

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.

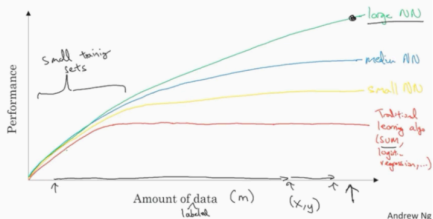
Введение

Данные в DL

“Data is the New Oil.”

— Andrew Ng, 2017

Scale drives deep learning progress



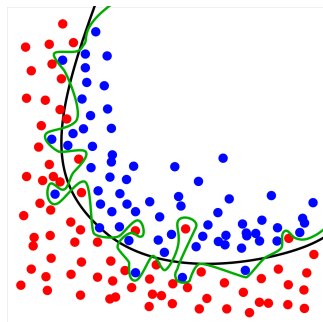
Введение

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.

Данные

Условия задач

- Разный контекст, форма, требования и сложность
 - Сложно выделить шаблон
- Собран вручную (~ 100)
- Запрос в Stepik
 - На английском
 - С тегами и темами
- Расширяемо за счет кравлеров по codeforces и hackerank



Given two integers n and m , not exceeding 100. Fill in a matrix of size $n \times m$ chequer-wise: the cells of one color should be filled with zeros, and of another color - with positive natural numbers top to bottom, left to right. Number 1 should be written in the top left corner.

Output data format

Output the resulting matrix, each element should take exactly 4 characters (including spaces).

Данные

Stackoverflow

- Есть готовый за 2008-2016
- 2016-2018 получаем с открытого api
- Хотя бы с одним тэгом *python*
 - Остальные тэги уже проставлены
- Замена нод с кодом на *CODE*
 - Не портим словарь
 - Не ищем/учитываем лишние ненужные зависимости
- X_{aux} , но слова и выражения используются в правильном значении



```
'I am in Python and I am using EasyGUI. I want to know how to keep a easygui.buttonbox window open after you click a button. DCNL Here is my code: DCNL CODE DCNL I would appreciate it if you would answer if you know how to do this. DC NL Thanks!')
```

Данные

Docstring

- Кравлер по гитхабу, *python* код
- ASCII text без вставок кода + ограничения по длине
- Похожий на английский
 - Готовая языковая модель
- Тэги
 - Entities
 - Связки сущ. + прилаг.
 - Top1000
- X_{aux} , но слова и выражения используются в правильном значении

Пример

Docstring:
S.lower() -> string

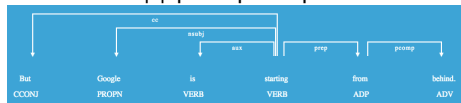
Return a copy of the string S converted to lowercase.
Type: builtin_function_or_method

Entities

Apple vs Google

But **Google** **ORG** is starting from behind. The company made a late push into hardware, and **Apple** **ORG** 's Siri, available on **iPhones** **PRODUCT**, and **Amazon** **ORG** 's **Alexa** **ORG** software, which runs on its **Echo** **APP** and **Dot** **ORG** devices, have clear leads in consumer adoption.

Дерево разбора



Тэги

```
str(entities.most_common(50))
```

```
"[('test', 16568), ('list', 7555), ('object', 7078), ('file', 6770), ('value', 6252),
```

```
'Return the address which this transport is pretending to be bound DCNL to.',
'common code for quickly building an ansible module in Python DCNL (although you can write modules in anything that
can return JSON) DCNL see library/* for examples',
'Get the stylesheet from the visitor. DCNL Ask the visitor to setup the page.']
```


Данные

Итого

$$X_{\text{data}} = X_1 \cup X_2 \cup X_3$$

Условия задачек



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

Stackoverflow



- Берем вопросы с тэгом *python*
- $|X_2 \in X_{\text{aux}}| = 600k$
- 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
```

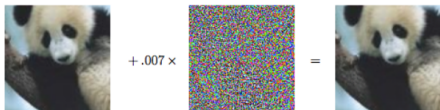
- $|X_3 \in X_{\text{aux}}| = 150k$
- Тэги = Entities
- Предобработка

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 ...

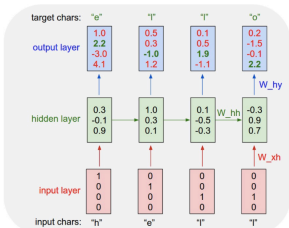
context focus word context

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

Введение

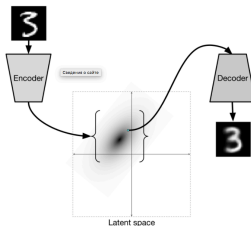
Генерация текста

RNN



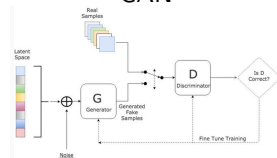
(Mikolov et al., 2011)

VAE



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

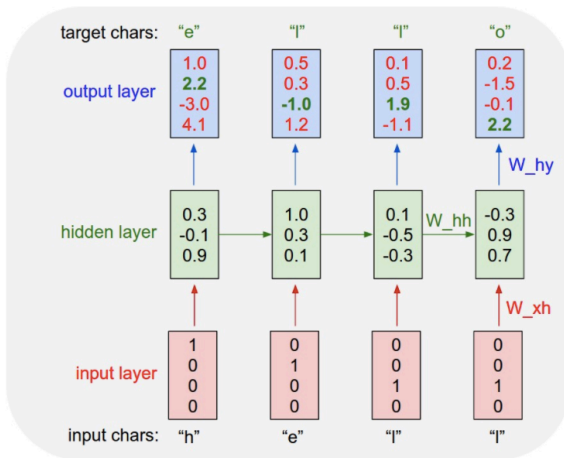
GAN



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

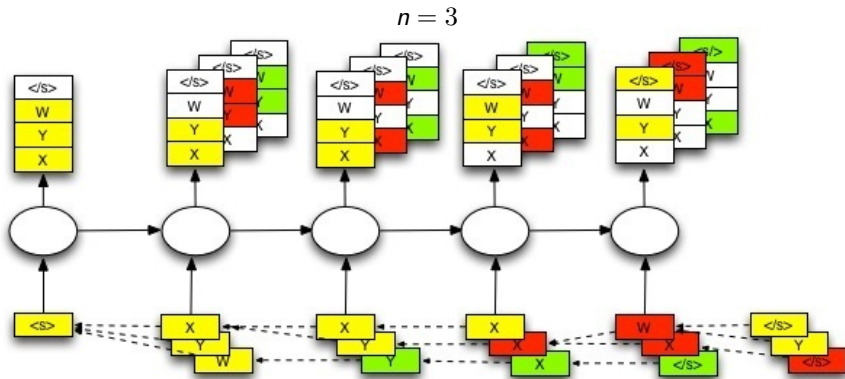
RNN

Обзор



RNN

Beam search



RNN

Seed

Как задать *RNN* начальные условия для генерации?

- Out-of-band (Chen et al., 2015; Lipton et al., 2015):
 - Конкатенация с векторным представлением начальных условий
 - На каждом шаге/один раз в начале
 - One-hot/LDA topic modeling/doc2vec
- In-band
 - Префикс
 - Префикс + суффикс

$\text{text} \Rightarrow \text{tags} + | + \text{text} + | + \text{tags}$

- Начинаем генерировать с нужного префикса
- Отрезаем суффикс

RNN

Реализация

Модель:

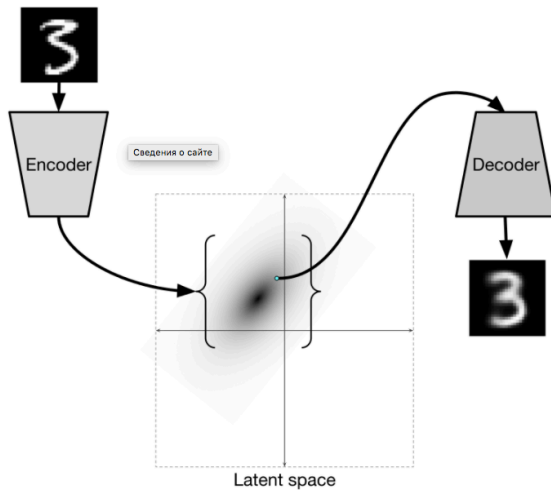
- MultiLayerLSTM, 2 слоя
- WordRNN/CharRNN/BPERNN
- Dropout=0.5 на первом слое
- 100 эпох
- Out-of-band/In-band

Результаты:

- Долго сходится и плохо интерпретируется
- Правдоподобная генерация, но в качестве Seed не получится передать близость к X
 - Out-of-band учится отделять X от X_{aux}
 - In-band путается

VAE

Обзор и применение



GAN

Обзор и применение

TODO: Написать

Оценивание

Метрики

Как можно оценить результат генерации? (Salimans et al., 16)

- Perplexity
- Assessors evaluation
 - MTurk, Я.Толока
 - DCG, MAP
- Самому
 - Generic-генерация
 - Генерация по заданным темам

Оценивание

Определение 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 - реально важная метрика)

Оценивание

Таблица perplexity

Test	RNN	VAE	CVAE	GAN
PTB	38.93	NaN	NaN	39.12
CMC	29.10	NaN	NaN	29.09
X_{test}	30.29	NaN	NaN	NaN
$X_{\text{test}} \cap X$	40.10	NaN	NaN	NaN

Таблица: Perplexity

Оценивание

Примеры

RNN (20 эпох)

```
generate_text(60, seed=['user', 'server'], beam=5) # prefix = 'user, server | '  
'Takes a user and service the service connection to server to'
```

Выводы

Результаты

- Анализ state-of-the-art методов генерации текста
 - Модификации для наших данных
 - Сравнение подходов
- Анализ влияние данных на генерацию
 - Каково влияние X_{aux} на генерацию?
 - Как соотносятся X и X_{aux} в терминах латентных представлений?
- Метрики и эмпирические проверки, позволяющие оценить сложность задачи



Выводы

Будущая работа

- Попытаться проинтерпретировать важной свойств
 - Seed для *RNN*
 - Латентное подпространство для $X_{\text{test}} \cap X$ из *VAE*
- Больше данных \Rightarrow выделить паттерн для генерации?
- Оптимизация скорости для тестирования:
 - *RNN* \Rightarrow *CNN*
- Попробовать GAN'ы
 - WC-GAN
 - SeqGAN
 - GumbelSoftmax
- Генерировать код решения по условию задачи

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>