Обучение рекуррентных нейронных сетей с использованием сопряженных уравнений

Студент: Хоменко Р.Д.
Руководитель: Ревизников Д.Л., профессор-совм. кафедры 810Б, д.ф.-м.н., профессор

Постановка задачи

Исследовать возможность применения Neural ODEs в задачах обработки естественного языка. Сравнить данный подход с классическими архитектурами на примере модели генерации текста.

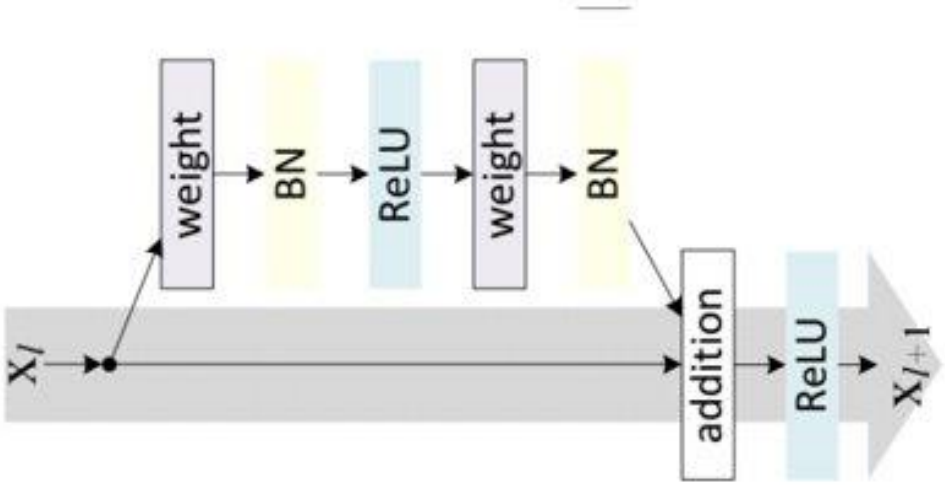
Обзор исследований

- Ricky T. Q. Chen, Yulia Rubanova, Jesse Bettencourt, David Duvenaud. **Neural Ordinary Differential Equations**
- Hanshu Yan, Jiawei Du, Vincent Y. F. Tan, Jiashi Feng. **On Robustness of Neural Ordinary Differential Equations**
- Edward De Brouwer, Jaak Simm, Adam Arany, Yves Moreau. **GRU-ODE-Bayes: Continuous modeling of sporadically-observed time series**
- Prasad Kawthekar, Raunaq Rewari, Suvrat Bhooshan. **Evaluating Generative Models for Text Generation**

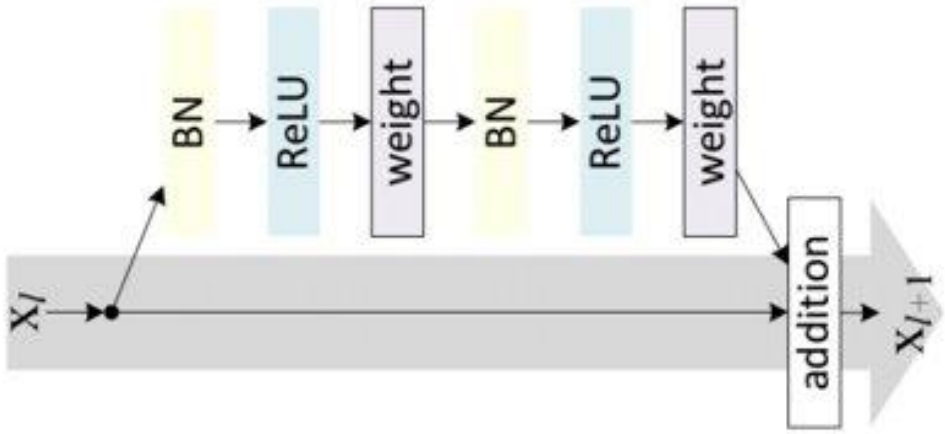
Актуальность темы исследования

- Ускорение обучения на больших объемах данных
- Контроль над точностью и вычислительными ресурсами
- Исследование комбинации Neural ODE с другими слоями
- Исследование модели (устойчивость решения и т.д.)

Архитектура ResNet

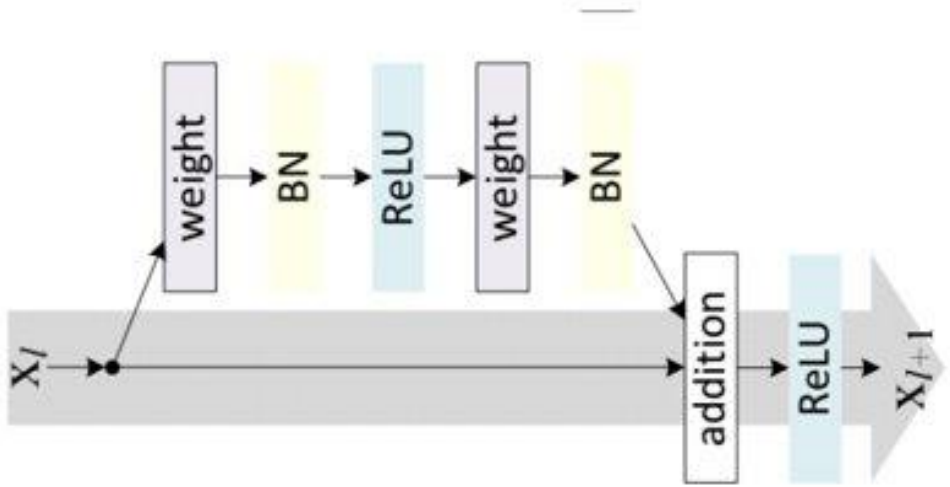


(а) ResNet блок

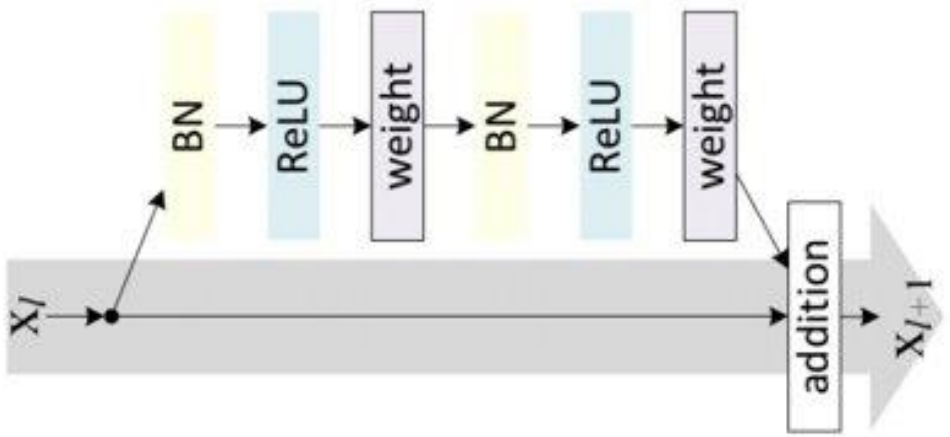


(б) ResNet блок с пред-активацией

Архитектура ResNet



(а) ResNet блок



(б) ResNet блок с пред-активацией

$$\begin{cases} \dfrac{\mathrm{d}x}{\mathrm{d}t} = f(x,t), \\ x(0) = x_0 \end{cases}$$

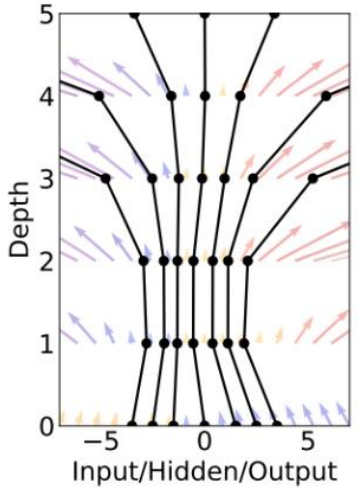
$$x_{\ell+1} = x_\ell + hf(x_\ell, t_\ell)$$

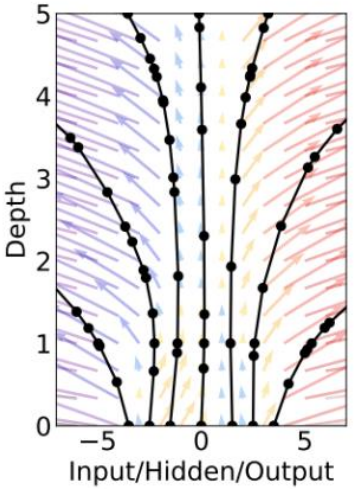
Neural ODEs

$$\begin{cases} \dfrac{\mathrm{d}z}{\mathrm{d}t} = f\left(z\left(t\right), \boldsymbol{\theta}\right), \quad t \in [t_0, t_1] \\ z(t_0) = z_0 \end{cases}$$

$$L(z\left(t_1\right)) \rightarrow \min_{\boldsymbol{\theta}}$$

- z_0 - вектор признаков
- $z(t_1)$ - выход сети
- θ – вектор параметров модели
- L – функция потерь
- $z(t) = z(t_0) + \int_{t0}^{t} f(z(t), t, \theta)\, dt$





Ajoint метод: условия оптимальности (1)

$L(z(t_1), \theta)$ $\quad \frac{\mathrm{d}z}{\mathrm{d}t} = f(z(t), \theta),\ z(t_0) = z_0$

$$\mathcal{L}(z(t), \theta, a(t)) = L(z(t_1), \theta) + \int_{t_0}^{t_1} a(t)^\top \left(\frac{\mathrm{d}z}{\mathrm{d}t} - f(z(t), \theta) \right) \mathrm{d}t + b \cdot (z(t_0) - z_0)$$

θ:

$$\delta_\theta \mathcal{L}(\theta, h) = \frac{\partial L}{\partial \theta}(z(t_1), \theta) \cdot h - \int_{t_0}^{t_1} a(t)^\top \frac{\partial f}{\partial \theta}(z(t), t, \theta)\, \mathrm{d}t \cdot h = 0 \quad \text{for any feasible variation } h$$

$$\begin{cases} \frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{\partial L}{\partial \theta}\right) = -a(t)^\top \frac{\partial f}{\partial \theta}(z(t), \theta) \\ \left.\frac{\partial L}{\partial \theta}\right|_{t=t_1} = \mathbf{0} \end{cases}$$

$\delta J(y, h) \triangleq \lim_{\varepsilon \to 0} \frac{J[y+\varepsilon h] - J[y]}{\varepsilon} = \frac{\partial}{\partial \varepsilon} J[y + \varepsilon h]\big|_{\varepsilon=0}$

Ajoint метод: условия оптимальности (2)

$$\mathcal{L}(z(t), \boldsymbol{\theta}, \boldsymbol{a}(t)) = L(z(t_1), \boldsymbol{\theta}) + \int_{t_0}^{t_1} \boldsymbol{a}(t)^\top \left(\frac{\mathrm{d}z}{\mathrm{d}t} - f(z(t), \boldsymbol{\theta}) \right) \mathrm{d}t + b \cdot (z(t_0) - z_0)$$

z:

$$\delta_z \mathcal{L}(z, \boldsymbol{h}) = \frac{\partial L}{\partial z}(z(t_1), \boldsymbol{\theta}) \cdot \boldsymbol{h}(t_1) + \lim_{\varepsilon \to 0} \frac{1}{\varepsilon} \int_{t_0}^{t_1} \boldsymbol{a}(t)^\top \left(\varepsilon \frac{\mathrm{d}\boldsymbol{h}(t)}{\mathrm{d}t} - \frac{\partial f}{\partial z}(z(t), \boldsymbol{\theta}) \cdot \varepsilon \boldsymbol{h}(t) + o(\varepsilon) \right) \mathrm{d}t$$

$$\int_{t_0}^{t_1} \boldsymbol{a}(t)^\top \frac{\mathrm{d}\boldsymbol{h}(t)}{\mathrm{d}t} \mathrm{d}t = \left[\boldsymbol{a}(t)^\top \boldsymbol{h}(t) \right] \Big|_{t_0}^{t_1} - \int_{t_0}^{t_1} \frac{\mathrm{d}\boldsymbol{a}(t)}{\mathrm{d}t} \boldsymbol{h}(t) \, \mathrm{d}t = -\boldsymbol{a}(t_1)^\top \boldsymbol{h}(t_1) - \int_{t_0}^{t_1} \frac{\mathrm{d}\boldsymbol{a}(t)}{\mathrm{d}t} \boldsymbol{h}(t) \, \mathrm{d}t$$

$$\delta_z \mathcal{L}(z, \boldsymbol{h}) = \left(\frac{\partial L}{\partial z}(z(t_1)) - \boldsymbol{a}(t_1) \right) \cdot \boldsymbol{h}(t_1) - \int_{t_0}^{t_1} \left\{ \frac{\mathrm{d}\boldsymbol{a}(t)}{\mathrm{d}t} + \boldsymbol{a}(t)^\top \frac{\partial f}{\partial z}(z(t), \boldsymbol{\theta}) \right\} \boldsymbol{h}(t) \mathrm{d}t = 0$$

$$\frac{\mathrm{d}\boldsymbol{a}(t)}{\mathrm{d}t} = -\boldsymbol{a}(t)^\top \frac{\partial f}{\partial z}(z(t), \boldsymbol{\theta}), \quad \boldsymbol{a}(t_1) = \frac{\partial L}{\partial z}(z(t_1))$$

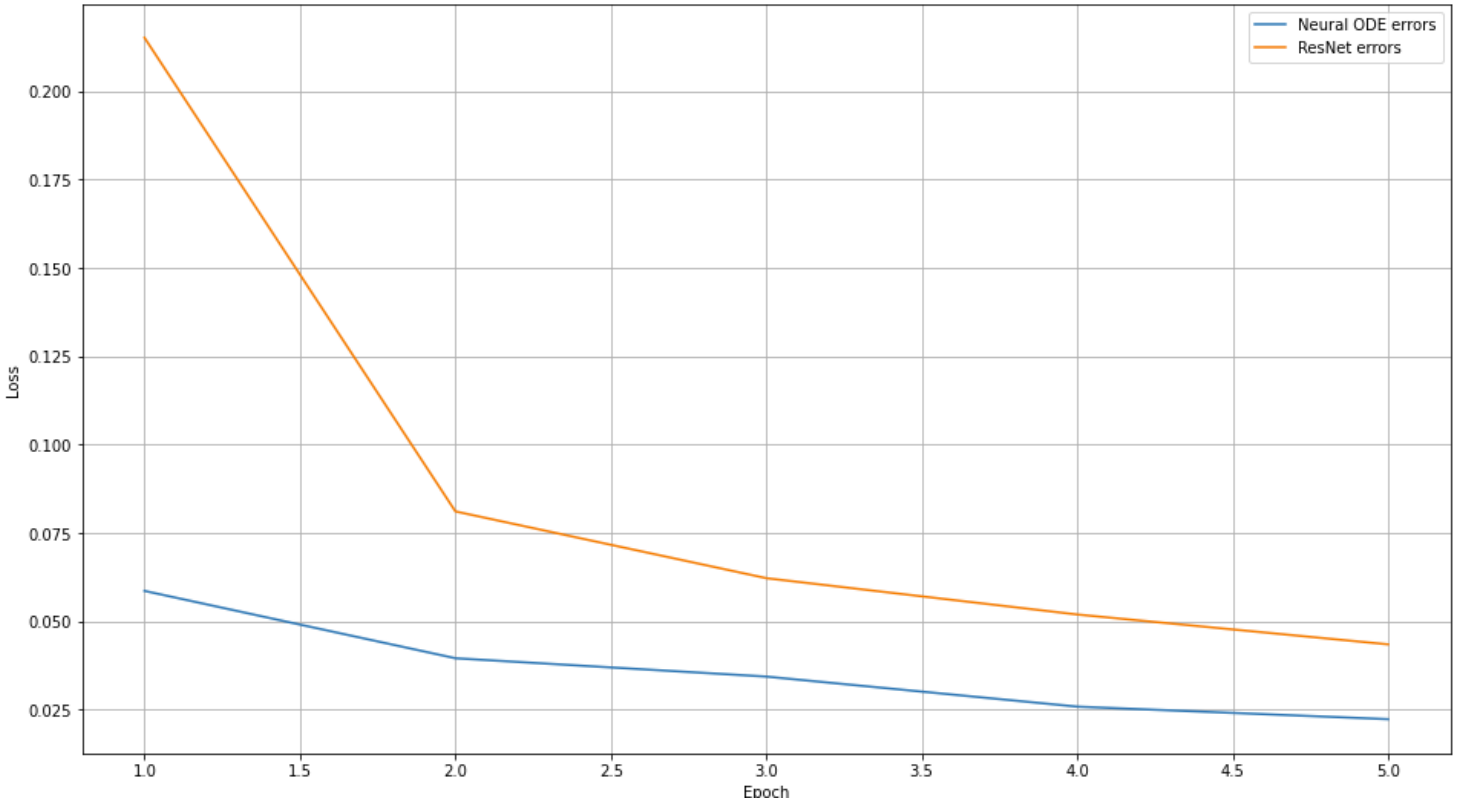
Ajoint метод

$$\begin{cases} \dfrac{\mathrm{d}}{\mathrm{d}t}\dfrac{\partial L}{\partial \boldsymbol{\theta}} = -\boldsymbol{a}(t)^{\top}\dfrac{\partial f}{\partial \boldsymbol{\theta}}(z(t), \boldsymbol{\theta}) \\ \dfrac{\partial L}{\partial \boldsymbol{\theta}}\Big|_{t=t_1} = \mathbf{0} \end{cases} \qquad \boxed{\begin{cases} \dfrac{\mathrm{d}z}{\mathrm{d}t} = f(z(t), \boldsymbol{\theta}) \\ z(t_1) = z_1 \end{cases}} \qquad \begin{cases} \dfrac{\mathrm{d}\boldsymbol{a}(t)}{\mathrm{d}t} = -\boldsymbol{a}(t)^{\top}\dfrac{\partial f}{\partial z}(z(t), \boldsymbol{\theta}) \\ \boldsymbol{a}(t_1) = \dfrac{\partial L}{\partial z}(z(t_1)) \end{cases}$$

Прямой проход

Обратный проход

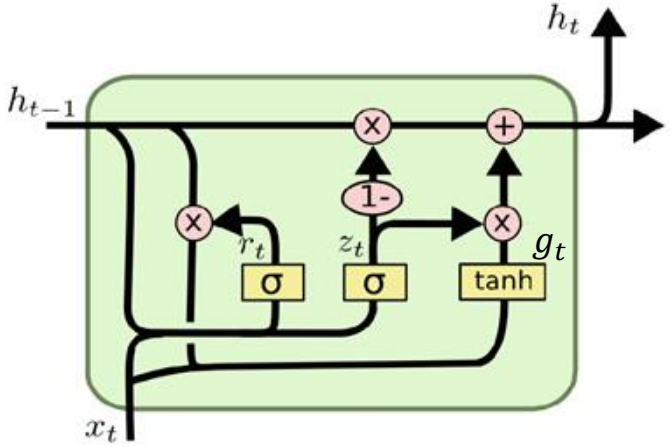
Ajoint метод: сравнение на архитектуре ResNet



GRU-ODE

GRU

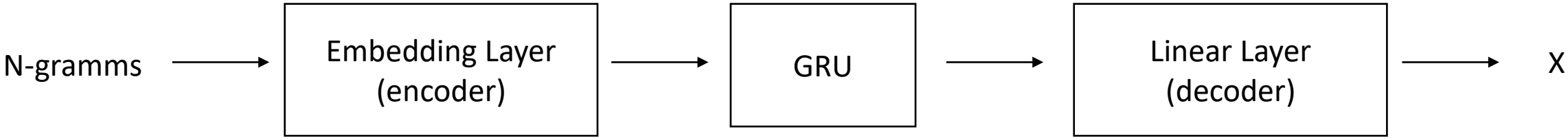
$$\mathbf{h}_t = \mathbf{z}_t \odot \mathbf{h}_{t-1} + (1 - \mathbf{z}_t) \odot \mathbf{g}_t$$
$$\mathbf{r}_t = \sigma(W_r \mathbf{x}_t + U_r \mathbf{h}_{t-1} + \mathbf{b}_r)$$
$$\mathbf{z}_t = \sigma(W_z \mathbf{x}_t + U_z \mathbf{h}_{t-1} + \mathbf{b}_z)$$
$$\mathbf{g}_t = \tanh(W_h \mathbf{x}_t + U_h(\mathbf{r}_t \odot \mathbf{h}_{t-1}) + \mathbf{b}_h)$$



GRU-ODE

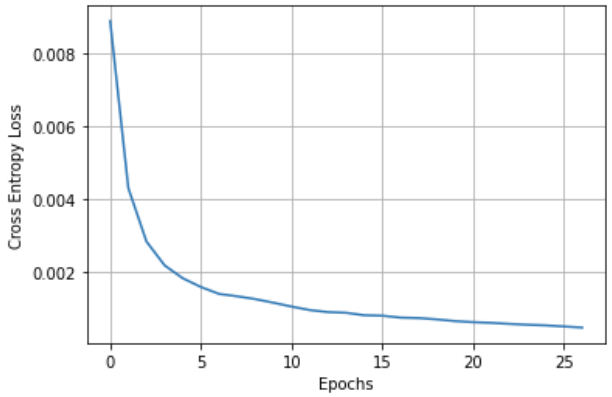
$$\frac{d\mathbf{h}(t)}{dt} = (1 - \mathbf{z}(t)) \odot (\mathbf{g}(t) - \mathbf{h}(t))$$
$$\mathbf{r}(t) = \sigma(W_r \mathbf{x}(t) + U_r \mathbf{h}(t) + \mathbf{b}_r)$$
$$\mathbf{z}(t) = \sigma(W_z \mathbf{x}(t) + U_z \mathbf{h}(t) + \mathbf{b}_z)$$
$$\mathbf{g}(t) = \tanh(W_h \mathbf{x}(t) + U_h(\mathbf{r}(t) \odot \mathbf{h}(t)) + \mathbf{b}_h)$$

Модель генерации текста

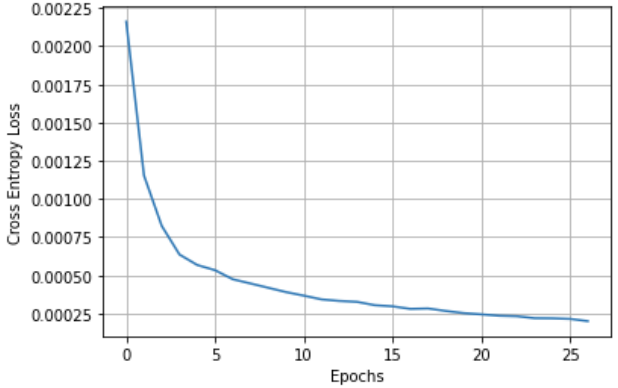


$$i_{word} = \text{multinomial}(\exp\left(\frac{X}{T}\right), 1)$$

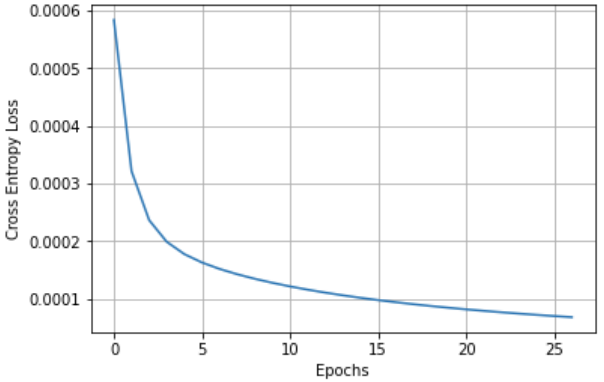
Обучение модели



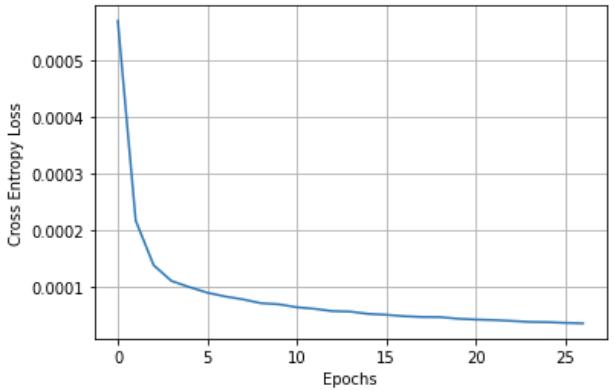
vanilla RNN



GRU

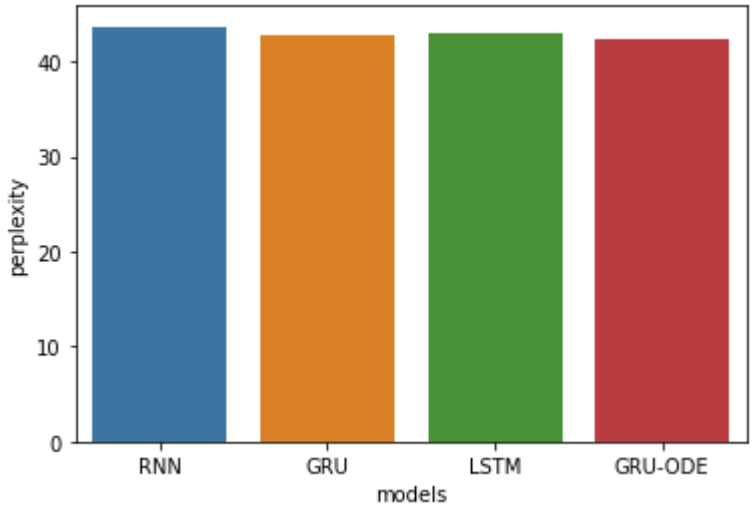


LSTM

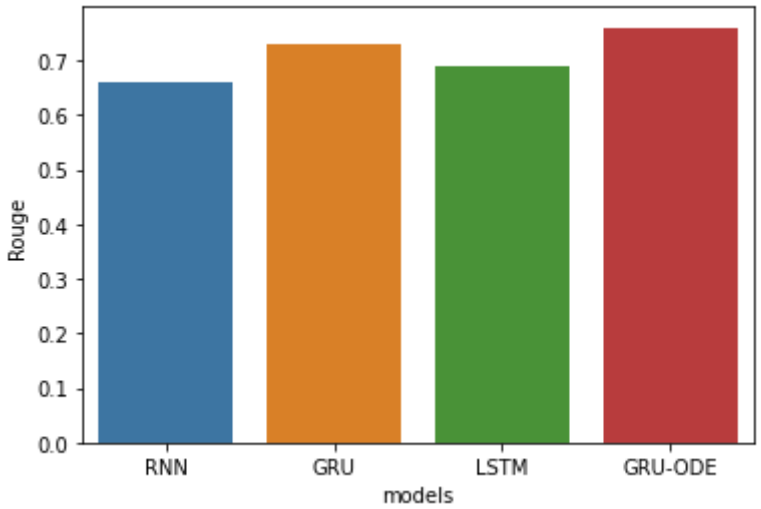


GRU-ODE

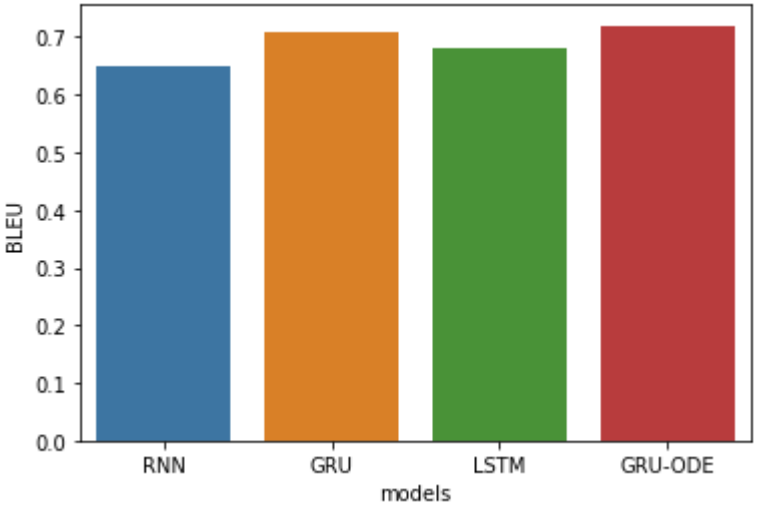
Метрики качества



Perplexity



Rouge



BLEU

Проблемы

- «жесткость» систем при backward pass
- некорректные результаты при использовании неявных методов
- оценка качества генеративных моделей

Методы решения

- применение специфических методов для backward pass
- использование методов с адаптивным шагом
- комбинированные метрики

Заключение

- Исследована модель нейронных ОДУ
- Реализован алгоритм обучения GRU-ячейки на основе нейронных ОДУ
- Реализация успешно протестирована на модели генерации текста

Нейронные ОДУ могут быть успешно интегрированы в существующих моделях.