领域自适应与新词处理

在说话人自适应章节种谈到领域自适应（Domain Adaptation），当已有模型与使用场景不一致的时候，我们需要对特定的领域做自适应。如果有相应的语音数据我们可以对声学模型做调整。

区分性训练

区分性训练通过定义某一目标函数，通常称准则，来近似一个与分类代价相关的度量，例如可以定义一个与分类错误相关的量并最小化它，或是定义一个与识别正确率相关的量，并最大化它。通过区分性训练，我们可以从一定程度上弱化模型假设错误所带来的影响。同时，由于区分性训练致力于优化与识别效果好坏相关的度量，因此也就为提高识别器性能提供了更直接的途径。区分性训练更重视模型之间的分类面，以更好的根据设定的目标函数对训练数据进行分类[1]。目前常用的区分性训练准则主要包括：最大互信息量准则（Maximum Mutual Information，MMI），最小因素错误准则（Minimum Phone Error，MPE），最小状态化错误（Stat Level Minimum Bayes Risk，sMBR）[2]。

我们使用区分训练来对特定领域做自适应，通常会把新数据集的特征作为区分性训练的输入，用大模型对新数据做 Alignment 和生成 Lattice，然后将生成的对齐文件、解码网络以及数据的特征文件归档为新的数据格式（egs），然后做区分性训练，如下图所示（为了清晰，删除了一部分参数设置）：

```
if [ $stage -le 1 ]; then
  # hardcode no-GPU for alignment, although you could use GPU [you wouldn't
  steps/nnet3/align.sh  --cmd "$decode_cmd" --use-gpu false \
    --online-ivector-dir $online_ivector_dir \
     --nj $nj $train_data_dir data/lang $srcdir ${srcdir}_ali ;
if [ -z "$lats_dir" ]; then
    steps/nnet3/make_denlats.sh --cmd "$decode_cmd" --determinize true \
      --online-ivector-dir $online_ivector_dir \
      --nj $nj --sub-split $subsplit --num-threads "$num_threads_denlats" --config conf/decode_dnn.config \
      $train_data_dir data/lang $srcdir ${lats_dir} ;
fi
if [ -z "$degs_dir" ]; then
    steps/nnet3/get_egs_discriminative.sh \
      --cmd "$decode_cmd --max-jobs-run $max_jobs --mem 20G" --stage $get_egs_stage --cmvn-opts "$cmvn_opts" \
      --online-ivector-dir $online_ivector_dir \
      --left-context $left_context --right-context $right_context \
      $frame_subsampling_opt \
      --frames-per-eg $frames_per_eg --frames-overlap-per-eg $frames_overlap_per_eg \
      $train_data_dir data/lang ${srcdir}_ali $lats_dir $srcdir/final.mdl $degs_dir ;
  fi
if [ $stage -le 4 ]; then
  steps/nnet3/train_discriminative.sh --cmd "$decode_cmd" \
    --stage $train_stage \
    --effective-lrate $effective_learning_rate --max-param-change $max_param_change \
    --criterion $criterion --drop-frames true \
    --num-epochs $num_epochs --one-silence-class $one_silence_class --minibatch-size $minibatch_size \
    --num-jobs-nnet $num_jobs_nnet --num-threads $num_threads \
    --regularization-opts "$regularization_opts" \
    ${degs_dir} $dir
fi
```

Figure 1 ： Kaldi 中的 wsj recipe local/nnet3/run_tdnn_discriminative.sh。提供的区分性训练的脚本。

我们在实际的使用区分性训练做自适应过程中发现存在一些现象，一个是模型的推广问题，区分性训练对新数据学习的过于精细，在未知测试集上难以到达到与在训练集上一样的提升效果，有时甚至还会变的更差，还有一个问题就是模型收敛过快，往往在前几个 epoch 的时候就已经收敛了，继续训练，测试效果并

没有提升反而是下降。所以在训练的时候，我们可以调小学习率、及时测试和查看 loss，发现模型表现效果变差或 loss 不再下降，就可以及时停止。

迁移学习

在文章[3]中提到两种迁移学习的方法。一个是跨语种的迁移学习，如图二左图所示，由于荷兰语的缺失（NL），使用英语的大模型的前几层作为小模型初始化的前几层，然后进行训练。

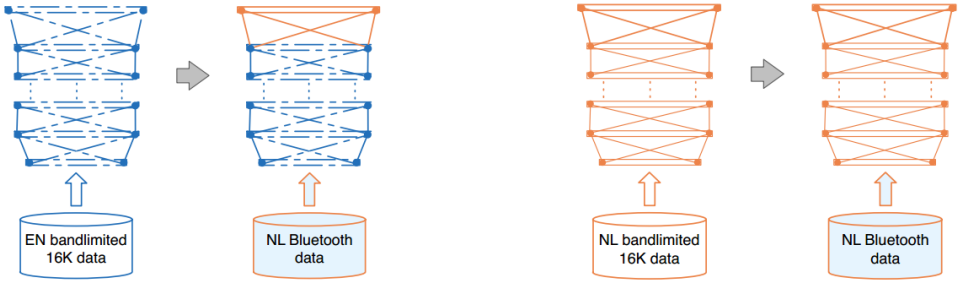


Figure 2 跨语种学习与跨频带学习

另一个如图二右所示，是针对同种语言不同频道的迁移学习，拿英语 16k 数据训练的大模型当做 8k 数据的初始模型，更详细的介绍请看文章。

在我们实际的应用过程中，对一些特殊领域数据做增强，我们通常采用借鉴大模型的前几层，作为小模型的初始化模型，通常会将借用的前几层模型固定参数，方法比如将学习率设置为 0 或者 FixAffineComponent 等，或者用大模型作为小模型的初始化模型，但参数不进行固定，这样做有一些不足之处，例如模型的结构不能改变、PDF-ID 数不能更改，但是实现方法较为容易。

还有一种迁移方法叫软迁移[4]，用大模型学到的知识引导小模型做训练，可以理解为以大模型的输出结果（soft-target）作为目标神经网络训练的过程。Soft-target 相对于传统 one-hot（hart-target），包含了更多的信息量，其用一个连续序列描述一种类别的信息，而非简单的 0/1 表示，这样更有普适意义，soft-target 使得神经网络学习空间更加连续平滑，排除掉了空间中的的突刺噪声，相对于 hard-target，更有益于训练的快速收敛，减少了走弯路或者陷入局部最优的可能性。

由于 soft-target 中不同类别共性部分的存在，会给分类任务带来一定的混淆性，而 hard-target 对于类别信息具有强烈的指向性，所以在神经网络训练时，要对二者结合使用。在一些复杂结构的神经网络训练中，训练前期可先利用一部分数据，只采用 soft-target，待神经网络收敛到某次最优点，再在此基础上结合 soft-target 与 hard-target 利用所有训练数据训练神经网络。关于软迁移的文章或者其他工具参考网址[5]。

新词添加与语言模型自适应

语音识别是计算机将人发出的语音转换为文字的系统，现有的成熟技术还不支持无约束语音解码，需要施以一定的语言结构或语法限制解码时的搜索路径，以此来提高识别率，所以语言模型成为了解码网络 HCLG 中的重要组成部分，是语音识别系统中把识别结果从字或词变通顺的语句的重要模块，一个好的语言模

型可以对语音识别的效果有质的提升。

语言模型的领域泛化可以在语料更新和新词处理等方面着力。如果有特定场景的文本预料，可以先基于此场景文本预料做语言模型，然后与原始的语言模型做差值融合，通过调整二者权重，达到对特定场景的自适应。如果没有语料，则需要先收集特定场景词表，再基于此词表收集文本语料构建相应的语言模型，也可定向的提高某些词在 HCLG 中的概率等，具体介绍如下所示：

热词表语言模型的构建

热词表语言模型的构建有两种方式，一种是基于语料的统计语言模型；另一种是基于语法结构语言模型[6]。

- 统计语言模型：根据用户热词表，通过网络搜索相关热词语料，然后基于热词表和原有词表相关词训练新的统计语言模型，再将其与原始语言模型做融合，得到新的语言模型。另一种情况下，若很难获取到热词表语料，而特定场景又有相对固定的语法结构，则可通过穷举句式的方式制作语料文本。再基于这些语料制作统计语言模型。
- 语法结构语言模型：在某些垂直领域应用场景，由于句式或命令词有限，则直接用基于语法结构语言模型即可，不同热词根据场景中使用频率可设定不同的权重。

上述热词表语言模型构建完成以后，与原始语言模型进行融合，arpa 格式直接用 ngram 进行加权平均；G.fst 或 HCLG.fst 格式用 fstunion 命令并联两个 fst，并修改两个 fst 进入边权重，实现通用场景与特定场景平衡。

基于相似词（similar pair）的热词表语言模型构建[7][8]

在原有此表中，针对每个热词选择一个相似词，再讲每个热词参照相似词插入到原有的语言模型中，操作可以在 lm.arpa 或 G.fst 或者 HCLG.fst 上进行，热词插入的权重参考原有相似词的权重（w）,另外修改其权重为aw或$a+w$，通过调剂a的值来改变新词的插入权重，若存在某些热词很难在原有此表中找到相似词，则需对其用原有词表进行分词，构建热词子词（sub-word）模型，权重操作如上所述段落所述，相似词如苹果和香蕉、中关村西路和学府路等，具有同一或相似属性的词便可称之为相似词。若原始词表中有苹果一词，而香蕉为新添加热词，则可用原有语言模型中苹果的统计概率来指代香蕉一词。

对于相似词对的选取，可以通过人工先验指定的方式，也可通过计算 td-idf/word-vec 向量计算其余弦距离得到。

基于类别（classes）信息的热词表语言模型构建

此种方法要求首先构建基于类别的统计语言模（BigClass.arpa/BigClass.G.fst），需要对原词表分类，如大类按名次、代词、形容词等，或只对部分常用类别，如人名、地名、水果类、器具类等分类，用统一的“<name>”、“<location>”、“<fruit>”等表示，每种类别可构建统计语言模型或语法结构语言模型

（SmallClass.arpa、SmallClass.G.fst），最后将其嵌入到大的语言模型（BigClass.arpa、BigClass.G.fst）中，形成最终的语言模型（lm.arpa、G.fst）。在对热词进行处理时，首先确定其类别，根据 BigClass.arpa 或 BigClass.fst 中的权重动态调其权重。

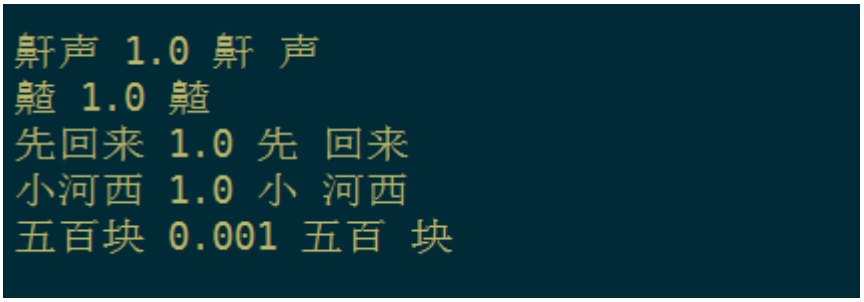
基于文本后处理的热词语音识别

此种方法是基于因素串匹配的热词识别方法，首先基于原词表与原语言模型对语音进行识别，待所有语音识别成文本后，将文本转成想用的因素串，同样的，对要添加热词也转成相应的因素串，遍历热词因素串，在识别文本因素串序列中搜寻最优匹配串，因素串匹配度可采用匹配个数比例、余弦距离（将因素串表示成向量格式）等，选取匹配度最高的热词因素串并转化成相应热词，替换掉原来识别语句中字词，来达到对热词的快速识别。

基于 fst 的可定义热词识别

热词识别的难点在与语言模型训练时，没有热词相关的概率信息，在解码时只能通过单字路径拼凑出热词，但单字存在路径稀疏及同音异形字，从而使得解码引擎识别时给出似是而非的结果，另外没在噪音背景较强时，声学模型区分性急剧降低，而语言模型部分有没有太多正确的信息约束，引入了更多的竞争性路径，造成了大量的错误识别。

以修改解码图 fst 的方式实现热词识别，首先定义热词映射字典：热词用词表内词进行描述，其他所有词表依旧映射到自身，如图所示，“鼾声”、“鼾”、“先”、“回来”、“小”等为原始词表里的词；“先回来”、“小河西”等为模拟热词；“1.0”、“0.001”为热词添加的权重，然后将热词映射字典编译成 C.fst，再与原有非热此词表加码图 HCLG.fst 进行组成新的 HCLGC.fst，新生成的解码图便具有了对热词的识别能力。且热词的权重可以调节。



小结

本文介绍了几种领域泛化的方法，分为声学模型和语言模型方面，声学模型方面主要是增强模型对特定场景的拟合能力，语言模型方法主要是增加特定场景的语料权重或者提高热词在搜索网络中的权重来提高识别效果。

[1] 鄢志杰. 声学模型区分性训练及其在自动语音识别中的应用[D].中国科学技术大学,2008.

[2] D Povey.Discriminative Training for Large Vocabulary Speech Recognition[D].Cambridge University.2004

[3] Zhuang, Xiaodan & Ghoshal, Arnab & Rosti, Antti-Veikko & Paulik, Matthias & Liu, Daben. (2017). Improving DNN Bluetooth Narrowband Acoustic Models by Cross-Bandwidth and Cross-Lingual Initialization. 2148-2152.10.21437/Interspeech.2017-1129.

[4] G. E. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," in *NIPS 2014 Deep Learning Workshop*, 2014

[5] https://github.com/dkozlov/awesome-knowledge-distillation

[6] https://www.w3.org/TR/2000/NOTE-jsgf-20000605/

[7] Ma X, Wang X, Wang D. Low-frequency word enhancement with similar pairs in speech recognition[C],Signal and Information Processing (ChinaSIP), 2015 IEEE China Summit and International Conference on. IEEE,2015: 343-347

[8] Ma X, Wang D, Tejedor J, et al. Similar Word Model for Unfrequent Word Enhancement in Speech Recognition[J]. IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP), 2016, 24(10): 1819-1830