI实验

Table 3. Performance improvements from i-vector and LFMMI training on the NIST 2000 CTS test set

Configuration	WER (%)							
	ReLU-DNN		ResNet-CNN		BLSTM		LACE	
	CH	SWB	CH	SWB	CH	SWB	CH	SWB
Baseline	21.9	13.4	17.5	11.1	17.3	10.3	16.9	10.4
i-vector	20.1	11.5	16.6	10.0	17.6	9.9	16.4	9.3
i-vector+LFMMI	17.9	10.2	15.2	8.6	16.3	8.9	15.2	8.5

- State level 3-gram for denominator
- LF-MMI also got promising gain.

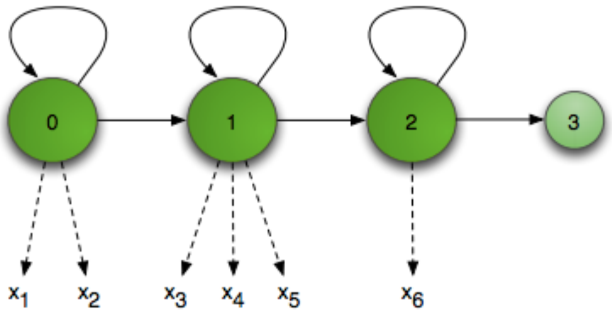
2016 - Xiong et al. - Achieving Human Parity in Conversational Speech Recognition

Kaldi chain model

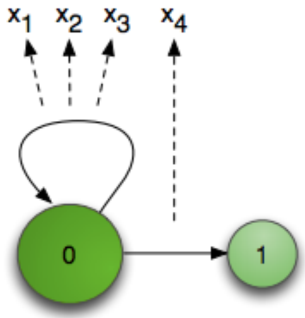
- Chain model: Lattice Free MMI from scratch, to make it
  - Better WER
  - Train faster
  - Decode faster
- But, with a lot of tricks
  - HMM Topology
  - Reduce frame rate(10ms to 30ms)
  - Numerator/Denominator all in FST framework, fixed chunk
  - CE Regularization
  - L2 Regularization
  - …
- It’s tricky, but it just works.

HMM Topology

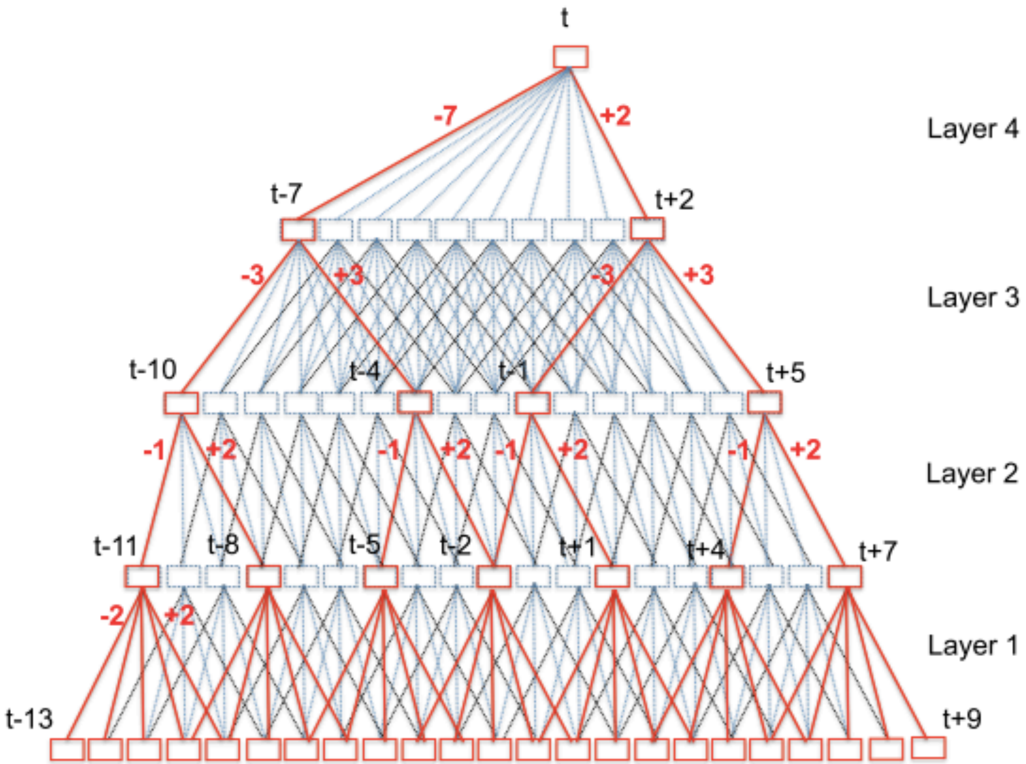
- 标准的3状态HMM topology



- LF-MMI topology



Reduce frame rate(10ms to 30ms)

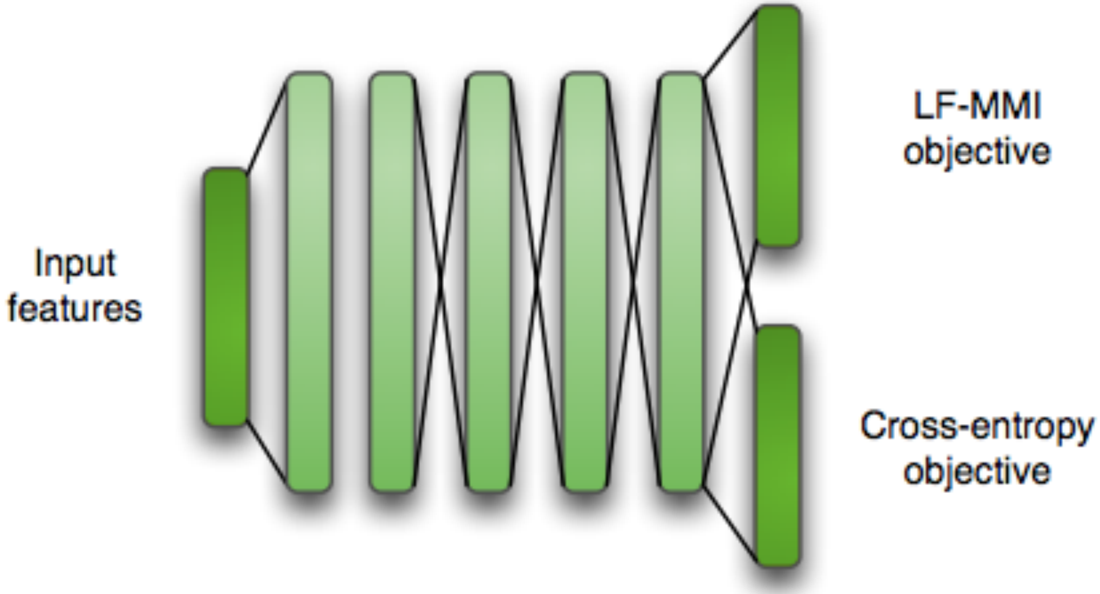


- 网络结构: TDNN
- 仅用网络输出的1/3计算loss function和梯度
- 即取t, t+3, t+6, t+9 ...
- 或者相邻的3个t中随机取一个就行

Numerator/Denominator

- Numerator
  - 使用在标注文本上生成的Lattice计算
  - 引入时间上的扰动，方便Fixed chunk切分数据。
- Denominator
  - Phone level 3-gram G, without back-off
  - Denominator FST H*C*G
  - C is bi-phone instead of tri-phone.
- Fixed Chunk
  - 将训练数据切分为固定大小的chunk(1.5s)训练

CE Regularization



使用CE作为第二个Task进行Multi-Task Learning

Kaldi chain model训练流程

- Chain prepare
  - LF-MMI Topology
  - bi-phone Tree
  - Numerator lattice
  - Denominator n-gram, FST
  - Fixed chunk

Kaldi Chain model实验

Table 4: Performance of LF-MMI on various LVCSR tasks with different amount of training data, using TDNN acoustic models

Database	Size	WER		
		CE	CE → sMBR	LF-MMI
AMI-IHM	80 hrs	25.1	23.8	$22.4^{\dagger}$
AMI-SDM	80 hrs	50.9	48.9	$46.1^{\dagger}$
TED-LIUM	118 hrs	12.1	11.3	11.2^{*}
Switchboard	300 hrs	18.2	16.9	15.5
Librispeech	1000 hrs	4.97	4.56	4.28
Fisher + SWBD	2100 hrs	15.4	14.5	13.3

- LF-MMI比CE->sMBR效果好
- LF-MM在不同数据集，不同大小的数据集上收益都很稳定
- **Currently, LF-MMI is the BEST and DEFAULT recipe in Kaldi**
- 更多的细节和Trick，请参考如下论文

2016 - Povey et al. - Purely sequence-trained neural networks for ASR based on lattice-free MMI

作业

- 跑通kaldi chain model在aishell(200h)数据集上的结果，理解Kaldi中训练chain model流程，理解Lattice free MMI。
    - 安装Kaldi
    - 数据(15G)
    - 硬件要求：带GPU的Linux服务器
    - 如何一键运行
        - cd your_kaldi_dir/egs/aishell/s5
        - bash run.sh

本章总结

数据准备 → 特征提取 → 单音素训练 → 三音素训练 → 对齐 → DNN训练 → 解码生成Lattice → 区分性训练 → 解码

数据准备 → 特征提取 → 单音素训练 → 三音素训练 → 对齐 → DNN训练 → Lattice Free MMI → 解码

数据准备 → 特征提取 → 单音素训练 → 三音素训练 → Chain Prepare → Lattice Free MMI → 解码

数据准备 → 特征提取 → **E2E Model (E2E-LF-MMI/CTC/LAS/RNN-T/Transformer)** → 解码

感谢各位聆听！

西工大音频语音与语言处理研究组