Обработка речевых сигналов

Блок 2. Автоматическое распознавание речи

Максим Кореневский
Старший научный сотрудник ООО «ЦРТ»,
к.ф.-м.н.

Настоящий блок лекций подготовлен при поддержке «ЦРТ | Группа компаний»

ЦРТ | ГРУППА КОМПАНИЙ

Блок 2. Автоматическое распознавание речи (Automatic Speech Recognition, ASR)

Часть 4. Традиционные системы распознавания речи на основе нейронных сетей

План лекции

- Использование нейронных сетей в качестве классификаторов
- Гибридные и тандемные системы распознавания
- Обучение DNN-HMM систем распознавания
- Последовательно-дискриминативное обучение
- Адаптация систем распознавания речи на основе нейронных сетей

План лекции

- Использование нейронных сетей в качестве классификаторов
- Гибридные и тандемные системы распознавания
- Обучение DNN-HMM систем распознавания
- Последовательно-дискриминативное обучение
- Адаптация систем распознавания речи на основе нейронных сетей

Использование нейронных сетей как классификаторов

Вспомним вероятностную постановку задачи распознавания:

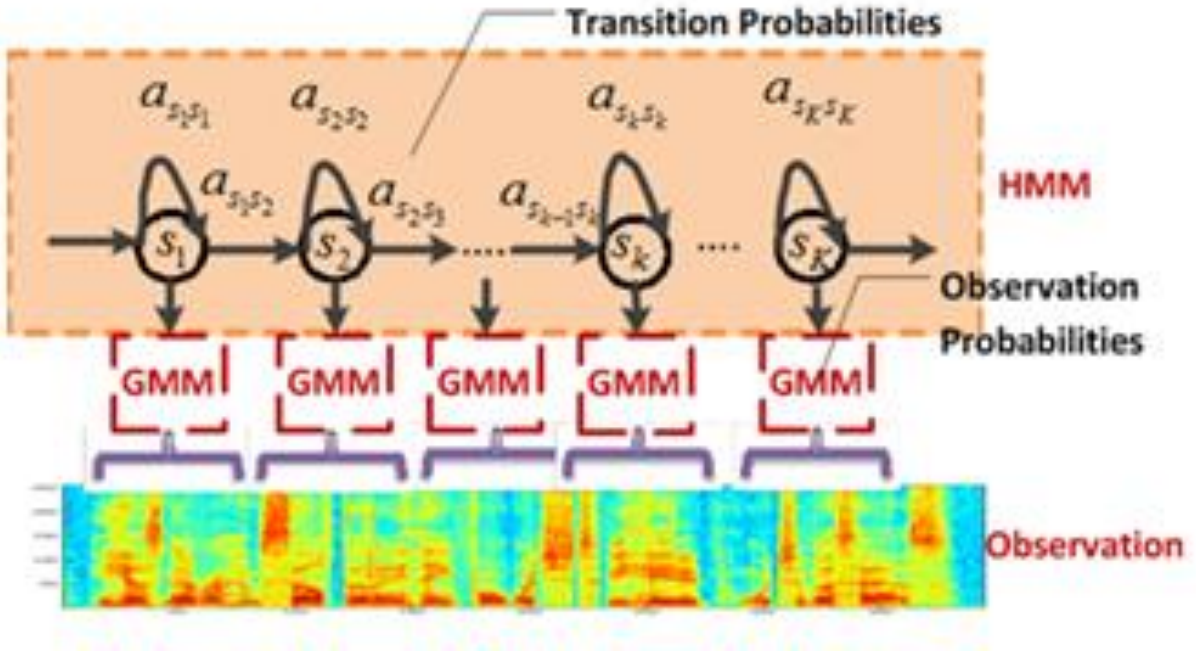
- Ищется последовательность слов W :

$$W = \arg\max_{W} p(O|W)P(W)$$

- Правдоподобие $p(O|W)$ - вычисляется акустической моделью через вероятности переходов в HMM и правдоподобия в состояниях HMM $b_i(o_t)$.
- Распространенный вариант распределения в состояниях – это GMM
- Фактически GMM работают, как классификаторы состояний по вектору наблюдений
- Альтернативный вариант – использовать искусственные нейронные сети (ANN)
- Нейронная сеть – тоже классификатор, только дискриминативный и не локальный
- Выдает результаты для всех состояний сразу

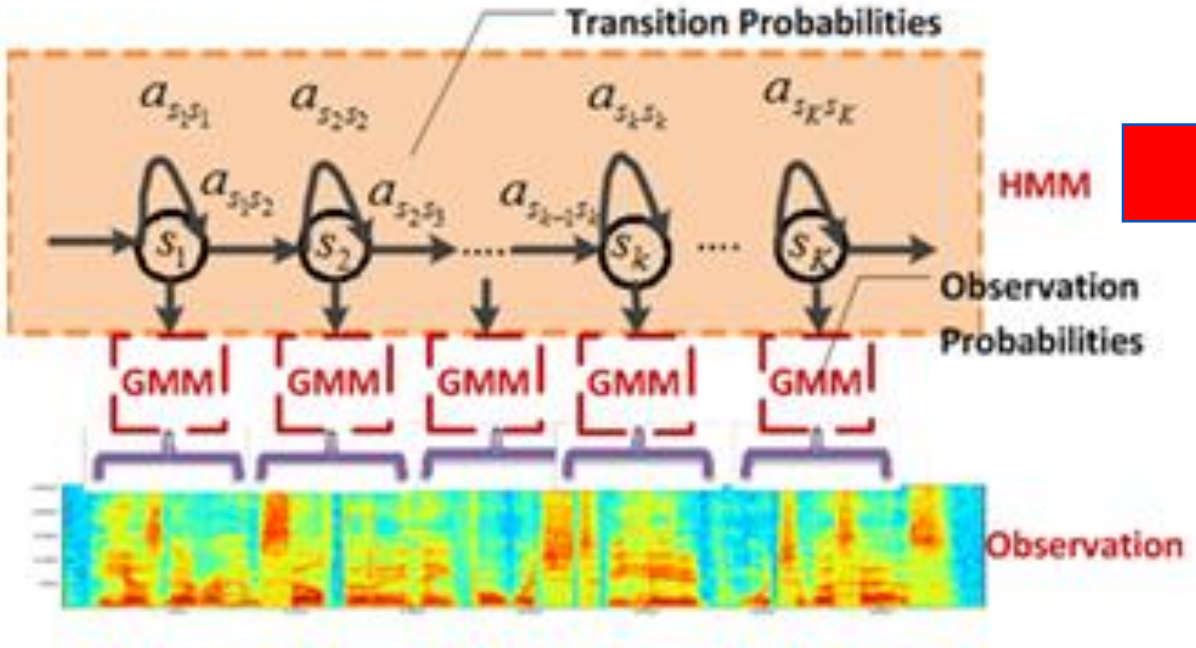
Использование нейронных сетей как классификаторов

Было:

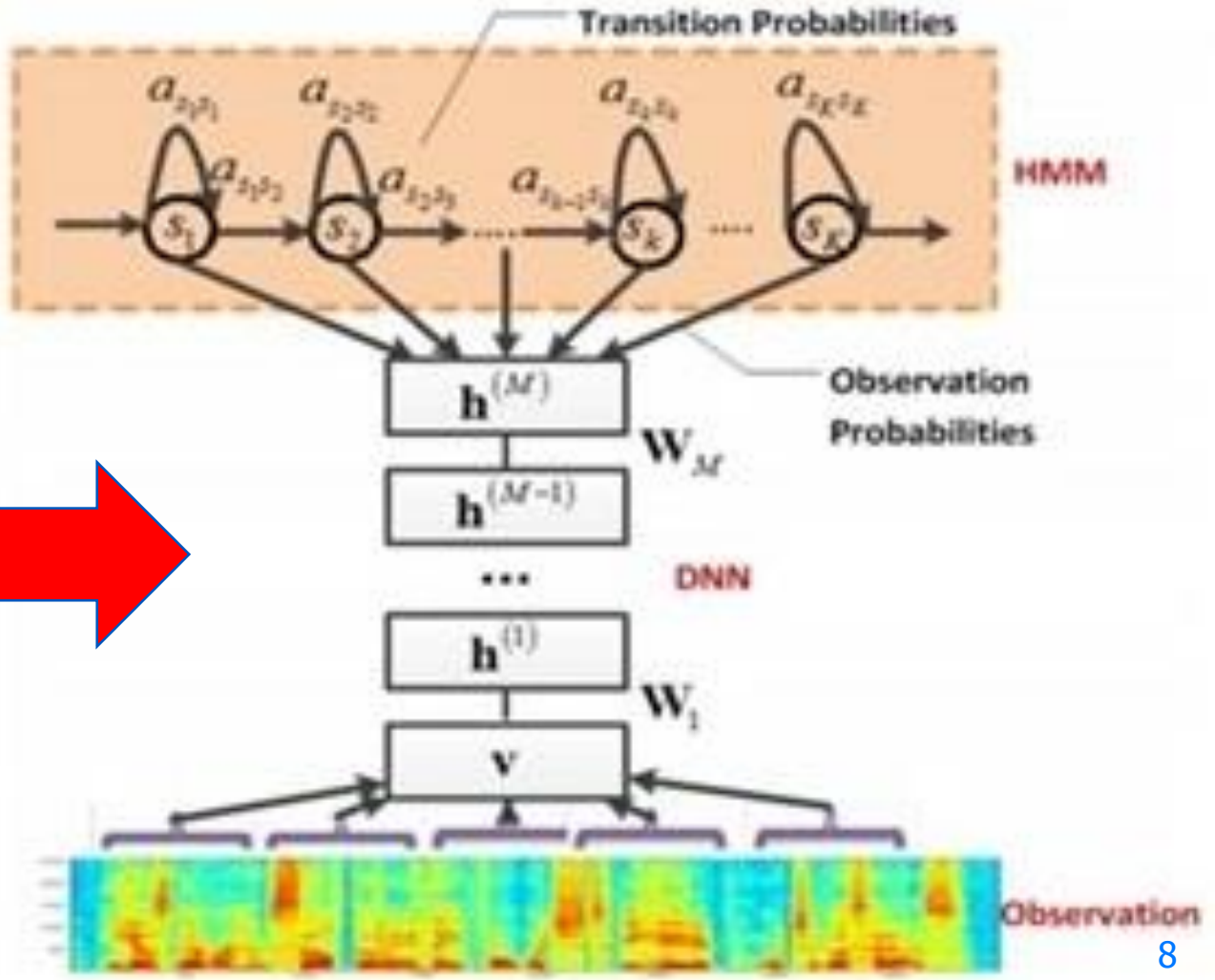


Использование нейронных сетей как классификаторов

Было:

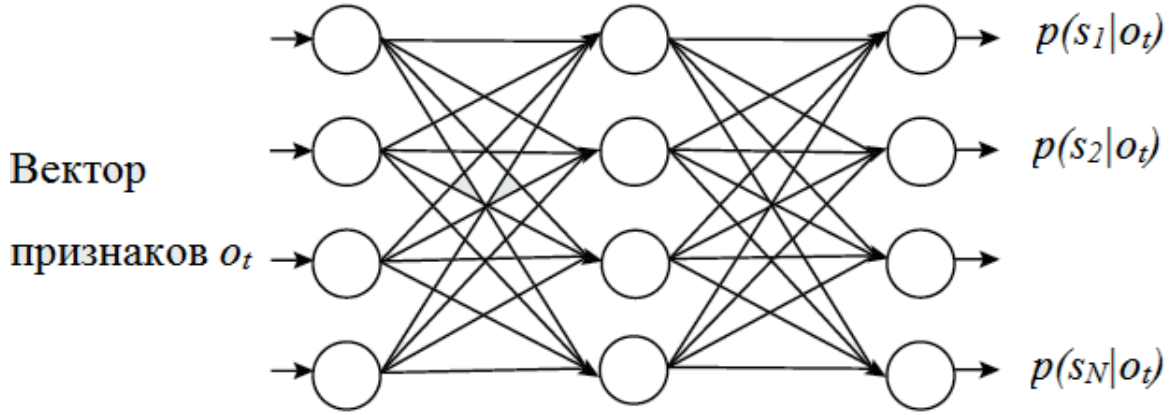


Стало:



Использование нейронных сетей как классификаторов

Как использовать ANN в системе распознавания?



- Выходы классификатора на базе ANN обычно трактуют, как апостериорные вероятности классов (в нашем случае – состояний HMM): $P(s_i|o_t)$
- А декодер должен получать от акустической модели правдоподобия: $p(o_t|s_i)$
- Как преобразовать одно в другое? Формула Байеса: $p(o_t|s_i) = \frac{P(s_i|o_t)}{P(s_i)} p(o_t)$
- Второй сомножитель не зависит от состояния! $\frac{P(s_i|o_t)}{P(s_i)}$ – псевдо-правдоподобия

Использование нейронных сетей как классификаторов

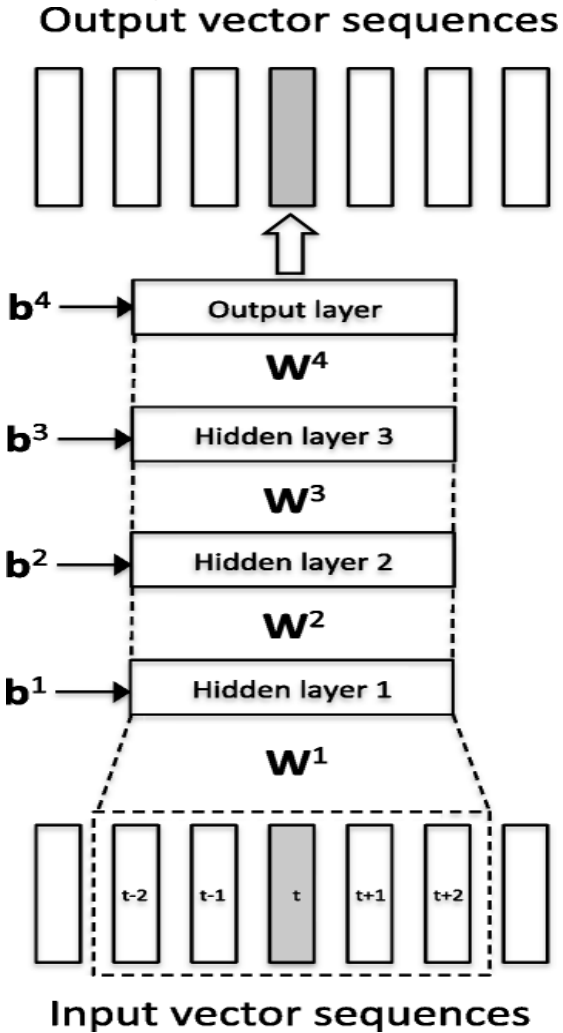
Как обучать ANN для системы распознавания речи?

- Подготовить базу обучающих фонограмм и их текстов
- Использовать существующую акустическую модель для РАЗМЕТКИ фонограмм на отдельные состояния HMM (Витерби-выравнивание, forced alignment)
- Сопоставить каждому вектору признаков выходной one-hot вектор, в котором 1 соответствует состоянию, указанному разметкой на данном кадре
- Обучить ANN по этому набору входов и выходов
- Оценить априорные вероятности состояний $p(s_i)$
  - Можно использовать частоты встречаемости состояний в разметке
  - Или посчитать среднюю апостериорную вероятность данного состояния на всех наблюдениях обучающей базы
- При необходимости переразметить и повторить

Использование нейронных сетей как классификаторов

Как обучать ANN для системы распознавания речи?

- Frame stacking (splicing):



План лекции

- Использование нейронных сетей в качестве классификаторов
- Гибридные и тандемные системы распознавания
- Обучение DNN-HMM систем распознавания
- Последовательно-дискриминативное обучение
- Адаптация систем распознавания речи на основе нейронных сетей

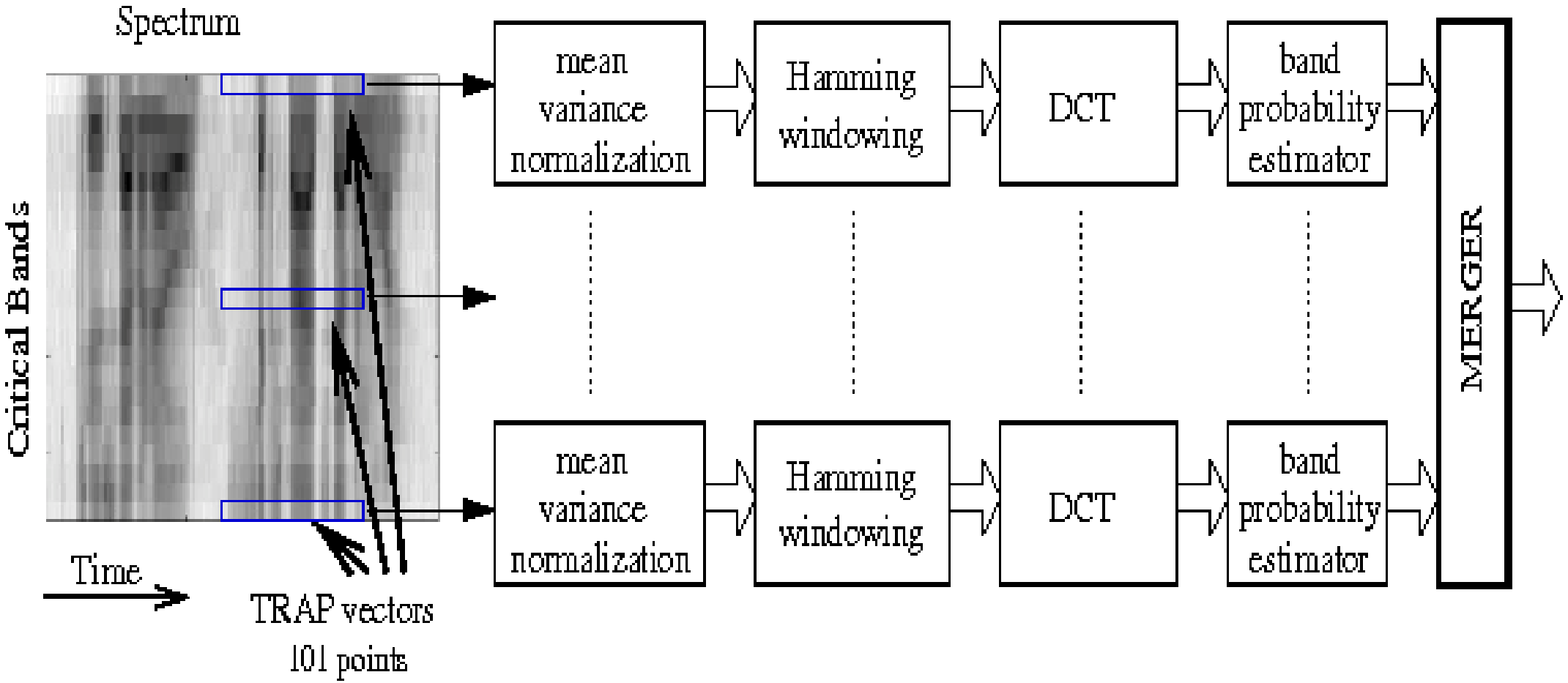
Гибридные и тандемные системы распознавания

Существуют два исторически различных подхода:

- В тандемной системе нейронная сеть используется в качестве генератора новых высокоуровневых признаков, а классификатор – GMM
- Известные варианты тандемных систем:
  - TRAP-признаки (H.Hermansky, 2003)
  - LC-RC-признаки (P. Schwarz, 2008)
  - Bottleneck-признаки
- В гибридной системе ANN используются непосредственно в качестве классификаторов состояний (псевдо-правдоподобия идут прямиком в декодер)
  - На 2017-й год большинство систем распознавания в мире – **гибриды** с разными типами ANN
  - До 2011-го года классами всегда были сами фонемы или состояния HMM для фонем
  - Первый гибрид с NN-классификатором связанных состояний трифонов: CD-DNN-HMM (G.Dahl etal, 2011)

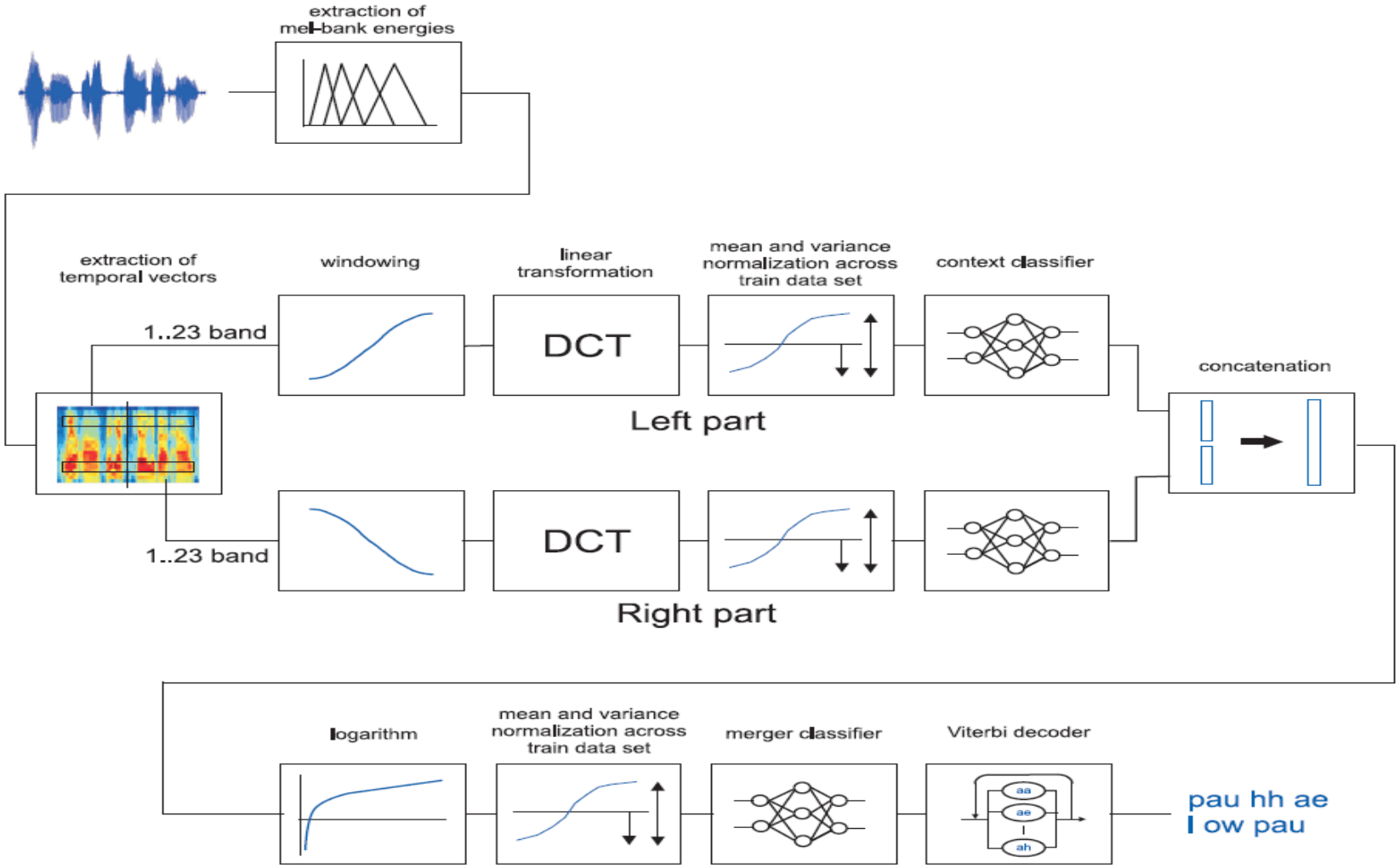
Тандемные системы распознавания

TempoRAl Patterns (TRAP) признаки (H.Hermansky, 2003):



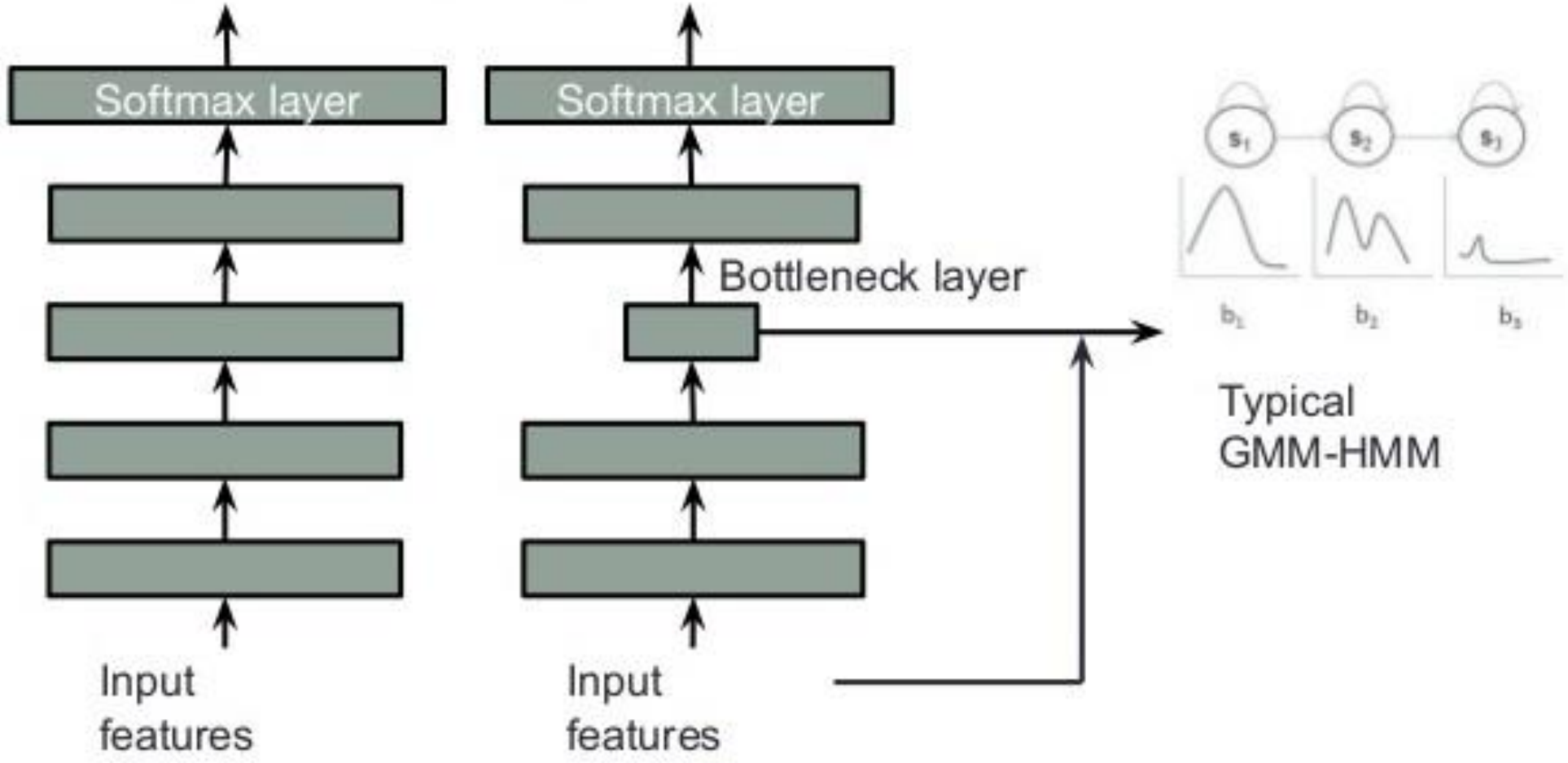
Тандемные системы распознавания

LC-RC-признаки (P. Schwarz, 2008):



Тандемные системы распознавания

Bottleneck-признаки (выходы слоя «узкого горла»):



Гибридные системы распознавания

CD-DNN-HMM-гибрид (G.Dahl etal, 2011) :

- В качестве классификатора используется полносвязная DNN на 5-10 слоев
- На входе MFCC или log-mel-fbanks с нескольких фреймов вокруг текущего
- К примеру 11 кадров 13-мерных MFCC дают размер входного слоя 11*13=143
- Размер выходного слоя = числу связанных состояний трифонных HMM (сенонов, 5-10K)
- Размеры скрытых слоев 512-2048
- Критерий обучения: кросс-энтропия, средняя на фрейм по всем обучающим данным

$$CE = \frac{1}{M}\sum_{k=1}^{M}\left(-\sum_{i=1}^{N} t_i^{(k)} \log y_i^{(k)}\right)$$

- где $y_i^{(k)}$- i-й выход DNN, $t_i^{(k)}$- целевое значение на k–м фрейме обучающей базы

Гибридные системы распознавания

CD-DNN-HMM-гибрид (G.Dahl etal, 2011) :

- Для обучения DNN требуется иметь разметку обучающей базы на сеноны. Поэтому предварительно необходимо обучить какие-то трифонные HMM (например, на базе GMM-HMM) и разметить ими базу (forced alignment, алгоритм Витерби)
- Вероятности переходов в HMM игнорируются (все переходы считаются равновероятными). Если их обучать, точность не повышается
- Сравнение результатов на базе Switchboard Hub5 eval2000 (Seide etal. 2011):

Модель	WER на монофонах	WER на сенонах (9304)
CD-GMM-HMM (BMMI)	---	23.6%
CD-DNN-HMM (7x2048)	34.9%	17.1%

План лекции

- Использование нейронных сетей в качестве классификаторов
- Гибридные и тандемные системы распознавания
- Обучение DNN-HMM систем распознавания
- Последовательно-дискриминативное обучение
- Адаптация систем распознавания речи на основе нейронных сетей

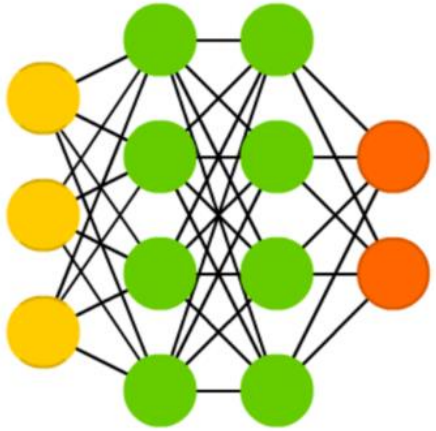
Обучение DNN-HMM для распознавания речи

- Что было до 2006 года:
  - Уже был известен алгоритм обратного распространения ошибки
  - Но: никто толком не умел учить многослойные перцептроны (с числом скрытых слоев >1)
  - Основная причина: использование сигмоидальных функций активации и вызываемый ими эффект gradient vanishing.
  - Идея: хорошо инициализировать сеть (в области весов достаточно близкой к оптимальной точке)
- Жадное послойное предобучение (J.Hinton etal., 2006):
  - Restricted Boltzmann Machine (RBM), Contrastive Divergence (CD) - обучение (J.Hinton etal., 2006-2008)
  - Denoising Autoencoders (Y.Bengio etal., 2007-2008)
  - Дискриминативное предобучение (F.Seide, 2011)
- Функции активации с незатухающими производными, линейные связи
  - ReLU и ее модификации (leaky ReLU, noisy ReLU, ELU), maxout и т.п.
  - Residual connections (ResNets), Highway Networks и т.д.
  - Гейтинг (gating) – модуляция линейного слоя сигмоидой от другого линейного слоя

Обучение DNN-HMM для распознавания речи

Распространенные архитектуры гибридных ASR-систем

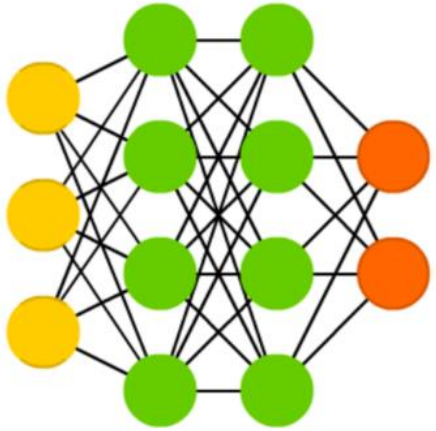
Полносвязные сети прямого распространения (MLP)

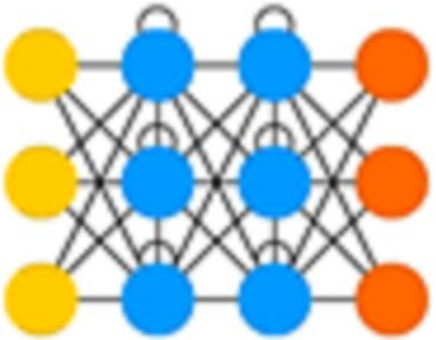


Обучение DNN-HMM для распознавания речи

Распространенные архитектуры гибридных ASR-систем

Полносвязные сети прямого распространения (MLP)



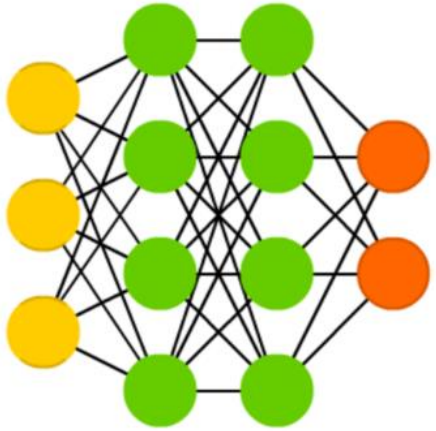


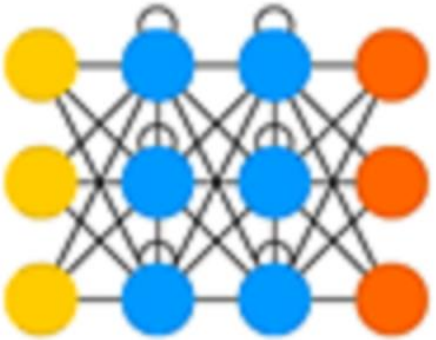
Рекуррентные сети (LTSM, GRU, BiLSTM), трансформеры

Обучение DNN-HMM для распознавания речи

Распространенные архитектуры гибридных ASR-систем

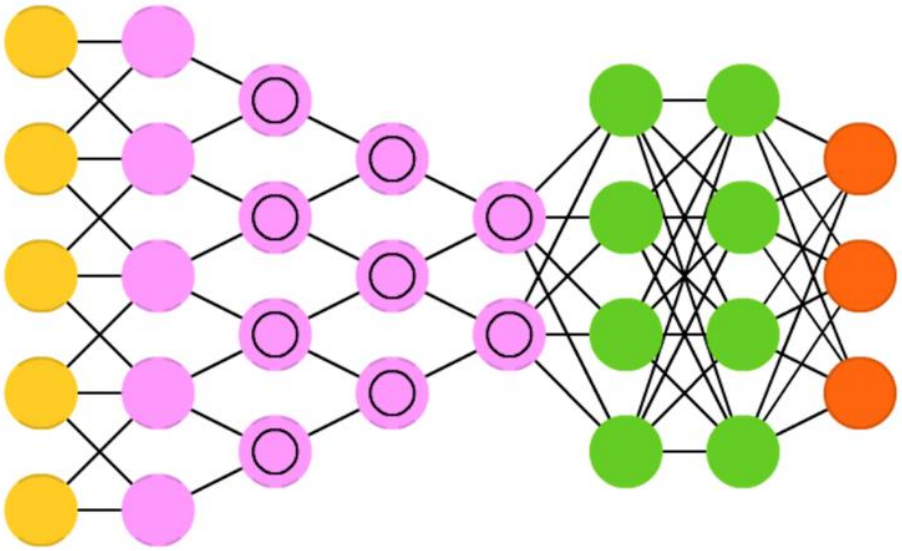
Полносвязные сети прямого распространения (MLP)





Рекуррентные сети (LTSM, GRU, BiLSTM), трансформеры

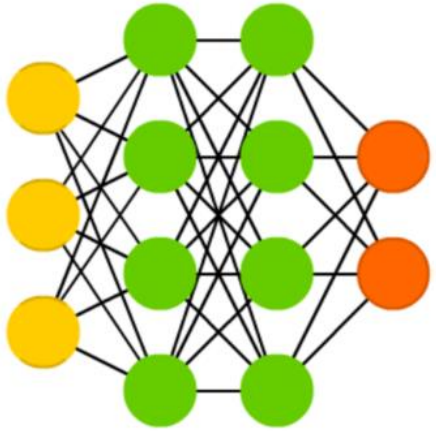
Сверточные сети

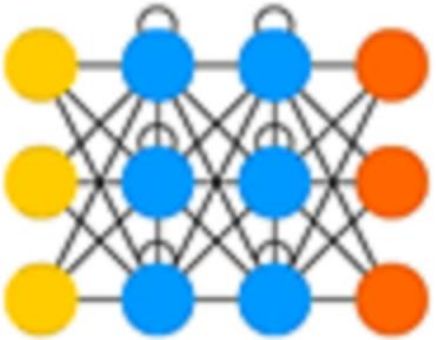


Обучение DNN-HMM для распознавания речи

Распространенные архитектуры гибридных ASR-систем

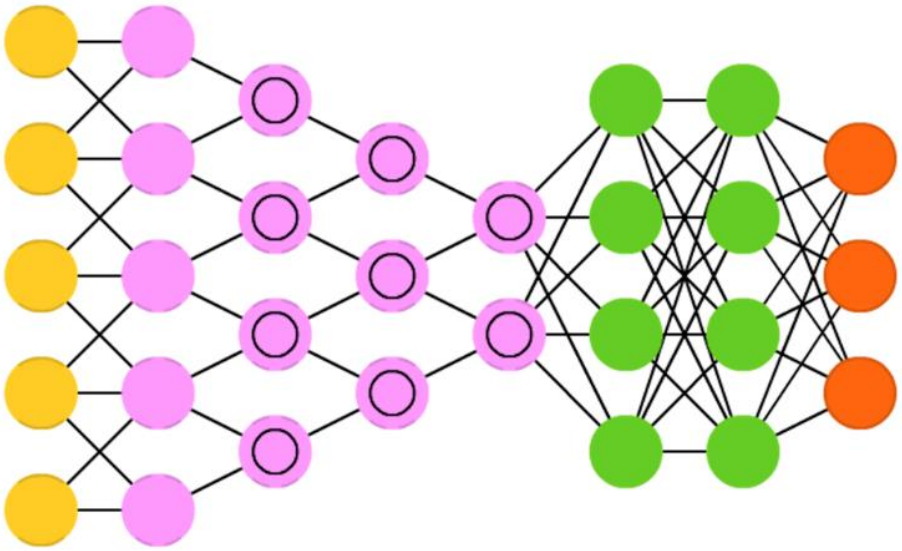
Полносвязные сети прямого распространения (MLP)





Рекуррентные сети (LTSM, GRU, BiLSTM), трансформеры

Сверточные сети

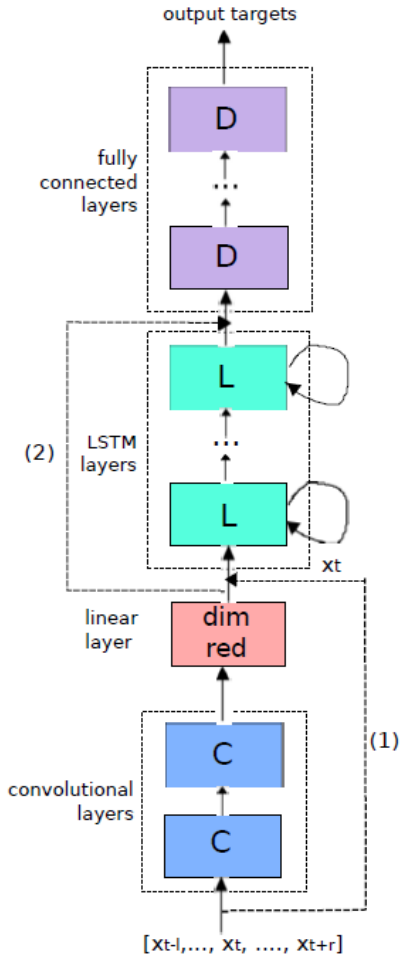


Различные комбинации вышеперечисленных вариантов

Обучение DNN-HMM для распознавания речи

Распространенные архитектуры гибридных ASR-систем

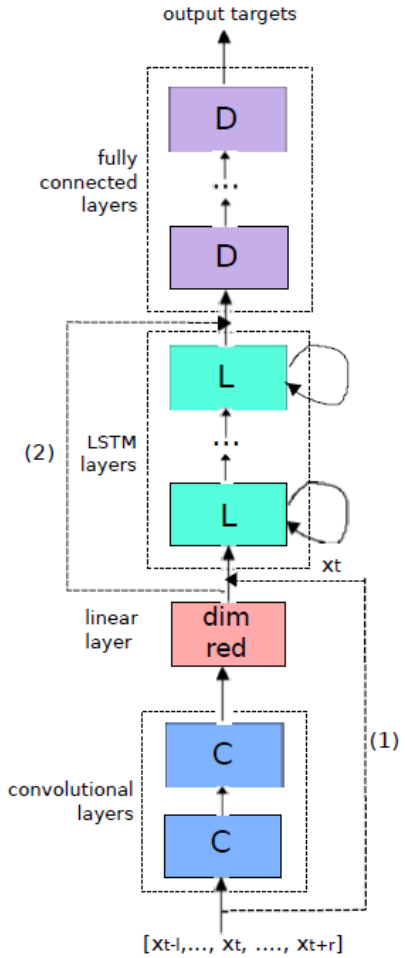
CLDNN (Google, 2015)



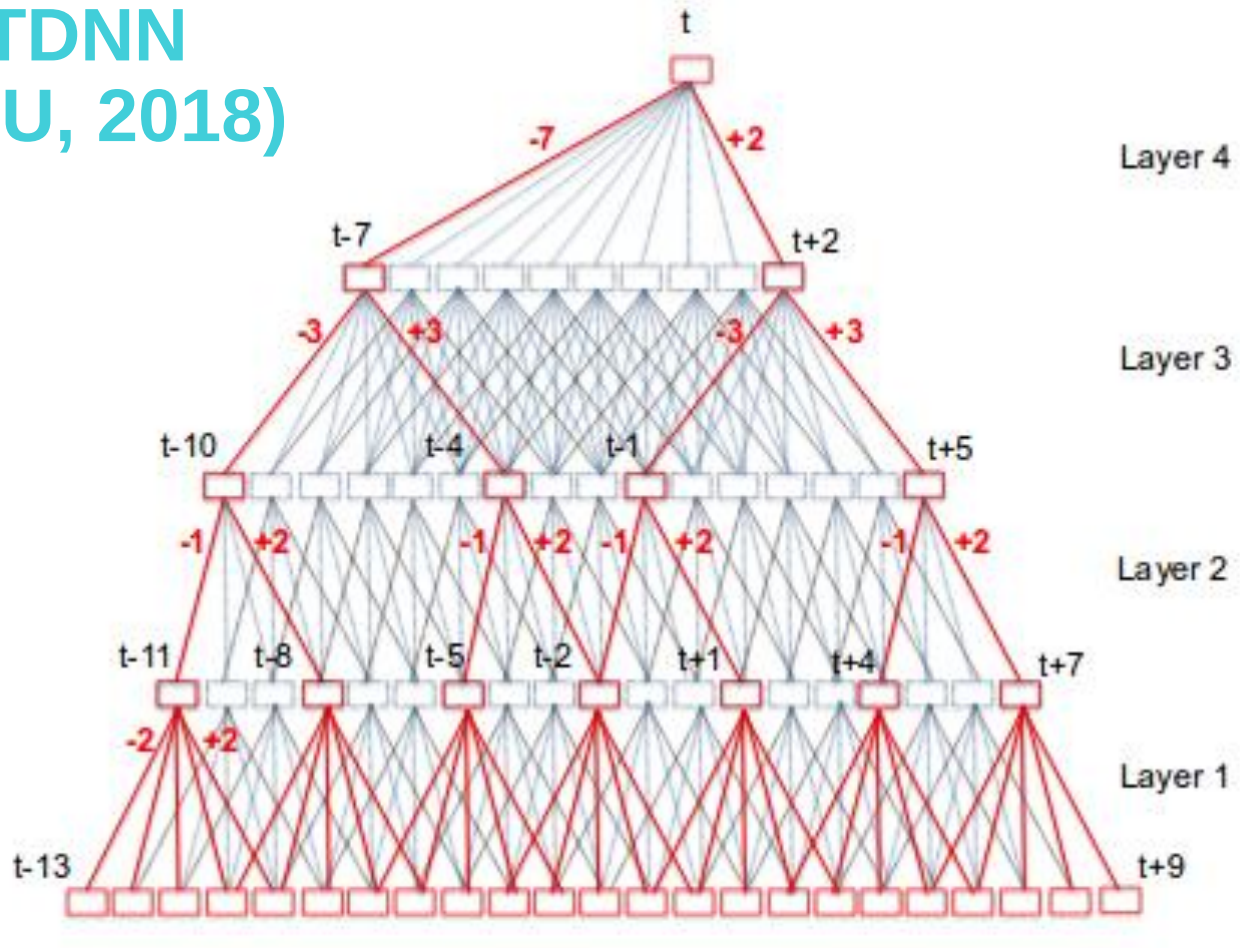
Обучение DNN-HMM для распознавания речи

Распространенные архитектуры гибридных ASR-систем

CLDNN (Google, 2015)



TDNN (JHU, 2018)



План лекции

- Использование нейронных сетей в качестве классификаторов
- Гибридные и тандемные системы распознавания
- Обучение DNN-HMM систем распознавания
- Последовательно-дискриминативное обучение
- Адаптация систем распознавания речи на основе нейронных сетей

Sequence discriminative training

- Кросс-энтропийное обучение минимизирует среднюю ошибку классификации на фрейм
- А хочется учить сеть так, чтобы минимизировать WER!
- Идея похожа на дискриминативное обучение GMM-HMM: минимизировать ошибку на целой последовательности слов, отдаляя истинную последовательность от остальных
- Распространенные критерии (функции потерь):
  - MMI минимизирует ошибку на уровне целых фраз
  - MWE/MPE минимизирует ошибку на уровне последовательностей слов/фонем
  - state-level Minimum Bayes Risk (sMBR) минимизирует ошибку на уровне последовательности состояний HMM
- Для вычисления критериев и их производных по выходам сети используются модификации Forward-Backward-алгоритма на сетях (lattice) числителя и знаменателя
- Типичный сценарий: 1) кросс-энтропийное обучение, 2) генерация word lattices для всей обучающей базы, 3) последовательно-дискриминативное **до**обучение.

Sequence discriminative training

Lattice-free MMI (LF-MMI) :

- Для расчета статистик знаменателя требуются сети знаменателя для каждой обучающей фразы. Их строить долго
- Идея: заменим все сети знаменателя на простой граф, соответствующий n-граммной фонемной языковой модели (как правило, n=4)
- Тогда можно учить последовательно-дискриминативную модель с нуля, минуя стадию кросс-энтропийного обучения
- Поскольку граф знаменателя общий для всех фраз, обучение можно эффективнее организовать на GPU
- Точность получается сравнимой или выше, чем в базовом сценарии CE=>MMI
- Недавно появились LF-аналоги других критериев, например, bMMI и sMBR

План лекции

- Использование нейронных сетей в качестве классификаторов
- Гибридные и тандемные системы распознавания
- Обучение DNN-HMM систем распознавания
- Последовательно-дискриминативное обучение
- Адаптация систем распознавания речи на основе нейронных сетей

Адаптация DNN-HMM систем распознавания речи

Все методы можно условно разделить на 3 класса

- Методы, не требующие изменения нейронной сети
  - Нормализация длины голосового тракта (VTLN)
  - fMLLR-преобразование признаков (требует наличия обученной GMM-HMM-модели распознавания!)
- Методы, меняющие архитектуру и/или параметры нейронной сети
  - Дообучение нейронной сети на адаптационных данных («консервативное» обучение)
  - Введение дополнительных слоев и обучение только их
  - Адаптация bottleneck-слоя
  - И т.д.
- Дикторо-осведомленное (speaker aware) обучение

Адаптация DNN-HMM систем распознавания речи

Консервативное обучение

- Данных для адаптации мало, а параметров много, потому важно избежать переобучения!
  - Ранний останов (требует выделения валидационного подмножества из адаптационных данных)
  - Уменьшение скорости обучения
  - Заморозка некоторых слоев (установка скорости обучения в ноль)
  - KLD-регуляризация: добавить в функцию потерь регуляризирующее слагаемое, штрафующее за «сильные отклонения» распределений на выходе адаптированной сети от распределений на выходе исходной (SI)

$$\hat{L} = (1-\rho)L\left(\theta; X^{adapt}, Y^{adapt}\right) + \rho L_{KLD}\left(\theta, \theta_{SI}; X^{adapt}, Y^{adapt}\right),$$

где

$$L_{KLD} = E_{x\sim X^{adapt}}\{KLD(P(y|x)||P_{SI}(y|x)\}$$

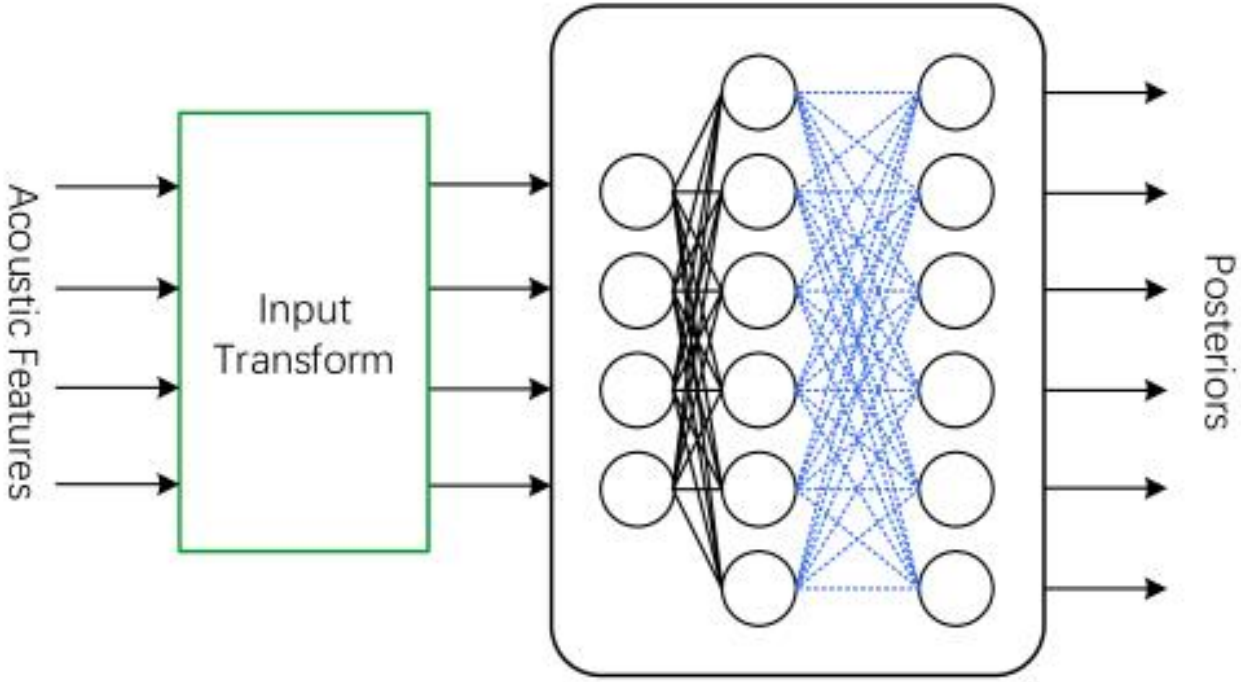
  - Если исходная функция потерь L – кросс-энтропийная, то это эквивалентно дообучению сети по адаптационным данным с soft-таргетами:

$$\hat{y}_{tj} = (1-\rho)y_{tj} + \rho P_{SI}(s_j|x)$$

Адаптация DNN-HMM систем распознавания речи

Адаптация с помощью дополнительных слоев:

- Linear Input Network (LIN):
  - Перед входным слоем вставляется дополнительный линейный слой
  - Инициализация единичной матрицей
  - Смысл: преобразовать входные признаки нового диктора к тому, к чему сеть «привыкла»
  - Все слои сети замораживаются и доучивается только входной
  - По смыслу очень похоже на fMLLR
  - Подходит для SAT-обучения



- Linear Hidden Network (LHN): линейный слой добавляется в середину сети
- Linear Output Network (LON): линейный слой добавляется прямо перед софтмаксом

Адаптация DNN-HMM систем распознавания речи

Адаптация bottleneck-слоя:

- Адаптируемый слой можно не внедрять в сеть искусственно, а «найти» в самой сети
- Это можно сделать с помощью сингулярного разложения (SVD) матрицы скрытого слоя:
- Пусть есть слой $y = f(Wx + b)$. Разложим его матрицу следующим образом:

$$W_{m\times n} = U_{m\times m}\Sigma_{m\times n}V^T_{n\times n} \approx \hat{U}_{m\times r}\hat{\Sigma}_{r\times r}\hat{V}^T_{r\times n}, \qquad r \ll m, n$$

- Тогда $y \approx f\left(\hat{U}\hat{\Sigma}\hat{V}x + b\right)$, т.е. исходный слой можно приближенно представить последовательностью из трех слоев: $\mathrm{v} = \hat{V}x$, $\mathrm{z} = \hat{\Sigma}v$, $\mathrm{y} = f(\hat{U}z + b)$
- Средний слой – линейный слой узкого горла (bottleneck layer)
- Можно адаптировать только его ($\hat{\Sigma}$), заморозив всю остальную сеть
- В нем мало слоев, поэтому риск переобучения минимален
- Можно адаптироваться на очень малых объемах целевых данных!

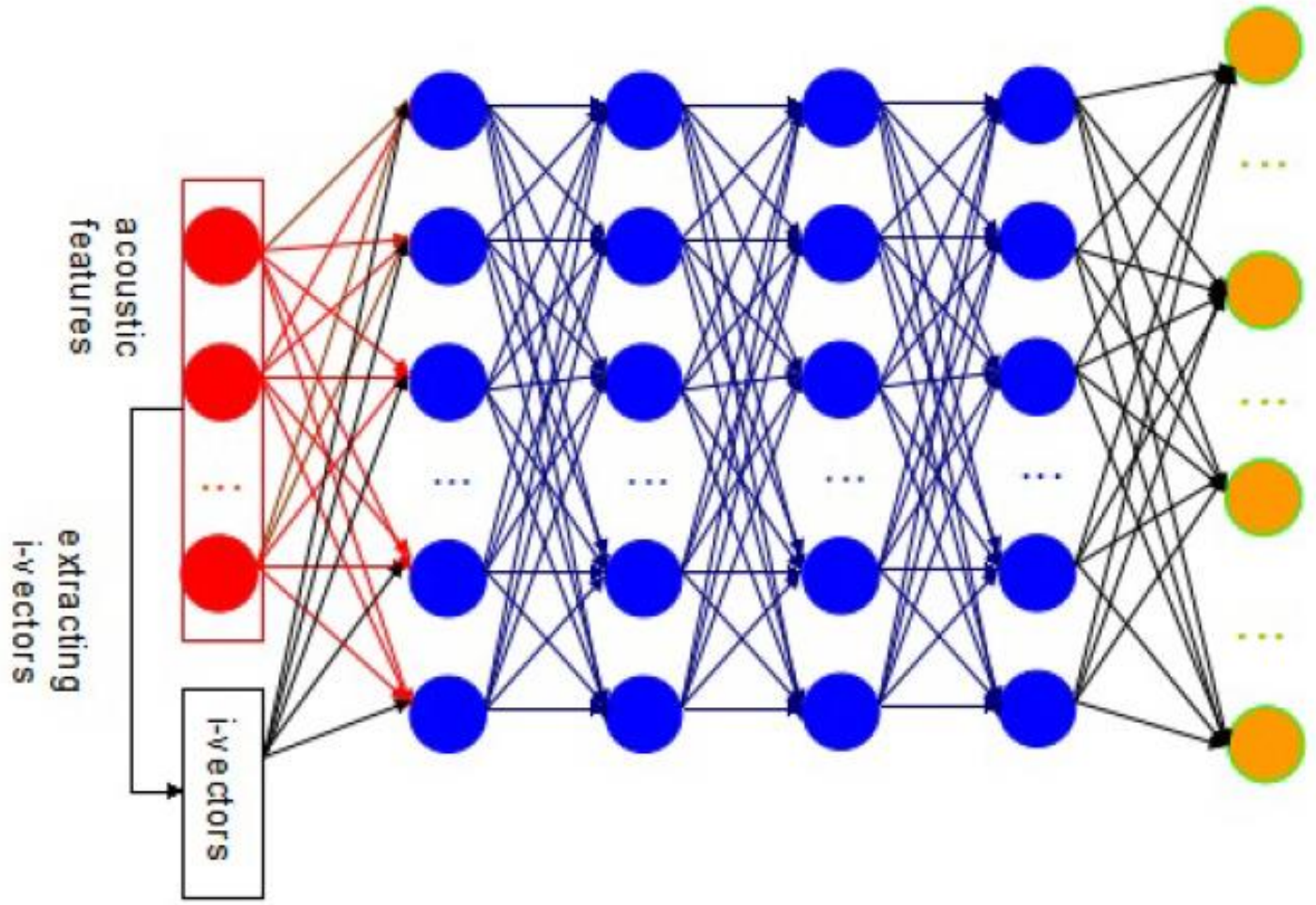
Адаптация DNN-HMM систем распознавания речи

Дикторооосведомленное обучение:

- Идея: если уметь извлекать из данных информацию о дикторе и подавать ее на вход сети, то адаптация не потребуется!
- Это надо делать и при обучении, и при распознавании
- Извлечение из речи информации о дикторе – центральная задача голосовой биометрии
- Можно воспользоваться достижениями из этой «родственной» области!
- Современные системы верификации/идентификации диктора извлекают «эмбеддинг», характеризующий голос диктора, всего по нескольким секундам речи
- Наиболее распространенные эмбеддинги: i-векторы, x-векторы
- Решение: расширяем входной слой сети и подаем эмбеддинг текущего диктора вместе с его признаками

Адаптация DNN-HMM систем распознавания речи

Дикторооосведомленное обучение:



Спасибо за внимание!

Вопросы?