Speech recognition

Сухарьков Александр, 171

То, что было

1. Скрытая Марковская Модель
2. Conditional Random Fields

Основные проблемы:

1. Требуются предварительные знания
2. Нужно делать (иногда сомнительные) предположения о зависимостях

Чем можно исправить?

RNN - без предварительных знаний, устойчива к шуму.

Но также имеет некоторые проблемы:

1. Данные должны быть пресегментированны
2. Выход должен быть обработан

Новый метод

Авторами статьи разработан метод, который использует RNN.

Его главное преимущество в том, что он не требует пресегментации данных и постобработки выхода.

Temporal Classification

$S \subset \mathcal{D}_{\mathcal{X} \times \mathcal{Z}}$

$\mathcal{X} = (\mathbb{R}^m)^* \quad \mathbf{x} = (x_1, x_2, ..., x_T)$

$\mathcal{Z} = L^* \quad \mathbf{z} = (z_1, z_2, ..., z_U)$

$h : \mathcal{X} \mapsto \mathcal{Z}$

$S' \subset \mathcal{D}_{\mathcal{X} \times \mathcal{Z}}$

$$LER(h, S') = \frac{1}{Z} \sum_{(\mathbf{x}, \mathbf{z}) \in S'} ED(h(\mathbf{x}))$$

Connectionist Temporal Classification

$\mathcal{N}_w : (\mathbb{R}^m)^T \mapsto (\mathbb{R}^n)^T$

$\mathbf{y} = \mathcal{N}_w(\mathbf{x})$

$L' = L \cup \{blank\}$

$$p(\pi|\mathbf{x}) = \prod_{t=1}^{T} y_{\pi_t}^t, \ \forall \pi \in L'^T \qquad (1)$$

Connectionist Temporal Classification

$$\mathcal{B} : {L'}^T \mapsto L^{\leq T}$$

$$\mathcal{B}(a - ab-) = \mathcal{B}(-aa - -abb) = aab$$

$$\mathbf{l} \in L^{\leq T}$$

$$p(\mathbf{l}|\mathbf{x}) = \sum_{\pi \in \mathcal{B}^{-1}(\mathbf{l})} p(\pi|\mathbf{x}) \quad (2)$$

$$h(\mathbf{x}) = \arg\max_{\mathbf{l} \in L^{\leq T}} p(\mathbf{l}|\mathbf{x})$$

Первый метод:

$$h(\mathbf{x}) \approx \mathcal{B}(\pi^*)$$
$$\text{where } \pi^* = \arg\max_{\pi \in N^t} p(\pi|\mathbf{x})$$

Прямая переменная

$$\alpha_t(s) \stackrel{\text{def}}{=} \sum_{\substack{\pi \in N^T: \\ \mathcal{B}(\pi_{1:t}) = \mathbf{l}_{1:s}}} \prod_{t'=1}^{t} y^{t'}_{\pi_{t'}} \quad (3)$$

$$p(\mathbf{l}|\mathbf{x}) = \alpha_T(|\mathbf{l}'|) + \alpha_T(|\mathbf{l}'| - 1) \quad (4)$$

Инициализация:

$$\alpha_1(1) = y^1_b$$
$$\alpha_1(2) = y^1_{\mathbf{l}_1}$$
$$\alpha_1(s) = 0, \ \forall s > 2$$

Рекурсия:

$$\alpha_t(s) = \begin{cases} \bar{\alpha}_t(s) y^t_{\mathbf{l}'_s} & \text{if } \mathbf{l}'_s = b \text{ or } \mathbf{l}'_{s-2} = \mathbf{l}'_s \\ \big(\bar{\alpha}_t(s) + \alpha_{t-1}(s-2)\big) y^t_{\mathbf{l}'_s} & \text{otherwise} \end{cases}$$

where

$$\bar{\alpha}_t(s) \stackrel{\text{def}}{=} \alpha_{t-1}(s) + \alpha_{t-1}(s-1).$$

Обратная переменная

$$\beta_t(s) \stackrel{\text{def}}{=} \sum_{\substack{\pi \in N^T: \\ \mathcal{B}(\pi_{t:T}) = \mathbf{l}_{s:|\mathbf{l}|}}} \prod_{t'=t}^{T} y_{\pi_{t'}}^{t'} \qquad (5)$$

Инициализация:

$$\begin{aligned} \beta_T(|\mathbf{l}'|) &= y_b^T \\ \beta_T(|\mathbf{l}'| - 1) &= y_{\mathbf{l}_{|\mathbf{l}|}}^T \\ \beta_T(s) &= 0, \ \forall s < |\mathbf{l}'| - 1 \end{aligned}$$

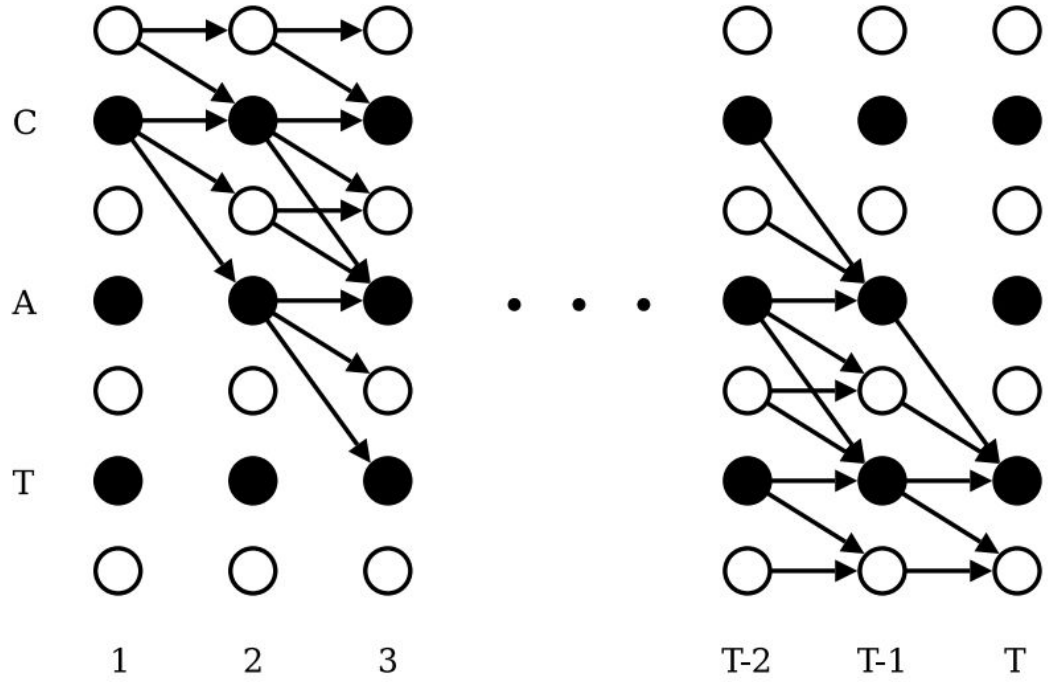
Рекурсия:

$$\beta_t(s) = \begin{cases} \bar{\beta}_t(s) y_{\mathbf{l}'_s}^t & \text{if } \mathbf{l}'_s = b \text{ or } \mathbf{l}'_{s+2} = \mathbf{l}'_s \\ \big(\bar{\beta}_t(s) + \beta_{t+1}(s+2)\big) y_{\mathbf{l}'_s}^t & \text{otherwise} \end{cases}$$

where

$$\bar{\beta}_t(s) \stackrel{\text{def}}{=} \beta_{t+1}(s) + \beta_{t+1}(s+1).$$

Иллюстрация алгоритма



Forward-Backward алгоритм

Переопределение для решения возможных проблем с рекурсией

$$C_t \stackrel{\text{def}}{=} \sum_s \alpha_t(s), \qquad \hat{\alpha}_t(s) \stackrel{\text{def}}{=} \frac{\alpha_t(s)}{C_t}$$

$$D_t \stackrel{\text{def}}{=} \sum_s \beta_t(s), \qquad \hat{\beta}_t(s) \stackrel{\text{def}}{=} \frac{\beta_t(s)}{D_t}$$

$$ln(p(\mathbf{l}|\mathbf{x})) = \sum_{t=1}^{T} ln(C_t)$$

Connectionist Temporal Classification

$$O^{ML}(S, \mathcal{N}_w) = -\sum_{(\mathbf{x},\mathbf{z})\in S} ln\big(p(\mathbf{z}|\mathbf{x})\big)$$

$$\frac{\partial O^{ML}(\{(\mathbf{x},\mathbf{z})\}, \mathcal{N}_w)}{\partial y_k^t} = -\frac{\partial ln(p(\mathbf{z}|\mathbf{x}))}{\partial y_k^t} \quad (6)$$

из (3) и (5) получаем:

$$\alpha_t(s)\beta_t(s) = \sum_{\substack{\pi\in\mathcal{B}^{-1}(\mathbf{l}):\\ \pi_t=\mathbf{l}'_s}} y^t_{\mathbf{l}'_s} \prod_{t=1}^{T} y^t_{\pi_t}$$

используя (1), получаем:

$$\frac{\alpha_t(s)\beta_t(s)}{y^t_{\mathbf{l}'_s}} = \sum_{\substack{\pi\in\mathcal{B}^{-1}(\mathbf{l}):\\ \pi_t=\mathbf{l}'_s}} p(\pi|\mathbf{x})$$

Connectionist Temporal Classification

$$p(\mathbf{l}|\mathbf{x}) = \sum_{s=1}^{|\mathbf{l}'|} \frac{\alpha_t(s)\beta_t(s)}{y^t_{\mathbf{l}'_s}}$$

дифференцируем: $$\frac{\partial p(\mathbf{l}|\mathbf{x})}{\partial y^t_k} = \frac{1}{{y^t_k}^2} \sum_{s \in lab(\mathbf{l},k)} \alpha_t(s)\beta_t(s) \qquad (7)$$

$\mathbf{l} = \mathbf{z}$ и подставляя (4) и (7) в (6): $$\frac{\partial O^{ML}(\{(\mathbf{x},\mathbf{z})\}, \mathcal{N}_w)}{\partial u^t_k} = y^t_k - \frac{1}{y^t_k Z_t} \sum_{s \in lab(\mathbf{z},k)} \hat{\alpha}_t(s)\hat{\beta}_t(s)$$

$$Z_t \overset{\text{def}}{=} \sum_{s=1}^{|\mathbf{l}'|} \frac{\hat{\alpha}_t(s)\hat{\beta}_t(s)}{y^t_{\mathbf{l}'_s}}$$

Deep Speech RNN

$$\mathcal{X} = \{(x^{(1)}, y^{(1)}), (x^{(2)}, y^{(2)}), \dots\}$$

$$x_t^{(i)}, t = 1, \dots, T^{(i)}$$

$$\hat{y}_t = \mathbb{P}(c_t | x)$$

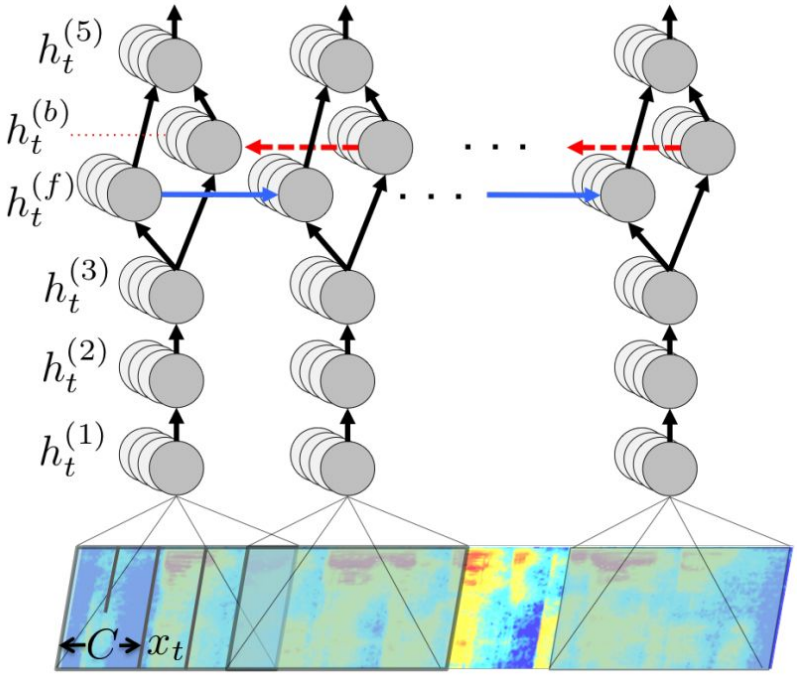
Первые три слоя: $h_t^{(l)} = g(W^{(l)} h_t^{(l-1)} + b^{(l)})$

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

Пятый слой: $h_t^{(5)} = g(W^{(5)} h_t^{(4)} + b^{(5)})$ where $h_t^{(4)} = h_t^{(f)} + h_t^{(b)}$

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

Структура RNN



Результаты экспериментов

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**

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**