Automatic speech recognition

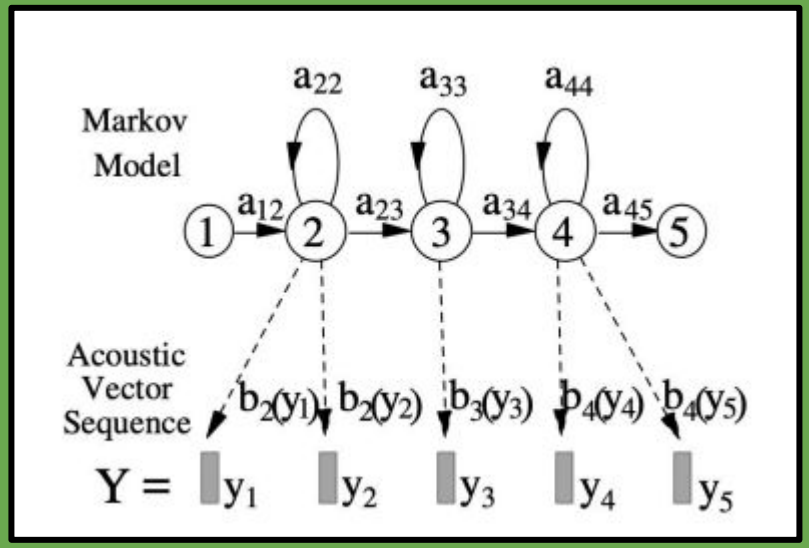
Сизов Михаил, 212

Содержание

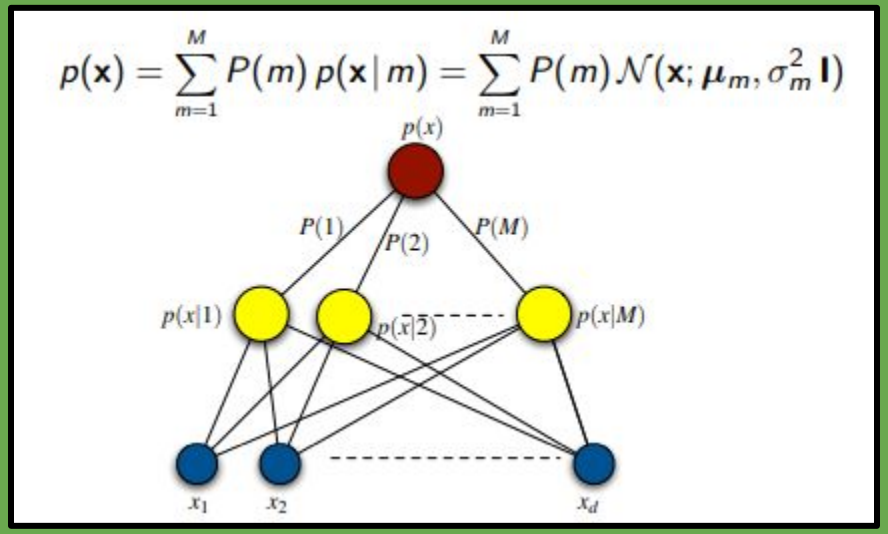
- Что было до нейросетей
- Как устроен Deep Speech
- CTC Loss
- Beam Search
- Языковая модель
- Результаты экспериментов

Традиционные методы распознавания речи

Марковские цепи



Модели Гауссовской смеси



Минусы традиционных методов

- Требуют принудительного выравнивания
- Низкая точность
- Неустойчивость к шумам
- Необходимо подготовить фонетический набор

Как устроен Deep Speech

Структура Deep Speech

Тренировочные данные: $\mathcal{X} = \{(x^{(1)}, y^{(1)}), (x^{(2)}, y^{(2)}), \ldots\}$

Вход: $x_t^{(i)}, t = 1, \ldots, T^{(i)}$

Выход:

$\hat{y}_t = \mathbb{P}(c_t|x)$, where $c_t \in \{\text{a,b,c}, \ldots, \text{z}, \textit{space}, \textit{apostrophe}, \textit{blank}\}$

Первые 3 слоя:

$$h_t^{(l)} = g(W^{(l)} h_t^{(l-1)} + b^{(l)})$$

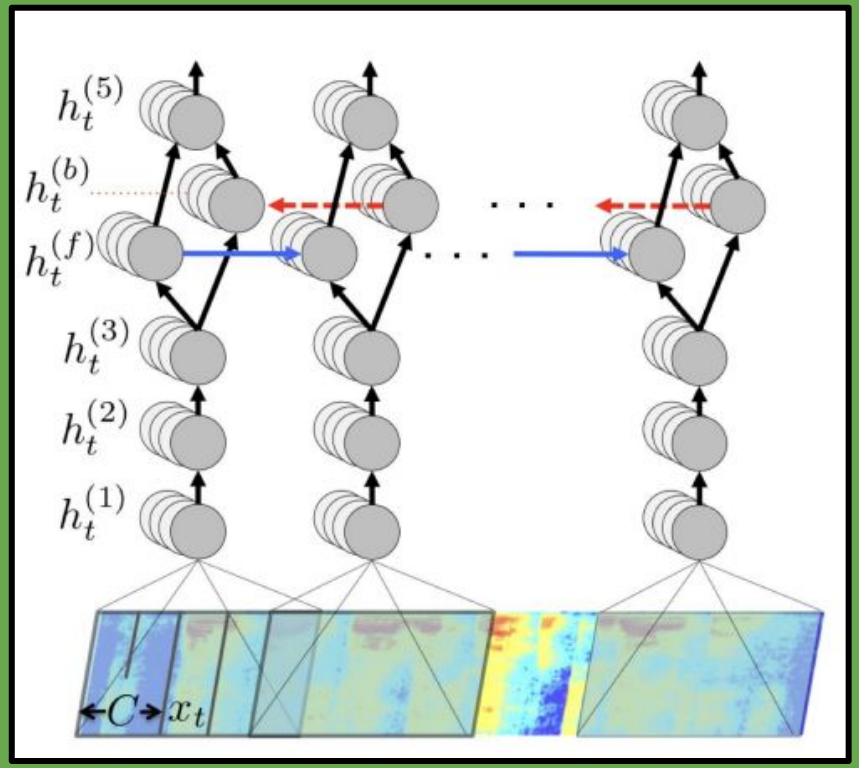
Четвертый слой:

$$h_t^{(f)} = g(W^{(4)} h_t^{(3)} + W_r^{(f)} h_{t-1}^{(f)} + b^{(4)})$$
$$h_t^{(b)} = g(W^{(4)} h_t^{(3)} + W_r^{(b)} h_{t+1}^{(b)} + b^{(4)})$$
$$h_t^{(4)} = h_t^{(f)} + h_t^{(b)}$$

Пятый слой:

$$h_t^{(5)} = g\left(W^{(5)} h_t^{(4)} + b^{(5)}\right)$$

Где $g(z) = \min\{\max\{0, z\}, 20\}$



Структура Deep Speech

Шестой слой:

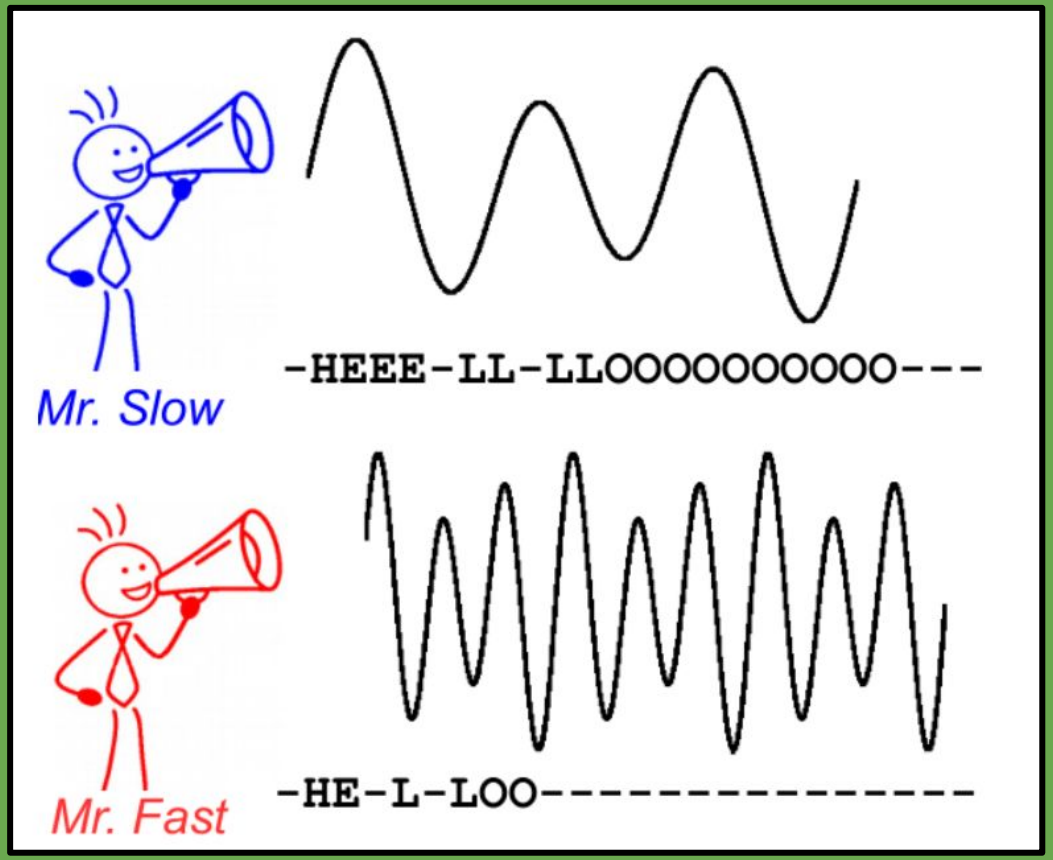
$$h_{t,k}^{(6)} = \hat{y}_{t,k} \equiv \mathbb{P}(c_t = k|x) = \frac{\exp(W_k^{(6)} h_t^{(5)} + b_k^{(6)})}{\sum_j \exp(W_j^{(6)} h_t^{(5)} + b_j^{(6)})}$$

Регуляризация: Dropout 5-10%

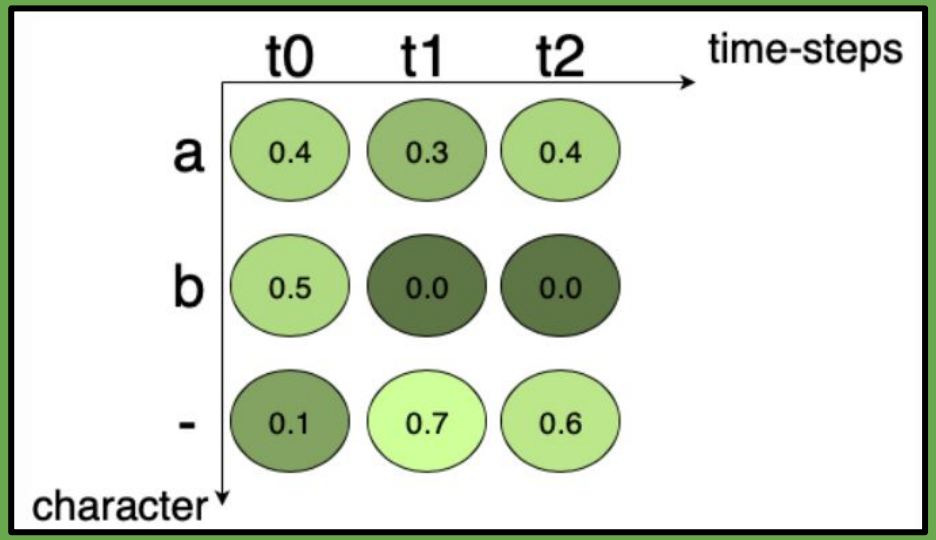
Оптимизатор: Nesterov gradient method

CTC Loss

Нарезка слова по timestamp



Матрица вероятностей



P(aaa) = 0.4 * 0.3 * 0.4 = 0.048

P(aa-) = 0.4 * 0.3 * 0.6 = 0.072

P(a--) = 0.4 * 0.7 * 0.6 = 0.168

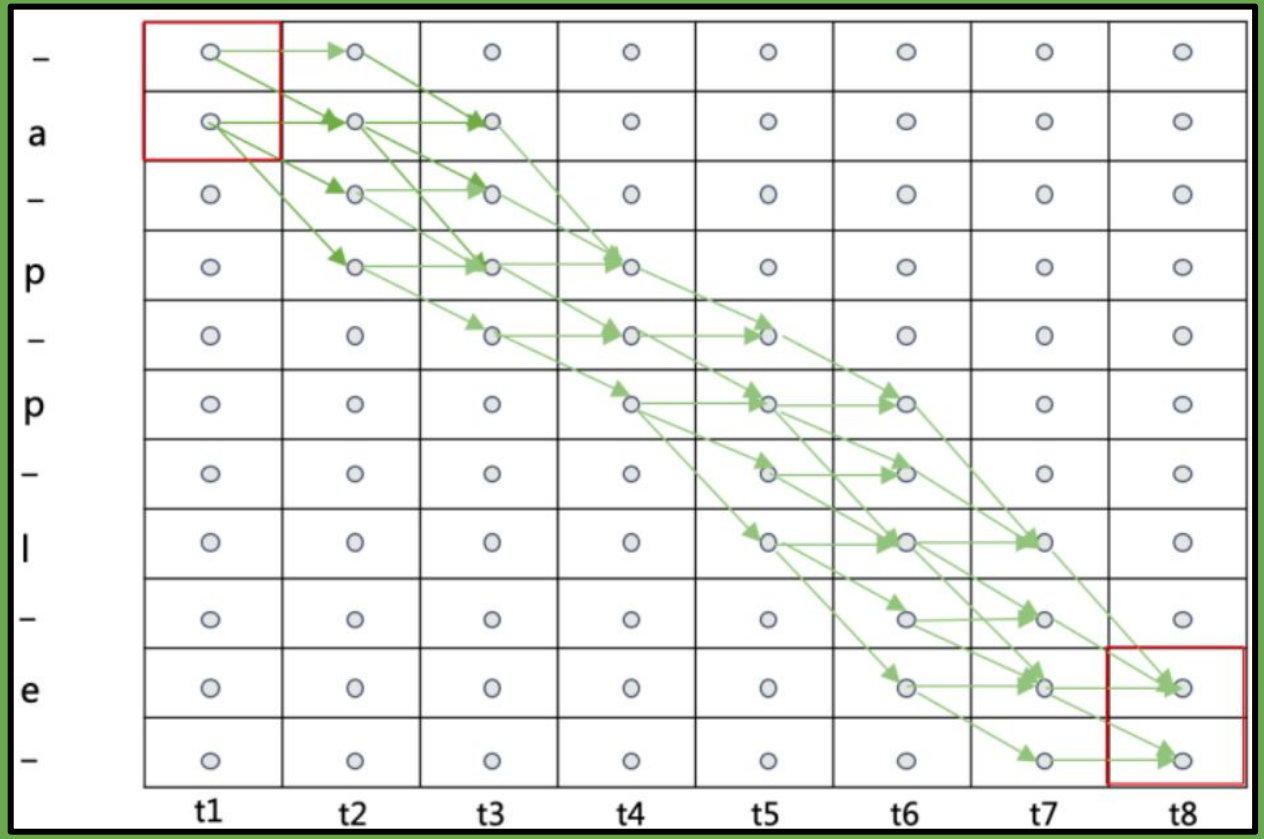
P(-a-) = 0.1 * 0.3 * 0.6 = 0.018

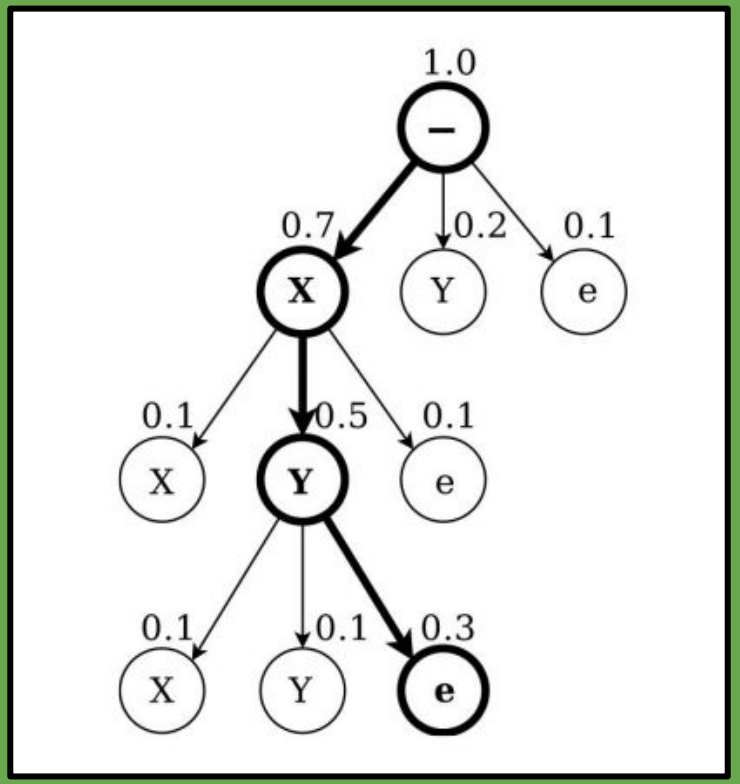
P(-aa) = 0.1 * 0.3 * 0.4 = 0.012

P(--a) = 0.1 * 0.7 * 0.4 = 0.028

Ответ для GT = a: 0.048 + 0.072 + 0.168 + 0.018 + 0.012 + 0.028 = 0.0346

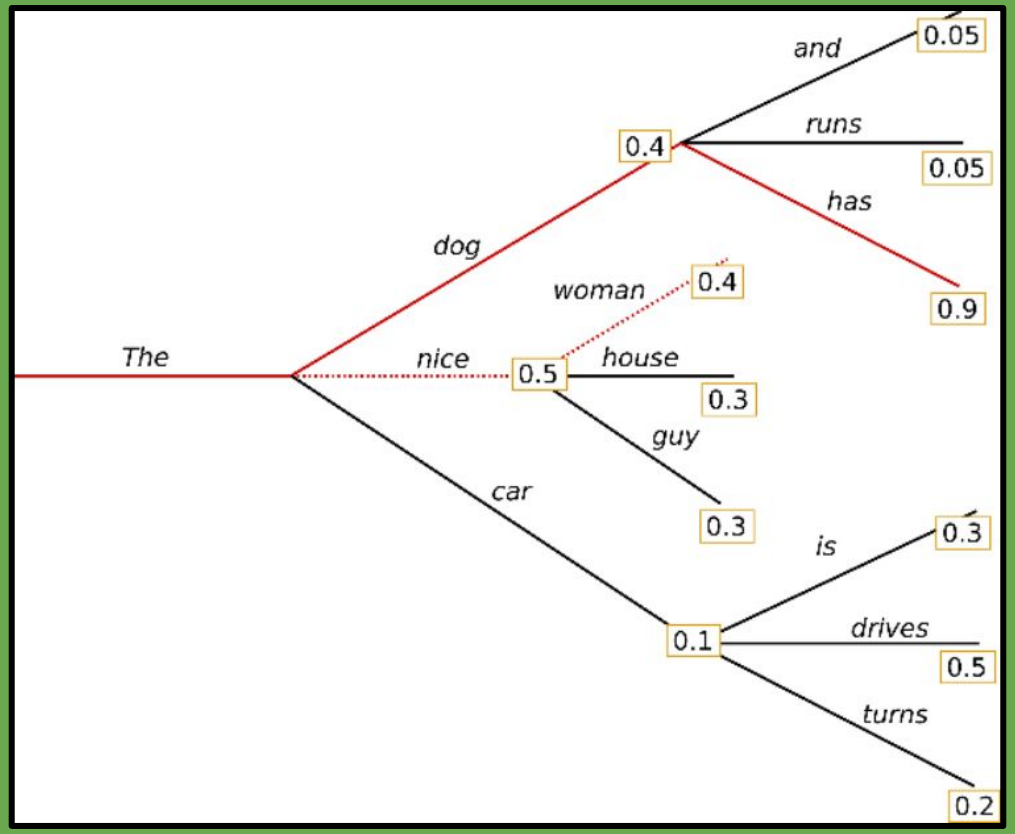
Декодирование





Префиксный алгоритм

Beam search и языковые модели


The
dog
nice
car
0.4
0.5
0.1
and
runs
has
woman
house
guy
is
drives
turns
0.05
0.05
0.9
0.4
0.3
0.3
0.3
0.5
0.2

Beam search

Language model

RNN output	Decoded Transcription
what is the weather like in bostin right now	what is the weather like in boston right now
prime miniter nerenr modi	prime minister narendra modi
arther n tickets for the game	are there any tickets for the game

$$Q(c) = \log(\mathbb{P}(c|x)) + \alpha \log(\mathbb{P}_{\text{lm}}(c)) + \beta \, \text{word\_count}(c)$$

Вероятность по CTC

Вероятность с точки зрения языковой модели

Максимизируем с помощью beam search, коэффициенты α и β оцениваем по кросс-валидации

Model	SWB	CH	Full
Vesely et al. (GMM-HMM BMMI) [44]	18.6	33.0	25.8
Vesely et al. (DNN-HMM sMBR) [44]	12.6	24.1	18.4
Maas et al. (DNN-HMM SWB) [28]	14.6	26.3	20.5
Maas et al. (DNN-HMM FSH) [28]	16.0	23.7	19.9
Seide et al. (CD-DNN) [39]	16.1	n/a	n/a
Kingsbury et al. (DNN-HMM sMBR HF) [22]	13.3	n/a	n/a
Sainath et al. (CNN-HMM) [36]	11.5	n/a	n/a
Soltau et al. (MLP/CNN+I-Vector) [40]	**10.4**	n/a	n/a
Deep Speech SWB	20.0	31.8	25.9
Deep Speech SWB + FSH	12.6	**19.3**	**16.0**

Результаты экспериментов

System	Clean (94)	Noisy (82)	Combined (176)
Apple Dictation	14.24	43.76	26.73
Bing Speech	11.73	36.12	22.05
Google API	6.64	30.47	16.72
wit.ai	7.94	35.06	19.41
Deep Speech	**6.56**	**19.06**	**11.85**


СПАСИБО ЗА ВНИМАНИЕ!