

Санкт-Петербургский Государственный Университет

Курсовая работа на тему
“Методика автоматической генерации названий картин на основе глубокого
обучения”

Студент 331 группы:
Алтынова Анна Юрьевна

Научный руководитель:
к.ф.-м.н, доцент кафедры информатики Григорьев Д.А.

Зачтено!

→ Григорьев Д.А.

26 декабря 2020 г.

1 Содержание

1. Введение
2. Постановка задачи
3. Методы
 - (a) Датасет
 - (b) Обучение с переносом
 - (c) Модель NIC
 - Описание
 - Шифровщик
 - Дешифровщик
 - (d) Внимание
 - (e) BLEU метрика
4. Результаты
5. Ограничения подхода
6. Дальнейшая работа
7. Список литературы

2 Введение

В настоящее время ощущается нехватка методов машинного обучения для ускоренного и одновременно глубокого изучения сложных аспектов искусствоведения среди участников арт-рынка, не имеющих соответствующего образования и компетенций. Прежде всего, трудности вызывают неоднозначные названия произведений искусства, а также разъясняющие (на практике же нередко усложняющие понимание и восприятие) подписи самих авторов к ним. Таким образом, задача машинного накопления знаний в области искусства тесно связана с задачей открытия доступа к арт-рынку экономистам, инвесторам, бизнес-аналитикам, ученым-исследователям.

3 Постановка задачи

Целью работы является разработка программного обеспечения, способного, получая на вход изображение картины, сгенерировать для неё подходящее название, максимально отражающее ее содержание и смысловую ценность.

Для достижения поставленной задачи и оценки полученных результатов были выбраны следующие методы:

1. Датасет WikiArt
2. Нейронная сеть с архитектурой шифрования-дешифрования NIC
3. Нейронная сеть с механизмом внимания
4. Метрика BLEU

Работа выполнена на языке Python 3.6 в среде Google Colab с использованием библиотеки Keras

4 Методы

4.1 Датасет

WikiArt датасет состоит из более чем 80.000 изображений изобразительного искусства, датированных от 15 века до настоящего времени, содержит 27 стилей и 45 различных жанров. Он составлен на основе содержания сайта [wikiart\[1\]](https://www.wikiart.org/) и на текущий момент это самый большой датасет, состоящий из картин. Тем не менее, для тренировки модели в данной работе было использовано только множество из 10тыс. картин в стиле реализм, так как тренировочный процесс требует времени и вычислительных ресурсов. Такой размер выборки соответствует стандартному для задачи генерирования подписей к изображениям, частным случаем которой является эта работа. В это же время, реализм предполагает большее сходство изображаемых предметов с реальными объектами, что позволяет использовать в ходе обучения модели, распознающие объекты на фотографиях.

	Author	Name	Year	Style	Filename
count	78587	78587	46517.000000	78587	78587
unique	1105	64711	NaN	27	77298
top	vincent van gogh	self portrait	NaN	Impressionism	viktor-vasnetsov_crucified-christ-1896.jpg
freq	1889	493	NaN	12987	2
mean	NaN	NaN	1867.140508	NaN	NaN
std	NaN	NaN	122.398838	NaN	NaN
min	NaN	NaN	1029.000000	NaN	NaN
25%	NaN	NaN	1874.000000	NaN	NaN
50%	NaN	NaN	1902.000000	NaN	NaN
75%	NaN	NaN	1926.000000	NaN	NaN
max	NaN	NaN	2012.000000	NaN	NaN

Рис. 1: Основная информация по датасету WikiArt

В датасете изображения представлены в высоком разрешении. Модели, которые мы построим, работают с меньшими размерами, поэтому необходима предварительная обработка входных данных. Нам потребуются два размера изображений для разных моделей: 224x224 и 299x299 в формате RGB. После обработки размер набора из 10тыс изображений составляет 1.49Гб и 2.64Гб соответственно.

4.2 Обучение с переносом

Для решения поставленной задачи нам необходимо научиться определять предметы, изображенные на картинах. Это задача классификации изображений, решение которой требует специально подготовленных данных, отсортированных по классам. Её решают с помощью обучения глубоких нейронных сетей на огромных базах, например, база ImageNet состоит из более чем 14 миллионов изображений, распределенных по 1000 классам. Чтобы избежать долгой тренировки такой сети, мы будем использовать одну из уже обученных. Для этого нам потребуется технология обучения с переносом.

Идея обучения с переносом(Transfer learning) кроется в передаче знаний (признаков, весов) натренированной модели, которые можно применить для обучения новых моделей.

Модели глубокого обучения - представители индуктивного обучения, т.е такого, где целью является построение отображения с помощью исходных данных, например, в случае классификации это отображение между векторами признаков и классами.

Для работы с новыми данными модели используют предположения о распределении данных, называемые индукционным смещением или inductive bias.

Например, предположение о ближайшем соседе: допустим, что большинство точек из достаточно малой окрестности принадлежат к одному классу. Тогда, получив новую точку, для определения ее класса достаточно проверить класс k

ближайших к ней точек для некоторого k .

Inductive transfer learning применяет индукционное смещение, использованное в задаче-источнике для решения целевой задачи.

Это можно интерпретировать следующим образом. Представим индукционное обучение как поиск модели в некотором пространстве допустимых гипотез. Тогда смещение можно настроить под новую задачу с помощью сужения этого пространства и изменения процесса поиска.

Модели глубокого обучения - многослойные архитектуры, которые вычисляют различные признаки входных данных на различных уровнях. Эти слои затем соединяются с последним слоем, на котором вычисляется результат (например, классификация).

Такое строение позволяет "отрезать" последние слои и использовать оставшиеся ("замороженные") для получения новых признаков данных целевой задачи. Таким образом, ключевая идея обучения с переносом - использовать некоторые слои модели, не изменяя весов на этих слоях во время обучения на новой задаче.

4.3 Модель NIC

4.3.1 Описание

Для достижения поставленной задачи будем использовать модель, построенную на основе статьи Show and Tell: A Neural Image Caption Generator[2]. Нейросети с таким строением успешно используются для генерации подписей к фотографиям с момента публикации работы.

Модель NIC (Neural Image Caption) имеет глубокую рекуррентную архитектуру и совмещает в себе механизмы компьютерного зрения и машинного перевода. Эксперименты, проведенные исследователями на многих датасетах, подтверждают ее аккуратность, т.е. соответствие подписи изображению, и высокое качество получаемого текста.

Принцип работы можно описать следующим образом:

Пусть S_i - слова из словаря, построенного на тренировочной выборке,
 $S^I = S_1, \dots, S_n$ - последовательность слов, наиболее точно описывающая содержание картины I ,
 H - множество всех возможных последовательностей на словаре,
 Тогда задача модели - максимизировать вероятность $p(S^I|I)$,
 то есть $p(S^I|I) = \max_{S \in H} (p(S|I))$

Модель вдохновлена архитектурой генераторов машинного перевода, использующих технологию шифрования - дешифрования, роль которых выполняют 2 различные рекуррентные нейронные сети, соединенные на общем векторном про-

странстве. В случае NIC шифровщик заменяется на сверточную нейронную сеть, предобученную классифицировать объекты на изображении. От нее нам требуется только предпоследний слой, содержащий наиболее полную информацию о изображенных объектах. Он преобразуется определенным образом и подается на вход дешифровщику - рекуррентной нейронной сети, генерирующей текст.

4.3.2 Шифровщик

В модели NIC в качестве шифровщика возьмем глубокую сверточную нейронную сеть VGG16[3]. Для тренировки модели с вниманием возьмем более новую CNN InceptionResNetV2[4]. Обе сети натренированы на задаче классификации изображений на датасете ImageNet. VGG16 способна определять один из 1000 классов, к которому относится предмет на изображении, с вероятностью почти 93%, точность InceptionResNetV2 заявлена как 97%.

Последний скрытый слой CNN содержит информацию об объектах и их расположении на картине, выход этого слоя преобразуется в вектор фиксированной длины, соответствующий размерности матрицы представлений словаря (embedding matrix) и подается на вход дешифровщику.

4.3.3 Дешифровщик

На шаге тренировки модель максимизирует вероятность правильного описания: $\theta_1 = \operatorname{argmax}_{\theta} \sum_{(I,S)} \log(p(S|I, \theta))$, где θ, I, S - параметры модели, изображение, правильное описание соответственно.

Если длина предложения(которая может разниться по сэмплам) $S = N$, то можем считать

$$\log(p(S|I, \theta)) = \sum_{t=0}^N \log(p(S_t|I, S_0, \dots, S_{t-1}, \theta))$$

Таким образом, на тренировочном шаге (I, S) дается на вход, и мы оптимизируем сумму логарифмов вероятностей по всему тренировочному множеству (с помощью стохастического градиентного спуска.)

Покажем, что естественно моделировать $p(S_t|I, S_0, \dots, S_{t-1}, \theta)$ с помощью рекуррентной нейронной сети. Обновляющееся на каждом шаге количество слов в предложении обозначается с помощью так называемого скрытого вектора h_t . Он обновляется после появления новых входных x данных с помощью нелинейной f , т.ч $f(x, h_t) = h_{t+1}$

Таким образом, пусть f является необходимым дешифровщиком, рекурсивно генерирующим слова последовательности для описания, а x_i - выходные данные шифровщика. Дешифровщик должен предсказывать новые слова на основе уже сгенерированной части предложения, значит, он должен ее помнить. Поэтому в качестве f используются рекуррентные нейронные сети, в частности модель долгой краткосрочной памяти (LSTM) и управляемые рекуррентные блоки (GRU),

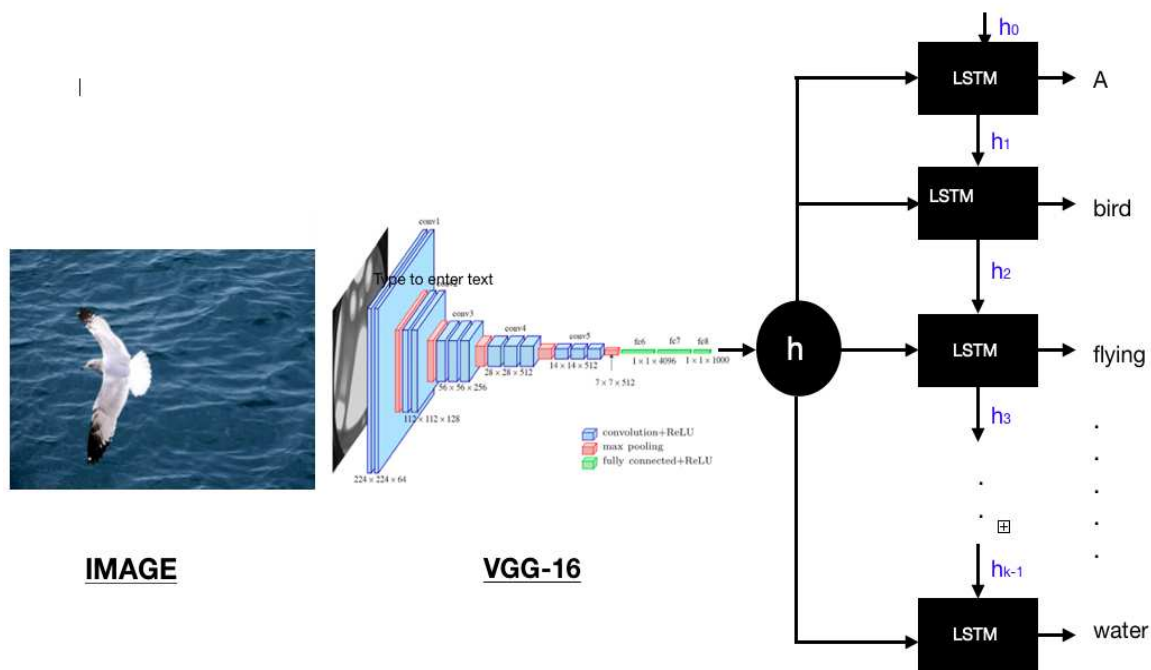


Рис. 2: классическая модель NIC

которые разработаны для подобного рода задач.

LSTM (long short-term memory, дословно долгая краткосрочная память — тип рекуррентной нейронной сети, способный обучаться долгосрочным зависимостям. LSTM-сеть содержит LSTM-модули вместо или в дополнение к другим сетевым модулям. LSTM-модуль — это рекуррентный модуль сети, способный запоминать значения как на короткие, так и на длинные промежутки времени. Ключом к данной возможности является то, что LSTM-модуль не использует функцию активации внутри своих рекуррентных компонентов. Таким образом, хранимое значение не размывается во времени, и градиент не исчезает при использовании метода обратного распространения ошибки во времени при тренировке сети.

GRU(Gated Recurrent Units, дословно управляемые рекуррентные блоки) - новое поколение рекуррентных нейронных сетей, очень похожее на LSTM. По сравнению с LSTM у данного механизма меньше параметров, отсутствуют некоторые из составляющих модуля LSTM. Показывают результаты, сравнимые с проведенными на LSTM или лучше(особенно на рязряженных датасетах небольшого размера).

4.4 Сети с механизмом внимания

При генерации подписей классическим методом шифровщика-дешифровщика, сеть обычно не может охватить все детали изображения, так как фокусируется на его основных частях.

С механизмом внимания (attention) поступающее на вход изображение сначала делится на n частей h_1, \dots, h_n , затем эти части отправляются к шифровщику, и выходные векторы так же по частям поступают в RNN. Таким образом, при предсказании следующего слова RNN описывает отдельную часть изображения. Пусть мы предсказали для изображения i слов, тогда скрытое состояние LSTM на данном шаге h_i . Для предсказания следующего слова механизм внимания использует h_i как контекст для выбора части изображения, которое подходит по содержанию. Внимание выбирает часть z_i , которая подается теперь на RNN, она предсказывает следующее слово, и скрытое состояние обновляется до h_{i+1} .

Механизм:

$e_{jt} = f_{Att}(s_{t-1}, h_j)$, где

e_{jt} означает, как важен пиксель j на шаге t дешифровщика

f_{Att} - функция внимания - простая нейросеть с прямой связью, выходное значение - скаляр

s_{t-1} - предыдущее состояние дешифровщика (предсказанная последовательность)

h_i - состояние шифровщика (i -я часть изображения, обработанного сверточной сетью)

далее для получения вероятностного распределения

$$a_{jt} = \text{Softmax}(e_{jt}) = \frac{\exp(e_{jt})}{\sum_{k=1}^T \exp(e_{kt})}$$

теперь считаем взвешенную сумму, получаем контекстный вектор

$$c_t = \sum a_{jt} h_j$$

на основании которого дешифровщик (RNN) предсказывает следующее слово y_t , получая таким образом последовательность s_t :

$$s_t = \text{RNN}(s_{t-1}, y_{t-1}, c_t)$$

Мы будем использовать Bahdanau Attention - механизм локального внимания, когда контекстный вектор суммирует не все n штук h_j , а только с наибольшей вероятностью имеющих отношение к e_{it} , т.е с наибольшими a_{jt}

4.5 BLEU метрика

Качество полученных подписей будем измерять с помощью метрики BLEU (bilingual evaluation understudy - двуязычный дублер оценки), используемой как в задаче генерации подписей, так и в оценке качества машинного перевода (ее основное предназначение). Оценка BLEU высоко коррелирует с человеческой при оценке текста и широко используется несмотря на то что впервые была представлена [5] еще в 2002г.

Основная идея метрики BLEU заключается в подсчете совпадений фраз пере-

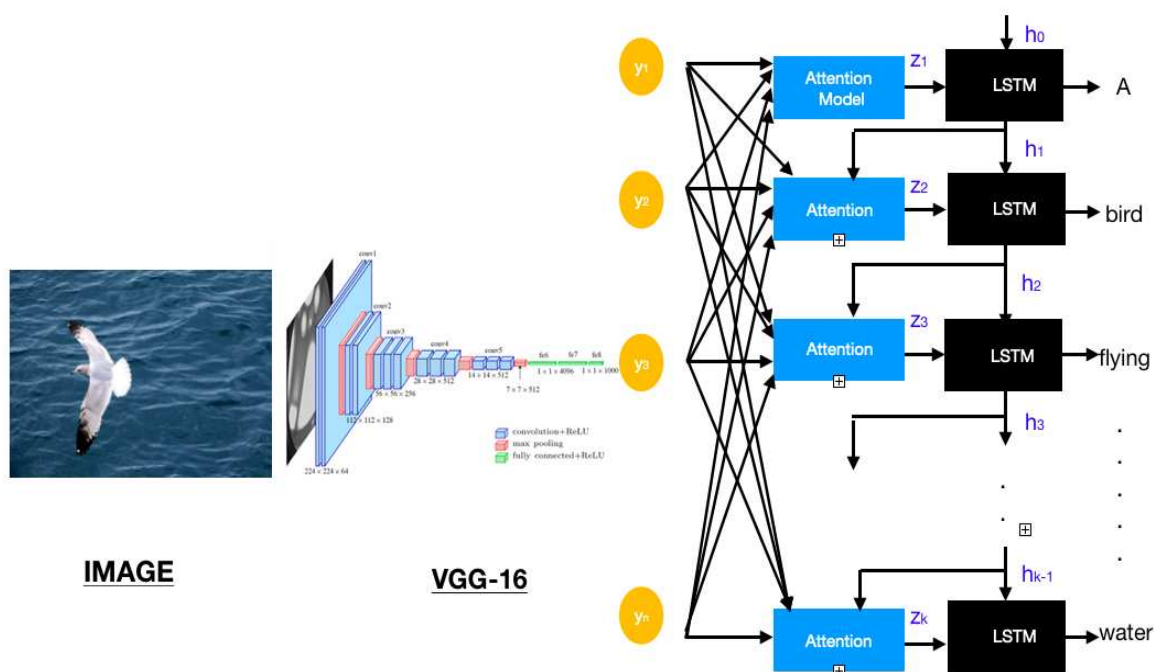


Рис. 3: Модель с Attention

менной длины (N-грамм) в сгенерированном и настоящем названии картины.

5 Результаты

5.1 NIC

Модель NIC на основе шифровщика-дешифровщика(без внимания):

VGG16 + LSTM, batch size = 1, learning rate = 0.00001, оптимизатор Adam, размер тренировочной и валидационной выборки 7975 и 2659, метрика accuracy, потери categorical crossentropy, 30 эпох.

Средний BLEU для юниграмм(N-грамм длины 1) = 0.15




5.2 Attention

Модель с Bahdanau Attention:

InceptionResNetV2 + GRU, batch size=32, оптимизатор Adam, размер тренировочной и валидационной выборки 7975 и 2659, метрика accuracy, потери categorical crossentropy, 10 эпох.

Средний BLEU для юниграмм = 0.21

Таблица 1: Примеры

Изображение	Результат
	<p>Оригинальное название: peasant woman digging Предсказанное название1: a woman and a basket Bleu1: 0.66 Предсказанное название2: girl carrying on a basket digging Bleu2: 0.63</p>
	<p>Оригинальное название: portrait of dr edward anthony spitzka Предсказанное название1: head of the artist Bleu1: 0.42 Предсказанное название2: priest of grain Bleu2: 0.27</p>
	<p>Оригинальное название: the beach at trouville at low tide Предсказанное название1: landscape in the river Bleu1: 0.27 Предсказанное название2: road in the chestnut figures Bleu2: 0.44</p>

Предсказанное название1 - NIC,
Предсказанное название2 - Attention

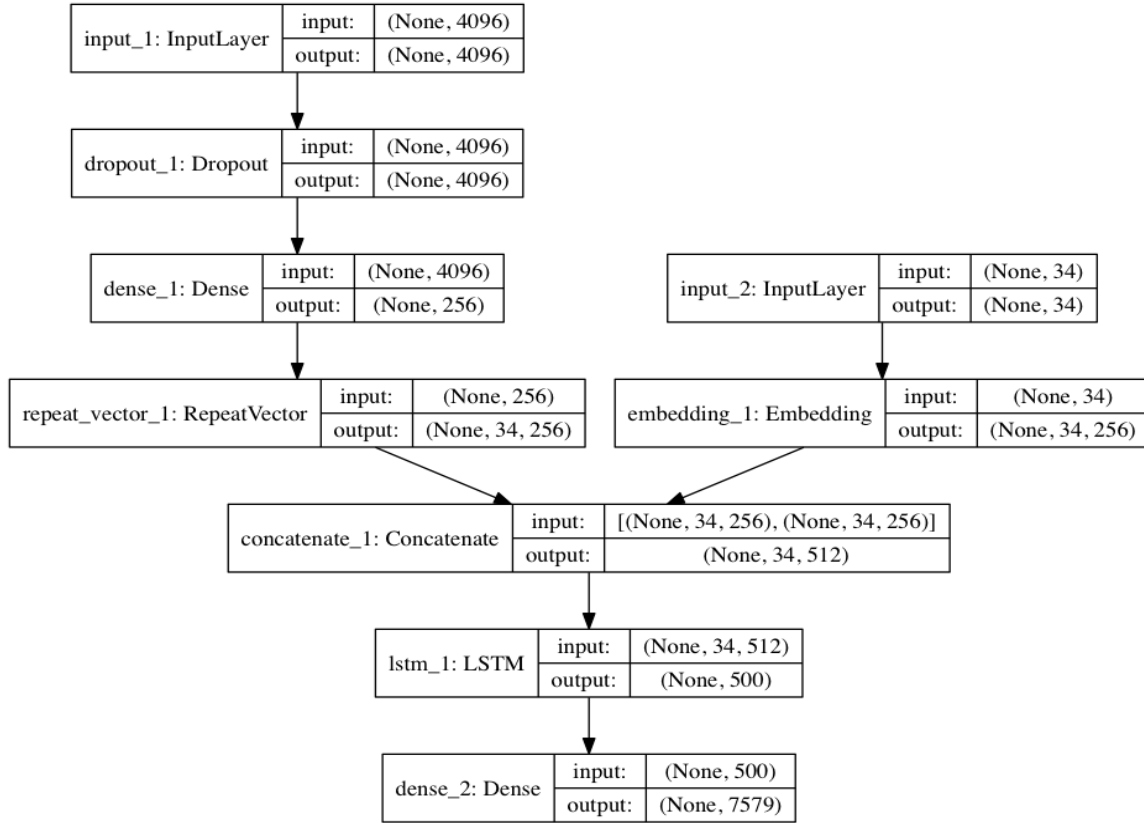


Рис. 4: Строение модели NIC

6 Ограничения подхода

1. Метрика BLEU не всегда точно отражает качество полученного названия, так как нацелена на прямое сравнение слов. Например, для картины Винсента Ван Гога "Orphans" сгенерированное название "village soldier" кажется вполне удачным, однако его Bleu score = 0, так как слова не пересекаются с оригинальным названием.

2. Еще одно препятствие - сами названия, а именно имена собственные и географические названия. Они присутствуют довольно часто и не всегда идентифицируемы. Обучение на таких данных приводит, например, к тому, что для картины Василия Верещагина "The three main gods in a chingacheling buddhist monastery in sikkim" генерируется описание "Pskov church" (также нулевой bleu.)

С другой стороны, авторы включают имена собственные в картины повсеместно, поэтому если модель не будет с ними работать, мы получим далекое от реаль-



Рис. 5: "Orphans"

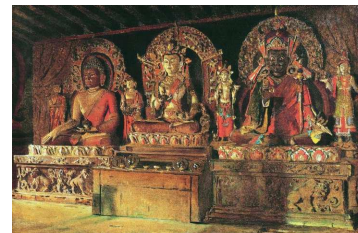


Рис. 6: "The three main gods in a chingacheling buddhist monastery in sikkim"

ности пространство генерируемых имён.

3. Также подход не позволяет адекватно описать картины в нереалистичных стилях, так как объекты на них плохо распознаются шифратором (ведь он обучен на фотографиях).

7 Дальнейшая работа

1. Обучить модель на всем датасете
2. Применить другие техники, такие как мультиобучение и обучение с переносом
3. Использовать шифратор, обученный на стилизованном датасете
4. Подобрать более подходящие метрики для оценки

8 Список литературы

1. WikiArt <https://www.wikiart.org/>
2. Show and Tell: A Neural Image Caption Generator <https://arxiv.org/pdf/1411.4555.pdf>
3. VGG16 <https://arxiv.org/abs/1409.1556>
4. InceptionResNetV2 <https://arxiv.org/abs/1602.07261>
5. BLEU <https://www.aclweb.org/anthology/P02-1040.pdf>
6. Show, Attend and Tell: Neural Image Caption Generation with Visual Attention <https://arxiv.org/abs/1502.03044>
7. LSTM https://www.researchgate.net/publication/13853244_Long_Short-term_Memory