

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!*

*Д.А. Григорьев*

11 июня 2021 г.

## Содержание

Содержание.....	2
Введение.....	3
Описание и подготовка датасета .....	3
Описание подхода.....	3
Обзор литературы .....	5
Реализация моделей .....	8
Выводы и дальнейшая работа.....	9
Примеры.....	11
Список литературы.....	12

## **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.

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: датасет WikiArt, нейронная сеть с механизмом внимания Bahdanau Attention, Multihead Transformer, метрики BLEU, CIDEr, ROUGE-L и METEOR.

## **Описание и подготовка датасета**

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](http://wikiart.org)[1].

В результате очистки датасета от повторяющихся данных и изображений без названия осталось 76864 уникальных картин. Они были обрезаны до размера 224x224 для подачи на вход шифровщику, затем предварительно им обработаны с сохранением признаков с предпоследнего слоя на жестком диске. Это позволило сэкономить время для непосредственной тренировки модели. Полученные данные были разделены на тренировочную и тестовую выборку по 57648 и 19216 изображений соответственно.

## **Описание подхода**

Для генерации названий будем использовать две модели и сравним их результаты. Во первых, нам понадобится модель, построенная на основе статьи Show, Attend and Tell:

Neural Image Caption Generation with Visual Attention[2]. Нейросети с таким строением успешно используются для генерации подписей к фотографиям с момента публикации работы.

Модель 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.

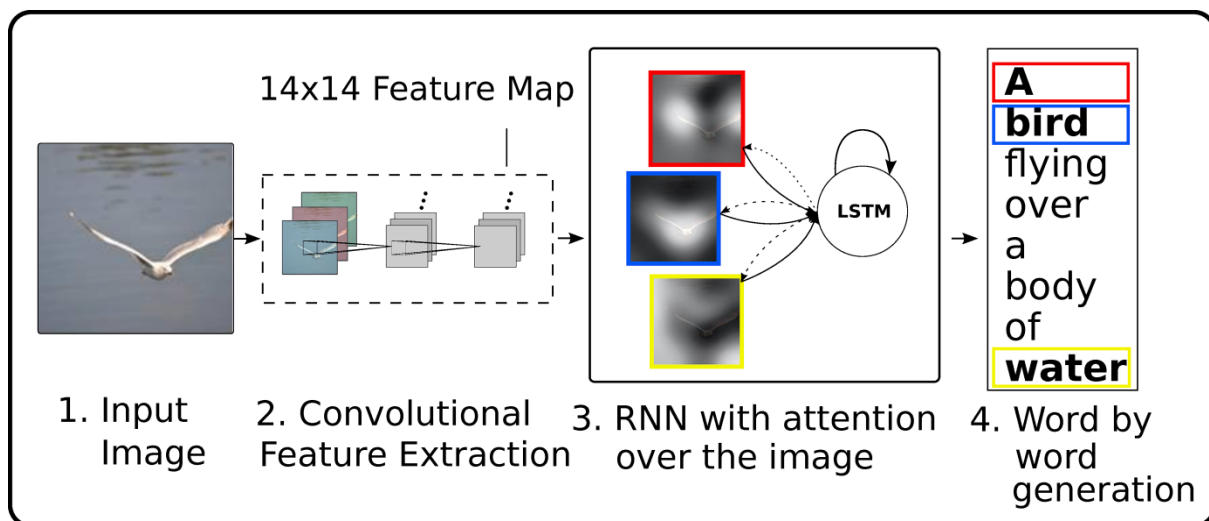
Модель вдохновлена архитектурой генераторов машинного перевода, использующих технологию шифрования - дешифрования, роль которых выполняют 2 различные рекуррентные нейронные сети, соединенные на общем векторном пространстве и взаимодействующие с помощью механизма внимания. В случае генерации подписей 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.

С механизмом локального внимания Bahdanau Attention поступающее на вход изображение сначала делится на  $n$  частей, затем эти части отправляются к шифровщику, и выходные векторы из него так же по частям поступают в RNN. Таким образом, при предсказании следующего слова RNN описывает отдельную часть изображения.

В качестве шифровщика возьмем глубокую сверточную нейронную сеть InceptionResNetV2 [3]. Сеть натренирована на задаче классификации изображений на датасете ImageNet. InceptionResNetV2 способна определять один из 1000 классов, к которому относится предмет на изображении, с точностью 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.

Дешифровщик должен рекурсивно генерировать слова последовательности для описания на основе уже сгенерированной части предложения. В качестве дешифровщика мы будем использовать рекуррентную нейронную сеть на основе управляемых рекуррентных блоков GRU [4], которая была разработана для подобного рода задач. Особенностью сетей такого типа является способность обучаться долгосрочным зависимостям, которая нам необходима.



Во вторых, воспользуемся моделью на основе архитектуры трансформера, которая является логическим продолжением архитектуры Show, Attend and Tell, и впервые была представлена в Attention is all you need [6]. В трансформере нет рекурсивной сети в дешифровщике, он работает исключительно на механизме внимания, что позволяет избежать таких недостатков RNN, как “забывчивость” - даже сети с long-short term memory теряют связь между словами в длинных предложениях, и невозможность генерировать слова в предложении одновременно(рекуррентность). Таким образом, внимание используется в трансформере в трех местах: внутренняя в шифровщике, внутренняя в дешифровщике, и между ними, когда сгенерированное предложение “обращает внимание” на входную последовательность шифровщика.

### Обзор литературы

Архитектура Bahdanau Attention - логическое продолжение модели NIC(Neural Image Caption), представленной в Show and Tell: A Neural Image Caption Generator [5], которая также построена на основе шифрования-дешифрования, но в которой отсутствует механизм внимания. Без него алгоритм не фокусируется на отдельных объектах изображения, а пытается охватить все целиком, что часто приводит к менее точным предсказаниям.

На основе Трансформера и unsupervised pre-training был разработан подход GPT [7], авторы которого обучили трансформер на больших объемах текста, в результате чего GPT показывает отличные результаты на множестве самых разных задач по обработке естественного языка. По сути, такая нейронная сеть на своем выходе создает векторные представления для слов, и даже целых фраз. А с помощью fine-tuning на последних слоях такой языковой модели можно дообучить эту нейронную сеть на многие задачи.

Последним словом в этой цепочке, пожалуй, является BERT: Pre-training of Deep

Bidirectional Transformers for Language Understanding [8]. Bert совмещает идеи предобучения трансформера на большой выборке без учителя, контекстно-зависимых вложений (contextual representations) и двунаправленности, что позволяет модели показывать лучшие результаты в NLP задачах. BERT также легко тюнингуются с помощью добавления всего одного слоя, что позволяет адаптировать модель на множество задач.

Как уже было сказано выше, в данной работе рассматриваются архитектуры из Show, Attend and Tell [2] и Attention is all you need [6], так как они являются своего рода классикой и легко применимы для нашей задачи. В дальнейших этапах для сравнения результатов планируется и рассмотрение других упомянутых моделей.

Качество полученных подписей будем измерять с помощью метрики BLEU [9], используемой как в задаче генерации подписей, так и в оценке качества машинного перевода. Оценка BLEU высоко коррелирует [9] с человеческой и широко используется несмотря на то, что впервые была представлена еще в 2002г. BLEU использует измененную форму метрики precision: отношения

$$\text{True\_positives} / \text{All\_positives},$$
то есть в нашем случае

$$\text{Количество\_слов\_из\_перевода\_которые\_есть\_в\_оригинале} / \text{Длина\_перевода},$$
то есть для каждого слова из перевода смотрим, есть ли оно в оригинале (правильном переводе) и, если есть, прибавляем к числителю единицу.

Простая precision требует изменений, поскольку известно, что системы машинного перевода генерируют больше слов, чем в оригинальном тексте, и любят повторять одни и те же слова. В BLEU решено модифицировать сумму в числителе так, чтобы каждое уникальное слово засчитывалось не более того числа раз, которое оно встречается в оригинале.

Сказанное выше описывает принцип bleu-1, то есть подсчета юниграмм - отдельных слов. Так как для адекватной оценки текста этого недостаточно, метод распространяется на подсчет n-грамм, то есть вместо отдельного слова считается n идущих подряд слов. Так получается принцип метода bleu-n.

В задаче image captioning роль машинного перевода играет сгенерированная подпись к изображению, роль оригинала (человеческого перевода) - человеческое описание. Обычно в этой задаче рассматривают bleu1 – bleu4, и мы будем следовать этому примеру.

Несмотря на широкое признание и преимущества bleu, за время своего существования она неоднократно подвергалась критике. В работе [10] проведен обзор 34 анализов метрики, из которых автор заключает, что для многих задач за пределами области

машинного перевода bleu не пригодна. В работе [11] также проводится обзор метрик для image captioning со сравнением на стандартном датасете flickr8, из которого также видно, что bleu во многом проигрывает другим подходам.

Следующей метрикой на рассмотрении будет ROUGE [12], изначально предложенная как метод оценки summary. ROUGE означает Recall-Oriented Understudy for Gisting Evaluation, т.е. в отличие от Bleu, использующей модифицированную precision, rouge модифицирует recall: отношение

$\text{True\_positives} / \text{All\_relevant}$ ,

т.е. в нашем случае

$\text{Количество\_слов\_из\_перевода\_которые\_есть\_в\_оригинале} / \text{Длина\_оригинала}$ .

Если precision показывает, насколько результат точен, то recall - насколько он полон.

Модификация rouge похожа на bleu тем, что она также базируется на пересекающихся n-граммах. В дополнение к ним rouge в версии ROUGE-L также использует длины

максимальных совпадающих подстрок, что позволяет ей оценивать структуру построения предложений. ROUGE-L чаще всего используют в image captioning, мы также рассмотрим именно ее.

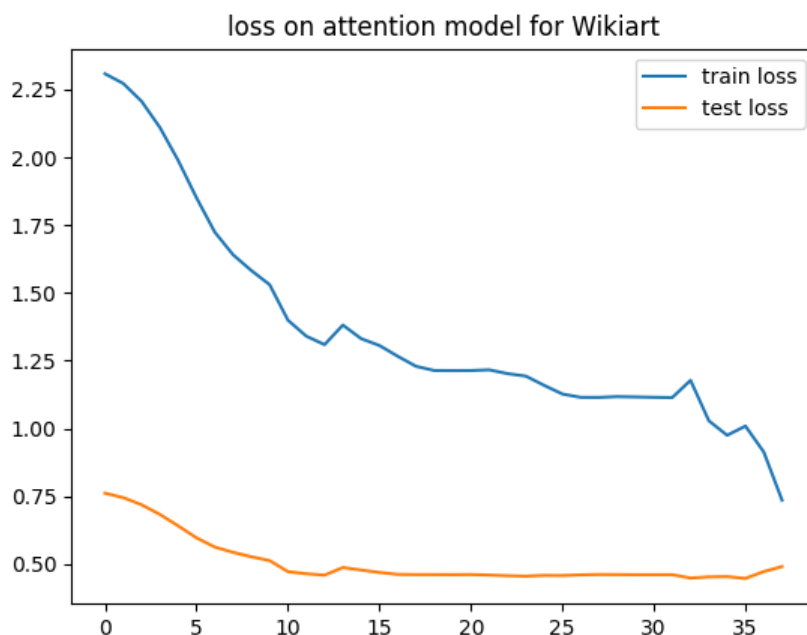
В некотором роде комбинацией этих двух метрик является METEOR [13], определяющийся как среднее гармоническое между precision и recall по юниграммам. Meteor также засчитывает синонимы, однокоренные слова и перифразу с помощью базы данных WordNet. Эта метрика разрабатывалась с целью превзойти метрику bleu за счет устранения ее недостатков, связанных с тем, что она не использует recall и не поощряет различные варианты одного и того же слова.

Мы также рассмотрим метрику CIDEr [14], предложенную в 2015 г. именно для оценки подписей к изображениям. Метрика основана на TF-IDF (TF — term frequency, IDF — inverse document frequency) — статистической мере, используемой для оценки важности слова в контексте документа. Вес некоторого слова пропорционален частоте употребления этого слова в документе и обратно пропорционален частоте употребления слова во всех документах коллекции. Tf-idf считается для n-грамм по n от 1 до 4, затем считается среднее по всем n. Перед подсчетом все слова урезаются до их морфологического корня, чтобы обеспечить поощрение однокоренным словам.

## Реализация модели 1

После тестирования параметров модели с Bahdanau attention(будем называть ее моделью 1) было установлено, что подходящим оптимизатором является Adam с верхним пределом learning rate = 0.0000003 на протяжении основной части тренировки и переходом к циклическому learning rate[15] в диапазоне [0.0000003 - 0.001] на плато. При этом размер бэтча (batch size) был выбран равным 32, количество нейронов на скрытом слое (units) = 512.

Лучшее качество генерируемых названий (по субъективной человеческой оценке на тестовой выборке из 50 изображений) удалось получить в результате тренировки модели в течение 35 эпох, при этом значение оптимизируемой функции потерь изменялось следующим образом:

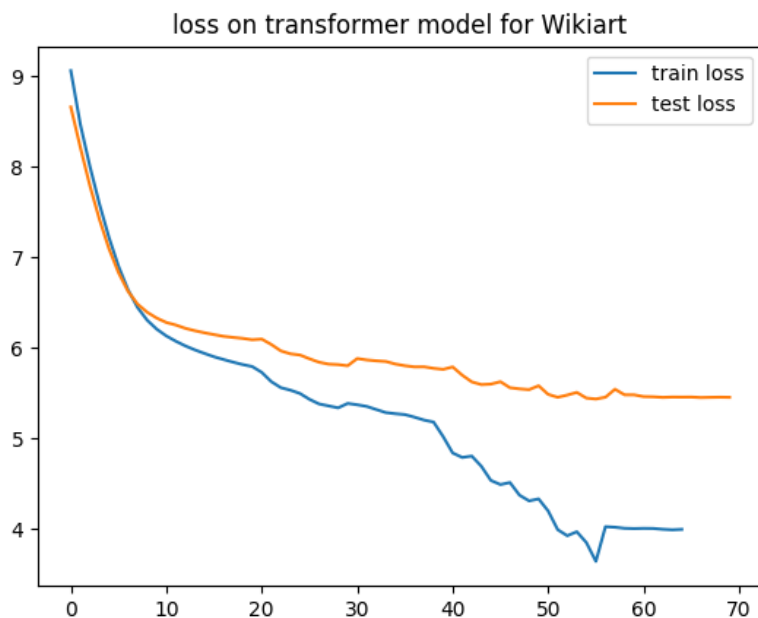


## Реализация модели 2

Для модели на основе архитектуры трансформера(назовем ее модель 2) на данный момент не удалось найти оптимальных параметров. Лучший результат достигается при learning rate =  $1e-6$  и количестве слоев = 6, количество “голов внимания”(number of heads)=16. При этом качество подписей не достигает качества модели 1, в большинстве своем выдавая повторяющиеся результаты. Это наиболее вероятно связано с “застреванием” алгоритма



обучения в локальном минимуме на уровне test loss  $\sim 5.8$ , из которого не удается выйти с помощью увеличения learning rate и изменения основных параметров. Планируется дальнейшая работа с этой моделью с целью получения желаемого результата.



### Выводы и дальнейшая работа

В силу описанных трудностей в работе модели 2, здесь и далее результаты будем приводить для первой модели.

Было отмечено, что в данной задаче стандартные метрики и особенно метрика BLEU не всегда точно отражают качество полученного названия. То же было отмечено в работе Iconographic Image Captioning for Artworks [16], где метрики использовались для оценки качества подписей к иконографическим изображениям.

Результаты для WikiArt(в %)

	модель 1
Bleu-1	5.2
Bleu-2	1.5
Bleu-3	0.4
Bleu-4	0.0000198
Rouge-L	4.2
Cider	1.9
Meteor	2.6



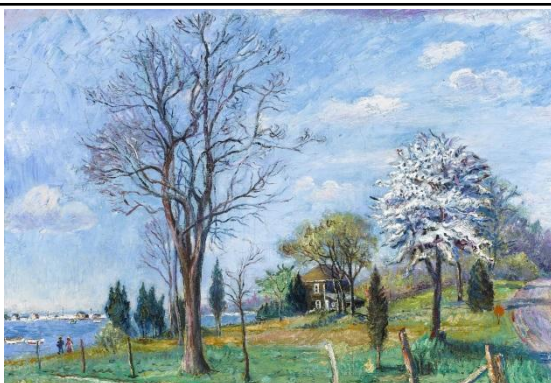
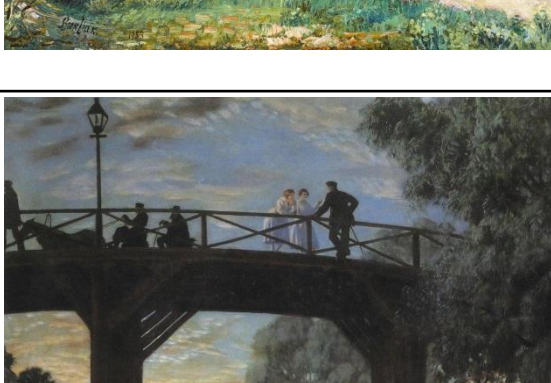
Так как ранее на датасете WikiArt измерения не проводились, сравним результаты с таковыми на стандартном датасете MS-COCO [17] (выполненными на модели с похожей архитектурой), и на датасете Iconclass, представленном в [16] (где использовалась предобученная Vision-language model):

	Ms-coco	Iconclass
Bleu-1	73.1	14.8
Bleu-2	56.2	12.2
Bleu-3	41	11.3
Bleu-4	32.6	10.0
Rouge-L	-	31.9
Cider	87.2	172.1
Meteor	26.1	11.7

Таким образом, при сравнении с MS-COCO с метрикой bleu-4 отрыв самый большой, разница с bleu-1 ~ 14 раз, с meteor ~в 10, с cider ~в 46 раз.

Чтобы понять, вызван такой разрыв особенностью метрик, датасета или же низким качеством подписей, необходима экспертная оценка качества сгенерированных названий и последующее ее сравнение с машинной оценкой. Пока же будем ориентироваться на meteor.

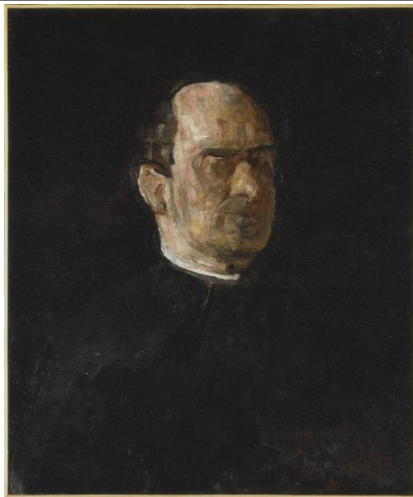
## Примеры:

	<p>Оригинал: the beach at trouville at low tide</p> <p>Предсказание: the fields of the thames</p>
	<p>Оригинал: forest horizon</p> <p>Предсказание: springtime</p>
	<p>Оригинал: a lakeshore</p> <p>Предсказание: landscape with bouquet</p>
	<p>Оригинал: bridge astrakhan</p> <p>Предсказание: sad woman harbor the temple</p>



Оригинал: glass with roses

Предсказание: roses on the season



Оригинал: portrait of dr edward anthony spitzka

Предсказание: portrait of the surrender



Оригинал: pharmacist

Предсказание: coniferous la fondation  
dancing man

## Список литературы

1. WikiArt <https://www.wikiart.org/>
2. Show, Attend and Tell: Neural Image Caption Generation with Visual Attention  
<https://arxiv.org/abs/1502.03044>
3. InceptionResNetV2 <https://arxiv.org/abs/1602.07261>
4. GRU <https://arxiv.org/abs/1406.1078>
5. Show and Tell: A Neural Image Caption Generator <https://arxiv.org/pdf/1411.4555.pdf>
6. Attention is all you need <https://arxiv.org/abs/1706.03762>
7. Improving Language Understanding by Generative Pre-Training  
[https://cdn.openai.com/research-covers/language-unsupervised/language\\_understanding\\_paper.pdf](https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf)
8. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding  
<https://arxiv.org/abs/1810.04805>
9. BLEU <https://www.aclweb.org/anthology/P02-1040.pdf>
10. A Structured Review of the Validity of BLEU  
<https://www.aclweb.org/anthology/J18-3002/>
11. Re-evaluating Automatic Metrics for Image Captioning  
<https://www.aclweb.org/anthology/E17-1019/>
12. ROUGE: A Package for Automatic Evaluation of Summaries  
<https://www.aclweb.org/anthology/W04-1013/>
13. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments <https://www.aclweb.org/anthology/W05-0909/>
14. CIDEr: Consensus-based Image Description Evaluation <https://arxiv.org/abs/1411.5726>
15. Cyclical Learning Rates for Training Neural Networks <https://arxiv.org/abs/1506.01186>
16. Iconographic Image Captioning for Artworks <https://arxiv.org/pdf/2102.03942.pdf>
17. Automatic Image Captioning Based on ResNet50 and LSTM with Soft Attention  
<https://www.hindawi.com/journals/wcmc/2020/8909458/>