

---

# Байесовская дистилляция моделей на базе трансформеров

---

A Preprint

Игорь Н. Игнашин\*

Кафедра интеллектуальных систем  
Национальный исследовательский университет «Московский физико-технический институт»  
Россия, 141701, г. Долгопрудный, Институтский пер., д. 9  
`ignashin.in@phystech.edu`

Elias D. Striatum

Department of Electrical Engineering  
Mount-Sheikh University  
Santa Narimana, Levand  
`stariate@ee.mount-sheikh.edu`

15.12.2023

В данной работе исследовано несколько способов дистилляции моделей на базе трансформеров, а также модели RNN. В статье в первую очередь уделяется внимание дистилляции посредством удаления слоя Attention из моделей. В качестве базовых моделей для дистилляции взяты модель RNN с аддитивным вниманием и модель трансформера seq2seq для задачи перевода текстов. Получены априорные распределения для инициализации дистиллированной модели RNN.

Ожидается лучшая сходимость, лучшее качество у дистиллированной модели по сравнению с моделью той же структуры, но с параметрами инициализации из произвольного нормального распределения при одинаковых условиях обучения.

Keywords Байесовская дистилляция · Трансформеры · RNN

## 1 Введение

Модели трансформеров и RNN часто используются во многих сложных задачах машинного обучения, в том числе, машинного перевода. Однако иногда модели являются переусложненными для своих задач. Потому возникает проблема уменьшения размерности этих моделей или дистилляции моделей. В данной статье поднимается вопрос дистилляции моделей посредством удаления слоев внимания. В статье ? разбирается дистилляция моделей за счет удаления линейных слоев. Подобно тому, как происходит дистилляция модели за счет удаления слоев в многослойной пересети, можно удалять слои внимания в трансформерах, использующих большое количество слоев внимания.

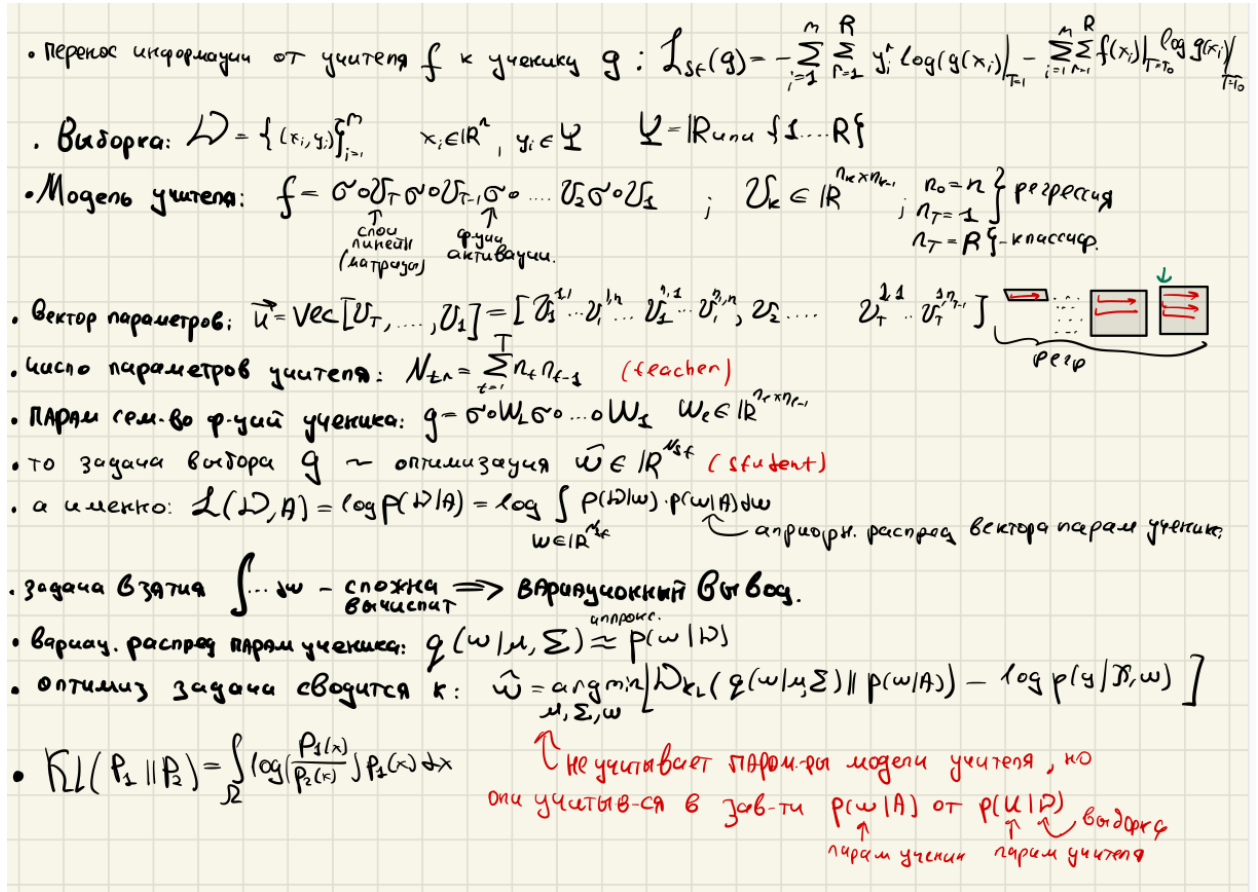


Рис. 1: Sample figure caption.

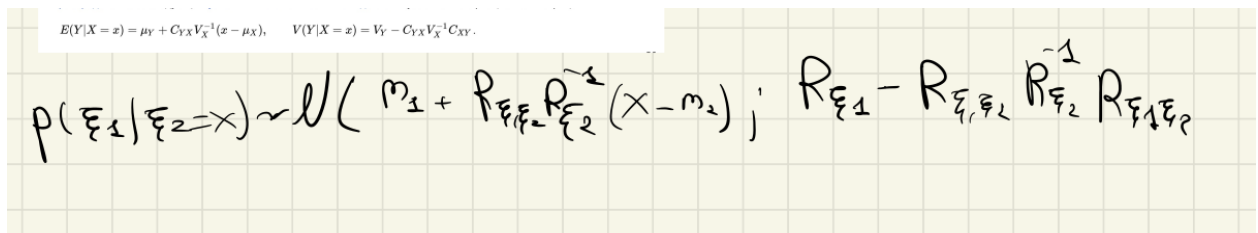


Рис. 2: Sample figure caption.

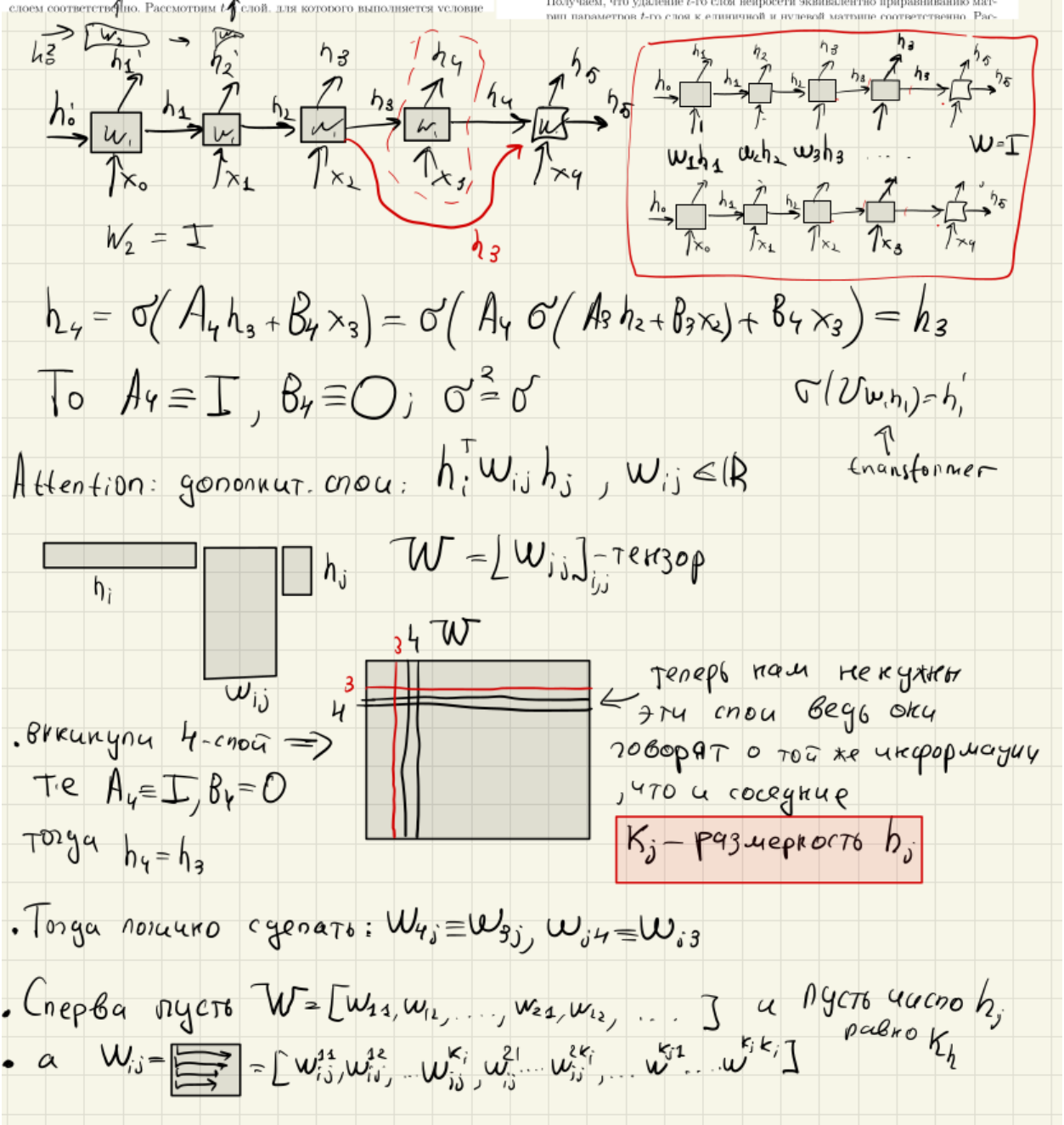
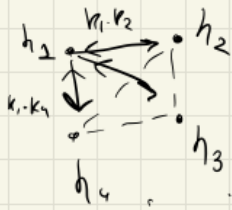


Рис. 3: Sample figure caption.

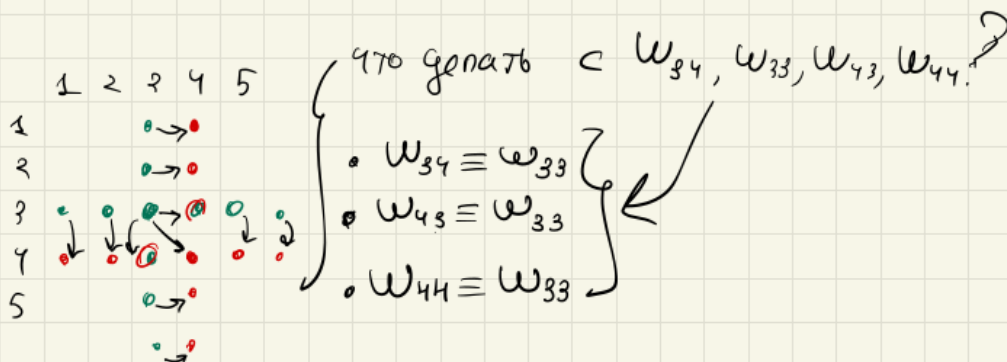
• Пусть  $W \sim N(m, \Sigma)$   $\dim m = (\sum k_i)(\sum k_i) - \sum k_i^2$



Тут наверх лучше оставить  
всю матрицу:  $w_{ij}$  и  $w_{ji}$  - разные  
(т.к. важен порядок  $h_i$  и  $h_j$ )

• Какое распред у  $W' = [w_{11}, w_{12}, w_{13}, w_{13}, \dots, w_{22}, w_{23}, w_{33}, w_{33}, \dots, w_{32}]$

$$W' = [\dots, w_{13}, \dots, w_{23}, \dots, w_{33}, \dots, w]$$



Тогда  $W' = [ \underbrace{w_{13}, w_{14}}_1, \underbrace{w_{23}, w_{24}}_2, \underbrace{w_{33}, w_{33}}_3, \underbrace{w_{31}, w_{32}, w_{33}, w_{33}, w_{45}, \dots}_{4}, \underbrace{w_{53}, w_{54}}_5 ]$

$$W = [V, \xi_1, \xi_2, w_{33}, \eta_1, \eta_2, w_{34}, w_{43}, w_{44}]$$

$$W' = [V, \xi_1, \xi_2, w_{33}, \xi_1, \xi_2, w_{33}, w_{33}, w_{33}]$$

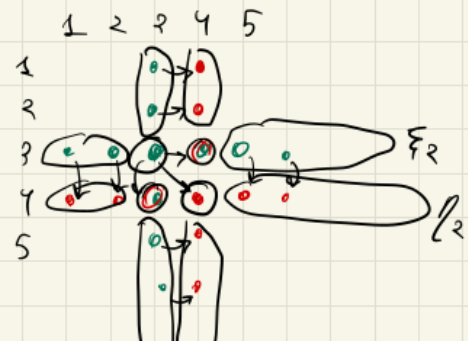


Рис. 4: Sample figure caption.

Lemma 1: Пусть  $\begin{bmatrix} \xi_1 \\ \xi_2 \end{bmatrix} \sim \mathcal{N}(\eta, \Sigma)$ , то  $\begin{bmatrix} \xi_1 \\ \xi_1 \end{bmatrix} \sim \mathcal{N}(\eta', \Sigma')$

□  $\begin{bmatrix} \xi_1 \\ \xi_2 \end{bmatrix} \xrightarrow{\text{avg}} \begin{bmatrix} \xi_1 \\ \xi_1 \end{bmatrix} = \begin{bmatrix} I_1 & 0 \\ I_2 & 0 \end{bmatrix} \begin{bmatrix} \xi_1 \\ \xi_2 \end{bmatrix}$ , то  $\boxed{\eta' = A\eta = \begin{bmatrix} \eta_1 \\ \eta_2 \end{bmatrix}}$

$\Sigma' = A \Sigma A^T = \begin{bmatrix} I_1 & 0 \\ I_2 & 0 \end{bmatrix} \underbrace{\begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}}_{\begin{smallmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{11} & \Sigma_{12} \end{smallmatrix}} \begin{bmatrix} I_1 & I_1 \\ 0 & 0 \end{bmatrix} = \begin{bmatrix} \Sigma_{11} & \Sigma_{11} \\ \Sigma_{11} & \Sigma_{11} \end{bmatrix}$

Th 2:  $W = [v, \xi_1, \xi_2, w_{33}, \eta_1, \eta_2, w_{34}, w_{43}, w_{44}] \sim \mathcal{N}(\eta, \Sigma)$

То же  $W' = [v, \xi_1, \xi_2, w_{33}, \xi_1, \xi_2, w_{33}, w_{33}, w_{33}] \sim \mathcal{N}(\eta', \Sigma')$

□

$A =$

	$I_1$							
		$I_2$						
			$I_3$					
	$I_1$							
		$I_2$						
			$I_3$					
			$I_3$					
			$I_3$					

$AW = [v, \xi_1, \xi_2, w_{33}, \xi_1, \xi_2, w_{33}, w_{33}, w_{33}]$

$\eta' = A\eta = [\eta_v, \eta_1, \eta_2, \eta_3, \eta_1, \eta_2, \eta_3, \eta_3, \eta_3]$

$\Sigma' = A \Sigma A^T$

$I_1$								
$I_1$								
	$I_2$							
		$I_3$						
	$I_1$							
		$I_2$						
			$I_3$					
			$I_3$					
			$I_3$					


$\Sigma$

$I_1$								
$I_1$								
	$I_2$							
		$I_3$						
			$I_4$					
				$I_5$				
					$I_6$			
						$I_7$		
							$I_8$	

$\Sigma_{vv} \Sigma_{11} \Sigma_{22} \Sigma_{33}$

Рис. 5: Sample figure caption.



$$f_{ij} = v_{\tanh}^T \{ W[h_j] \} = f_{ij}(h_j)$$

$$d_{ij} = \text{softmax} \{ f_{ij} \} = \text{const}$$

$$c_i = \sum d_{ij} h_j = \sum \text{softmax} \{ f_{ij} \} h_j = \sum \text{softmax} \{ v^T \sigma(w_1 h_i + w_2 h_j) \} h_j$$

$$\sum_j d_j h_j = \sum \frac{1}{T} h_j$$

$$\sum_j \text{softmax} \{ v^T \sigma(w_1 h_i + w_2 h_j) \} h_j = h_i$$

$$v^T \sigma \{ (w_1 + w_2) h_i \}$$

Рис. 8: Sample figure caption.

## 2 Постановка задачи

## 3 Дистилляция

### 3.1 Трансформер

### 3.2 RNN

## Список литературы

1) Пока 1 слой трансформера  $\rightarrow$  удаляем

2) без 2-го слоя  $RWU$

$$\begin{aligned} \boxed{\begin{aligned} h_j &= \sigma(A h_{j-1} + B x_{j-1}) \\ h'_j &= \sigma(w_j h_j) \end{aligned}} \Rightarrow \boxed{\begin{aligned} h_j &= \sigma(A h_{j-1} + B x_{j-1}) \\ h'_j &= \sigma(h_j) = h_j \\ \text{т.е. } W_j &= I \end{aligned}}$$

Индексация:  $W = [U, V]$   
 $\uparrow$  слой  $RWU$   $\uparrow$  слой трансформера

$$\begin{aligned} &\begin{matrix} \xrightarrow{\quad} & \xrightarrow{\quad} & \xrightarrow{\quad} \\ \xleftarrow{\quad} & \xleftarrow{\quad} & \xleftarrow{\quad} \end{matrix} \\ &A \in \mathbb{R}^{n \times n} \quad B \in \mathbb{R}^{n \times n} \end{aligned}$$

Предположение  $W \sim \mathcal{N}(\mu, \Sigma)$

Генерация:  $W = [U, i]$ ,  $i = \left[ \underbrace{1 \dots 0 \dots 0 \dots 0 \dots 1}_{m \times n}, \underbrace{-1 \dots -1}_{2}, \underbrace{-1 \dots -1}_{3}, \dots, m \right]$

$$p(\xi_1 | \xi_2 = x) \sim \mathcal{N} \left( m_1 + R_{\xi_1 \xi_2} R_{\xi_2}^{-1} (x - m_2); R_{\xi_1} - R_{\xi_1 \xi_2} R_{\xi_2}^{-1} R_{\xi_2 \xi_1} \right)$$

$$\text{тогда: } W \sim \mathcal{N} \left\{ \begin{aligned} &m_1 + R_{UV} R_V^{-1} (i - m_V) \\ &R_U - R_{UV} R_V^{-1} R_{VU} \end{aligned} \right\}$$

Рис. 9: Sample figure caption.



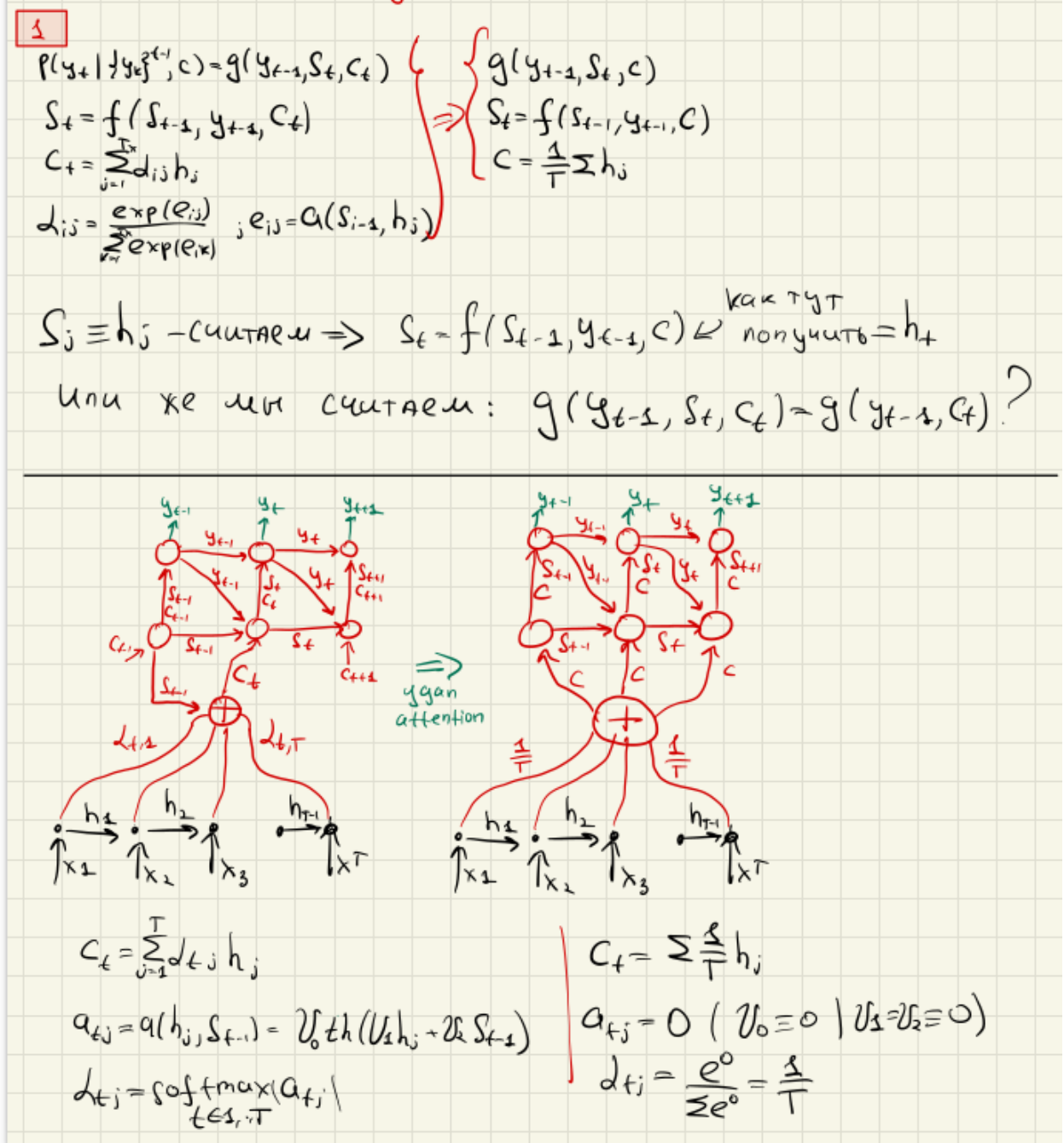


Рис. 10: Sample figure caption.

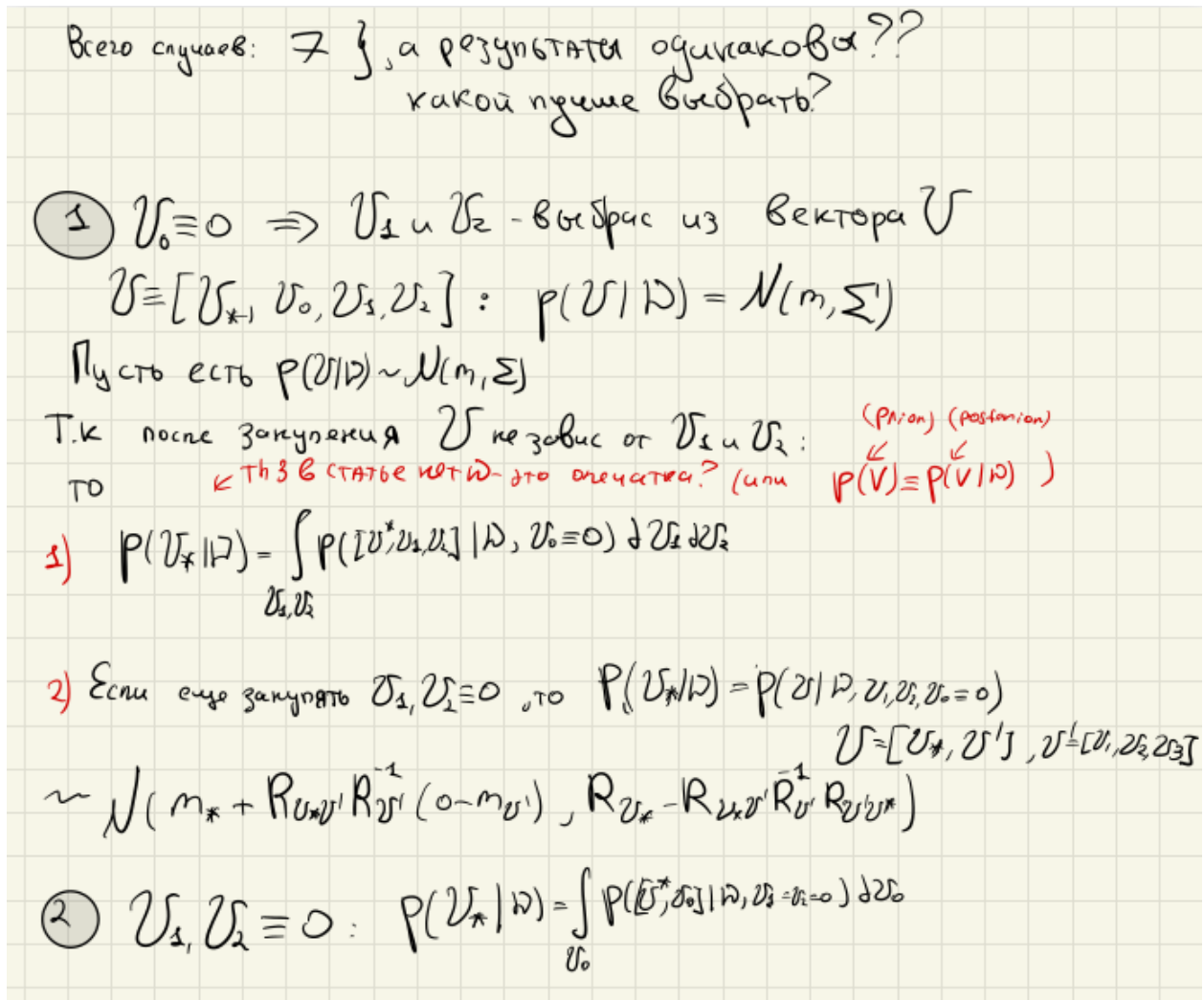


Рис. 11: Sample figure caption.