Seq2Seq + Attention (BiLSTM Encoder + Bahdanau Attention + LSTM Decoder)

Архитектура

- **Encoder:** двунаправленный LSTM (BiLSTM)
- **Attention:** Bahdanau (additive)
- **Decoder:** однонаправленный LSTM
- **Вход в декодер:** конкатенация эмбеддинга токена и контекстного вектора из attention
- **Нормализация:** LayerNorm после конкатенации, дальше масштабирование и сдвиг:

$$\tilde{x}_t = \mathrm{LayerNorm}(x_t) \cdot a + b$$

---

1. Forward Pass

1. **Encoder (BiLSTM)**
   Для каждого входного токена $x_j \in \mathbb{R}^{d_{in}}$:

$$\overrightarrow{h}_j = \mathrm{LSTM}_f(x_j, \overrightarrow{h}_{j-1}), \quad \overleftarrow{h}_j = \mathrm{LSTM}_b(x_j, \overleftarrow{h}_{j+1}), \quad h_j = [\overrightarrow{h}_j;\ \overleftarrow{h}_j] \in \mathbb{R}^{2h}.$$

2. **Attention (Bahdanau)**
   На шаге t декодера, имеем скрытое состояние $s_{t-1} \in \mathbb{R}^h$ и $\{h_j\}$:

$$e_{tj} = v^\top \tanh\big(W_s\, s_{t-1} + W_h\, h_j + b_e\big), \quad \alpha_{tj} = \frac{\exp(e_{tj})}{\sum_k \exp(e_{tk})}, \quad c_t = \sum_j \alpha_{tj}\, h_j.$$

3. **Decoder Input**
   Эмбеддинг предыдущего токена $e_t \in \mathbb{R}^d$ и контекст $c_t \in \mathbb{R}^{2h}$:

$$x_t = \begin{bmatrix} e_t \\ c_t \end{bmatrix}, \quad \tilde{x}_t = \mathrm{LayerNorm}(x_t)\, a + b.$$

4. **Decoder (LSTM)**

$$(s_t, m_t) = \mathrm{LSTM}(\tilde{x}_t, (s_{t-1}, m_{t-1})),$$

   где $s_t \in \mathbb{R}^h, m_t \in \mathbb{R}^h$.

5. **Output & Loss**

$$z_t = W_{out}\, s_t + c_{out}, \quad \hat{y}_t = \mathrm{softmax}(z_t), \quad L_t = -y_t^\top \log(\hat{y}_t).$$

---

2. Backward Pass (шаг t)

2.1 Output Layer

$$\delta_t^{(y)} = \hat{y}_t - y_t, \quad \frac{\partial L_t}{\partial W_{out}} = \delta_t^{(y)}\, s_t^\top, \quad \frac{\partial L_t}{\partial c_{out}} = \delta_t^{(y)}, \quad \delta_t^{(s)} = W_{out}^\top \delta_t^{(y)}.$$

2.2 Decoder LSTM

Обозначения:
$a_t = W_{dec}\tilde{x}_t + U_{dec}s_{t-1} + b_{dec}$
разбиваем на гейты a^i, a^f, a^o, a^g .

Локальные ошибки гейтов:

$$\delta_t^{(o)} = (\delta_t^{(s)} \odot \tanh(m_t)) \odot (o_t \odot (1 - o_t)), \quad \delta_t^{(m)} = \delta_t^{(s)} \odot o_t \odot (1 - \tanh^2(m_t)),$$
$$\delta_t^{(g)} = (\delta_t^{(m)} \odot i_t) \odot (1 - g_t^2), \quad \delta_t^{(i)} = (\delta_t^{(m)} \odot g_t) \odot (i_t \odot (1 - i_t)), \quad \delta_t^{(f)} = (\delta_t^{(m)} \odot m_{t-1}) \odot (f_t \odot (1 - f_t)).$$

Собираем в один вектор $\delta_t^{(gates)} \in \mathbb{R}^{4h}$.

Градиенты параметров:

$$\frac{\partial L_t}{\partial W_{dec}} = \delta_t^{(gates)}\, \tilde{x}_t^\top, \quad \frac{\partial L_t}{\partial U_{dec}} = \delta_t^{(gates)}\, s_{t-1}^\top, \quad \frac{\partial L_t}{\partial b_{dec}} = \delta_t^{(gates)}.$$

Градиент по нормализованному входу:

$$\delta_t^{(\tilde{x})} = W_{dec}^\top \delta_t^{(gates)}.$$

2.3 LayerNorm

Обозначим $d = d + 2h$. Пусть $\delta_t^{(\tilde{x})} \in \mathbb{R}^d$.
Стандартно:

$$\delta_t^{(\hat{x})} = \delta_t^{(\tilde{x})} \odot a, \quad \frac{\partial L_t}{\partial a} = \delta_t^{(\tilde{x})} \odot \hat{x}_t, \quad \frac{\partial L_t}{\partial b} = \delta_t^{(\tilde{x})},$$
$$\delta_{t,i}^{(x)} = \frac{1}{\sqrt{\sigma^2 + \epsilon}} \Big[\delta_{t,i}^{(\hat{x})} - \tfrac{1}{d}\sum_j \delta_{t,j}^{(\hat{x})} - \hat{x}_{t,i} \sum_j (\delta_{t,j}^{(\hat{x})} \hat{x}_{t,j})\Big].$$

Разделяем $x_t = [e_t; c_t]$:

$$\delta_t^{(e)} = [\delta_t^{(x)}]_{1:d}, \quad \delta_t^{(c)} = [\delta_t^{(x)}]_{d+1:d+2h}.$$

2.4 Attention (Bahdanau)

Из $\delta_t^{(c)}$ в $\delta_{tj}^{(\alpha)}$:

$$\delta_{tj}^{(\alpha)} = (\delta_t^{(c)})^\top h_j.$$

Через softmax:

$$\delta_{tj}^{(e)} = \sum_k \delta_{tk}^{(\alpha)} (\mathbb{I}_{j=k} - \alpha_{tk}).$$

Далее $r_{tj} = \tanh(u_{tj}), u_{tj} = W_s h_j + W_h s_{t-1} + b_e$:

$$\delta_{tj}^{(u)} = \delta_{tj}^{(e)} \, v \odot (1 - r_{tj}^2),$$

Параметры:

$$\frac{\partial L_t}{\partial W_s} = \sum_j \delta_{tj}^{(u)} h_j^\top, \quad \frac{\partial L_t}{\partial W_h} = \sum_j \delta_{tj}^{(u)} s_{t-1}^\top, \quad \frac{\partial L_t}{\partial b_e} = \sum_j \delta_{tj}^{(u)}, \quad \frac{\partial L_t}{\partial v} = \sum_j \delta_{tj}^{(e)} r_{tj}.$$

---

3. Encoder (BiLSTM) Gradients

Общая градиентная составляющая для h_j :

$$\delta^{(h_j)} = \delta_t^{(c)} \alpha_{tj} + W_s^\top \delta_{tj}^{(u)}.$$

Разделяем на директную и обратную ветки:

$$\delta^{(\overrightarrow{h}_j)} = [\delta^{(h_j)}]_{1:h}, \quad \delta^{(\overleftarrow{h}_j)} = [\delta^{(h_j)}]_{h+1:2h}.$$

3.1 Прямая ветка $(\rightarrow)$

Для каждого $j = 1..N$: Локальные гейты $\delta^{(i,j\rightarrow)}, \delta^{(f,j\rightarrow)}, \delta^{(o,j\rightarrow)}, \delta^{(g,j\rightarrow)}$ (аналогично декодеру), после:

$$\frac{\partial L_t}{\partial W_{enc}^{\rightarrow}} = \sum_j \delta^{(\text{gates},j\rightarrow)} (x_j^{enc})^\top, \quad \frac{\partial L_t}{\partial U_{enc}^{\rightarrow}} = \sum_j \delta^{(\text{gates},j\rightarrow)} (s_{j-1}^{enc,\rightarrow})^\top, \quad \frac{\partial L_t}{\partial b_{enc}^{\rightarrow}} = \sum_j \delta^{(\text{gates},j\rightarrow)}.$$

3.2 Обратная ветка $(\leftarrow)$

Аналогично, но $s_{j+1}^{enc,\leftarrow}$ вместо s_{j-1} , и суммирование по j .

---

4. Accumulation

Все $\partial L_t / \partial \theta$ аккумулируются по $t = 1..T$:

$$\frac{\partial L}{\partial \theta} = \sum_{t=1}^{T} \frac{\partial L_t}{\partial \theta}.$$