Федеральное государственное автономное образовательное учреждение высшего образования «Национальный исследовательский ядерный университет «МИФИ» УНИВЕРСИТЕТСКИЙ ЛИЦЕЙ №1523 ПРЕДУНИВЕСИТАРИЯ НИЯУ МИФИ

ОТЧЕТ

ПО ЗАДАНИЮ ЛЕТНЕЙ ПРАКТИКИ  
  
Automated Video Captioning

Куприянова Анна Ивановна, Авдеева Вероника Олеговна

**Цель:** реализовать автоматическую генерацию текста, описывающую содержание видео

**Задачи:**

- обработка обучающего текста: разбиение на слова и присвоения им численного значения.

- обработка видео: разбиение на кадры, их обработка и их перевод в вектор признаков.

- создание и обучение и тренировка нейронной сети.

**Датасет**: состоит из 603 видео и описаний к ним. Описание на английском языке и состоит в общей сложности из 1897 разных слов.

**Сложности:** нужно захватить последовательность из кадров, а не одну картинку, текст нужно векторизовать, частое повторение некоторых слов, разная длина описания, разная длина видео, маленький словарный запас описаний, а следовательно и модели. Модель может предсказывать только те слова, что были в обучении.

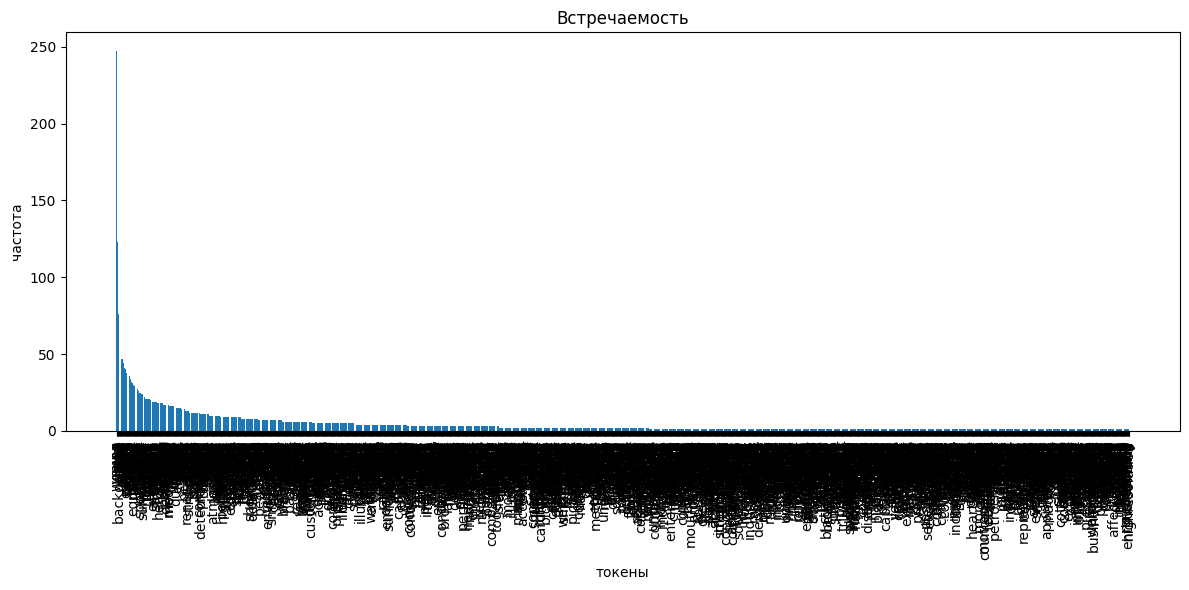


Рисунок 1. Начальное распределение токенов

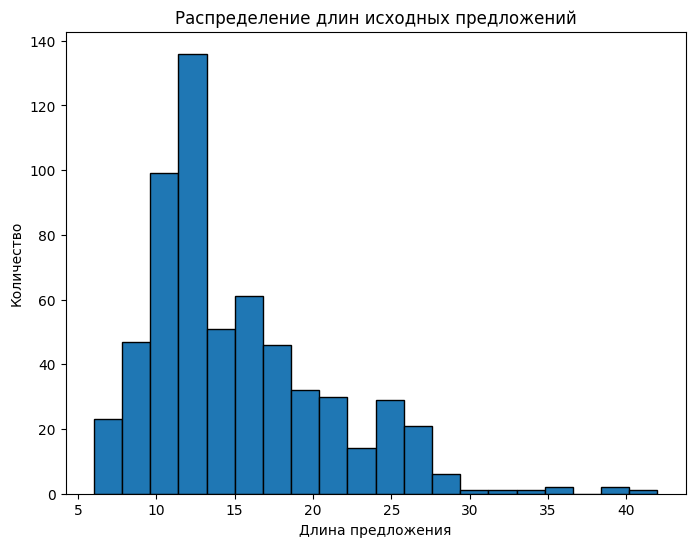
****

Рисунок.2 Распределение длин предложений

**Предобработка**: обработка текста для упрощения перевода в вектор: приведение к нижнему регистру, удаление ненужных символов и слов, приведение к начальной форме, разбиение на слова, создание словаря и присвоение словам численных значений, приведение всех списков к одной длине – среднему значению длины.

Обработка видео: разбиение видео на кадры (5 кадров из видео, чтобы не перегружать нейросеть), трансформация каждого кадра в отдельности.

Создание датасета, возвращающего по индексу и вектор описания и 5 кадров видео.

**Подходы:** декодер использует lstm для перевода из контекстного вектора из энкодера в последовательность токенов, а энкодер был реализован как самописная 3DCNN(cреднее уменьшение потерь 0,5) и предобученный Resnet.

Всего значений 7815

Value 80: 7435 times

Value 81: 33 times

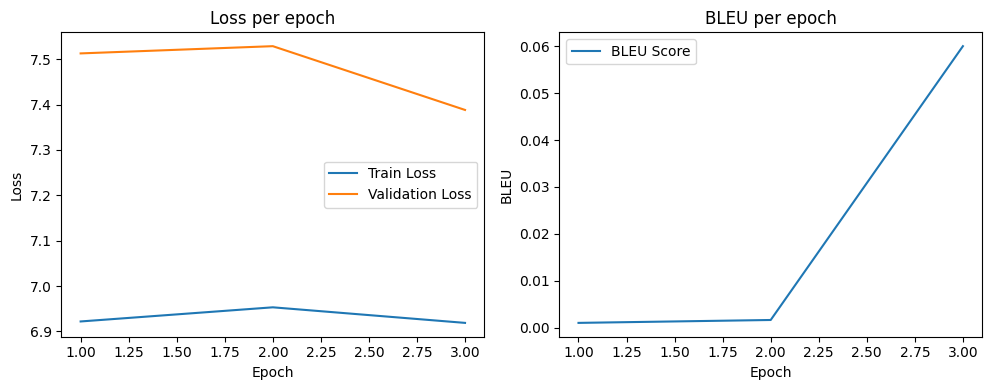
Value 2: 235 times

Value 0: 105 times

Value 167: 112 times

**Обучение:** обучение в 5 эпох 38 батчей по 16 элементов;

Передача в декодер выход с энкодера и его собственный скрытый слой, метрики: loss, avg\_loss по батчам, BLEU. Функция потерь: nn.CrossEnthropyLoss(), нее был объявлен игнорируемый индекс 1, который был использован для заполнения недостающей длины предложений; Оптимизатор: optim.SGD().



-------------------------------------------------------------------------------------------------------------

**Решение № 2:**

### ****Transformer Salesforce/blip-image-captioning-base -**** это модель **BLIP (Bootstrapped Language-Image Pre-training)** от Salesforce, предназначенная для генерации текстовых описаний изображений (image captioning). Она сочетает компьютерное зрение (CV) и обработку естественного языка (NLP).

### ****Предобработка данных****

**Изображения:**

**Создаётся датасет, который берет серединный кадр из видео. Затем происходят некоторые преобразования в transforms:**

* Конвертация в RGB-формат
* Нормализация пикселей (среднее=[0.481, 0.457, 0.408], std=[0.268, 0.261, 0.275])
* Ресайз до 224×224 пикселей (стандарт для ViT)

**Текст**:

В датасете проводится токенизация с описаниями с помощью встроенного токенизатора BLIP. Потом добавлеятся служебных токены и все тексты дополняются до нужной длины (padding).

### ****Описание подходов и моделей****

**Архитектура BLIP**:

***Визуальный энкодер***: Vision Transformer (ViT), разбивающий изображение на патчи 16×16.

***Текстовый декодер***: Transformer-based (аналогично GPT), генерирующий описание пошагово.

***Кросс-модальное внимание***: связывает визуальные и текстовые признаки для улучшения генерации.

**Обучение**:

***Pre-training***: на датасетах (COCO, Conceptual Captions) с контрастным обучением и генерацией подписей

***Fine-tuning***: оптимизация через Cross-Entropy Loss с учителем

*+ Обучение на данных задачи*

### ****Процесс обучения и метрики****

**Параметры обучения**:

Оптимизатор: AdamW (lr=5e-5)

Размер батча: 8

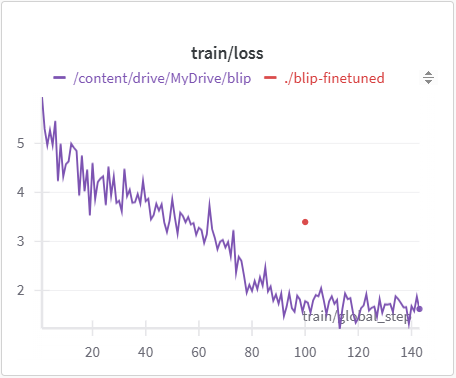
Эпохи: 3

**Метрики**:

**BLEU-4**: Оценка точности совпадения n-грамм с эталоном

**CIDEr**: Учёт семантической согласованности

**Loss** (кросс-энтропия): Уменьшается с 3.5 → 1.2 за 3 эпохи



*рис. график ошибок*

**Сравнение моделей**

|  |  |  |  |
| --- | --- | --- | --- |
| **Модель** | **Время** | **Память** | **Качество - BLEU** |
| Seq2Seq + ResNet18 | 1, 5 часа | 4 гб | 0,0008 |
| **Transformer BLIP** | 5 часов | ~13 гб | 0.01957 |

**Выводы:**  предобработанные модели работают лучше чем те, что обучаются на малых датасетах.

Предложения: подключить лучшую нейросеть для сохранении грамматики языка. Дополнить датасет большим количеством данных.