

Controllable text generation with small data using auxiliary in-domain enrichment

Беляев Станислав
Научный руководитель: Брыксин Тимофей

Санкт-Петербургский Академический Университет
stasbelyaev96@gmail.com

23 марта 2018

Введение

Обзор

“If a typical person can do a mental task with less than one second of thought, we can probably automate it using AI either now or in the near future.”

— Andrew Ng, 2017

Машины умеют:

- 1 Различать формы и объекты
- 2 Имитировать стиль изображений
- 3 Отвечать на простые вопросы

Машины НЕ умеют:

- 1 **Хорошо** подражать высшей нервной деятельности
- 2 Понимать и обобщать сложные категории
 - Этика (юмор, мораль, норма, ...)
 - Эстетика (книги, картины, ...)

Введение

Постановка задачи

МООС платформам нужна генерация контента:

- Дешево
- Быстро
- Ультимативная защита от списывания

Особенности:

- Generic характер генерации
- Примеров готового контента мало
- Набор текстовых свойств для единицы контента (курс, тема, тэги, сложность, ...)



Задача: По набору свойств $f = \{f_i \in F\}$ сгенерировать новые примеры текстовых данных из генеральной совокупности X , соответствующих f . Возьмем в качестве X условия задач по программированию.

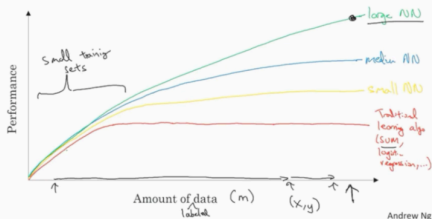
Введение

Данные в DL

“Data is the New Oil.”

— Andrew Ng, 2017

Scale drives deep learning progress



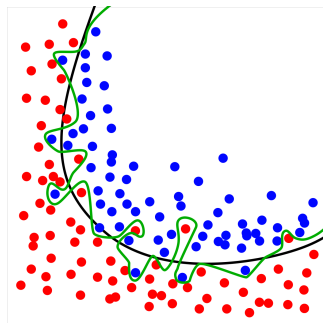
Введение

Проблема данных

Если мы не знаем паттерна для генерации и хотим уметь обобщать, то будем использовать DL и больше данных (Mikolov et al., 2010).

Но что делать, если данных мало?

- Мы не сможем обобщать
- Мы скорее всего переобучимся
- При генерации новые сэмплы будут слишком похожи на старые



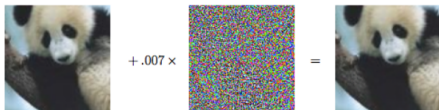
Решение: Искать похожие $X_{\text{aux}} \sim X$ in-domain данные из смежных областей.

Введение

Изображения vs текст

Изображения

$$f: \mathbb{R}^2 \rightarrow \mathbb{R}^M$$



- Непрерывное пространство
- Набор всевозможных преобразований как дифференцируемых функций
- Понятно, куда распространять градиент

Текст

...an efficient method for learning high quality distributed vector ...

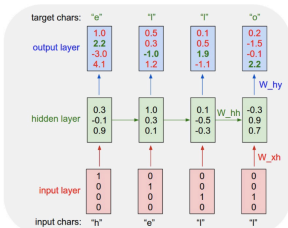
context focus word context

- Дискретное пространство
- Переменная длина
- Нет устойчивости к шуму
- Long-term зависимости
- Омонимия и контекст

Введение

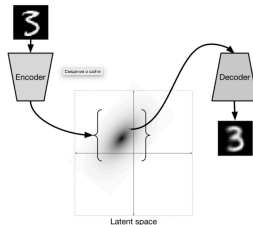
Генерация текста

RNN



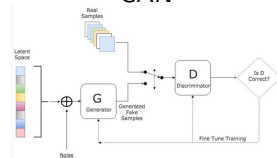
(Mikolov et al., 2011)

VAE



(Bowman et al., 2016; Hu et al., 2018)

GAN



(Yu et al., 2017; Fedus et al., 2018)

Данные

Источники

Условия задачек



- $|X_1 \in X| = 5k$
- Тэги (f) уже проставлены
- Собран вручную, но будет готовый

Stackoverflow



- Берем вопросы с тэгом *python*
- $|X_2 \in X_{aux}| = 600k$
- Тэги уже проставлены
- Предобработка

Docstring

```
def test_func(test, **kwargs):
    """Write a DataFrame to a Google BigQuery table.

    If the table exists, the DataFrame will be appended. If not, a new table
    will be created, in which case the schema will have to be specified. By
    default, rows will be written in the order they appear in the DataFrame,
    though the user may specify an alternative order.

    """
    print "test"
```

```
Docstring:
S.lower() -> string

Return a copy of the string S converted to lowercase.
Type: builtin_function_or_method
```

- $|X_3 \in X_{aux}| = 150k$
- Тэги = Entities
- Предобработка

Данные

Stackoverflow

TODO: Написать

Данные

Docstring

TODO: Написать

Данные

Итого

TODO: Написать

RNN

Обзор и применение

TODO: Написать

VAE

Обзор и применение

TODO: Написать

GAN

Обзор и применение

TODO: Написать...

Оценивание

Метрики

Как можно оценить результат генерации?

- Perplexity
- Assessors evaluation
 - MTurk
 - Я.Толока
- Самому

Оценивание

Perplexity

$X_{\text{train}}, X_{\text{test}}$ - разбили датасет $X_{\text{data}} \subset X \cup X_{\text{aux}}$.

Есть языковая модель M , обученная на X_{train} . Как оценить эффективность?

Посчитаем вероятность предложений $W \in X_{\text{test}}$.

Perplexity

$$PP(W) = P(w_1 w_2 w_3 \dots w_{|W|})^{-\frac{1}{|W|}}$$

Chain rule

$$PP(W) = \left[\prod_{i=1}^{|W|} \frac{1}{P(w_i | w_1 \dots w_{i-1})} \right]^{\frac{1}{|W|}}$$

- Нижний терм в произведении \Leftrightarrow очередной шаг алгоритма
- Чем меньше perplexity, тем больше $P(W)$, т.е. тем лучше.
- Отдельно посчитаем для $X_{\text{test}} \cap X$ (это - реальная метрика)

Выводы

Результаты

TODO: Это получилось, а это не получилось





Выводы

Будущая работа

TODO: Дальше я буду делать то то и то то

Ссылки

Статьи, код и контакты

- 1  [Antonio Valerio Miceli Barone \(2017\)](#)
A parallel corpus of Python functions and documentation strings for automated code documentation and code generation
- 2 [Karpathy, Andrej \(2015\). "The Unreasonable Effectiveness of Recurrent Neural Networks".](#)
- 3  [Samuel R. Bowman \(2016\)](#)
Generating Sentences from a Continuous Space
- 4  [Zhiting Hu \(2018\)](#)
Toward Controlled Generation of Text
- 5  [Heng Wang \(2017\)](#)
Text Generation Based on Generative Adversarial Nets with Latent Variable
- 6 <https://github.com/stasbel/task-gen> (Генерация)
- 7 <https://github.com/stasbel/bachelor-thesis> (Презентация)
- 8 <https://t.me/stasbel>