Выявление манипуляций и их мишеней в новостных текстах.

Лукьяненко Иван

Московский физико-технический институт
Факультет управления и прикладной математики
Кафедра интеллектуальных систем
Научный руководитель: д.ф.-м.н. Воронцов Константин Вячеславович

Лето 2023 г.

Propaganda Detection

Цель

Обобщить методологию задачи Propaganda Detection. Разработать базовые модели для задачи выявления манипуляции и их мишеней на русском языке.

Актуальность

Задача Propaganda Detection является актуальной в условиях повсеместной цифровизации. Отсутствие исследования данной задачи в новостях на русском языке.

Решение

Разработать базовые модели для решения задач в области Propaganda Detection.

Propaganda Detection

1. 2017: Analytical study of language in Propagandistic News[1]
2. 2019: Document-level Propaganda Detection[2]
3. 2019: Span Identification and Span Classification[3]
4. SemEval-20: Task 11: Detection of Propaganda Techniques in News Articles competition
5. SemEval-21: Task 6: Multimodal Propaganda Detection: Text and Images
6. SemEval-23: Task 3: Detecting the Category, the Framing, and the Persuasion Techniques in Online News in a Multi-lingual Setup

[1] **Truth of Varying Shades: Analyzing Language in Fake News and Political Fact-Checking**.

[2] **Proppy: Organizing the news based on their propagandistic content**.

[3] **Fine-Grained Analysis of Propaganda in News Articles**.

Span Identification

Постановка задачи

Подзадача Span Identification - как многоклассовой классификации токенов.

- $p(x^t, \theta)$ - параметрическое семейство моделей для описания распределения вероятности классов каждого токена, $\{x_i^t\}_{i=1}^N$ - последовательность токенов текста $t \in T$, где T - множество всех текстов
- параметр $\theta \in \Theta$, где Θ - пространство параметров модели
- $\{y_i^t\}_{i=1}^N$ - класс i-го токена формате one-hot в тексте t
- $L_1(y, \hat{y}) = \sum_{c \in C} y_c \log(\hat{y}_c)$, где вектор $\hat{y}$ - распределение вероятности классов, $c \in C$ - класс из множества допустимых классов

Задача поиска оптимальных параметров модели сформулирована как:

$$\frac{1}{|T|} \sum_{t \in T} \sum_{i \in t} L_1(y_i^t, p(x^t, \theta)_i) \to \min_{\theta \in \Theta}$$

Span Targeting

Постановка задачи

Подзадача Span Targeting - задача семантического сопоставления текстовых фрагментов.

- $p(x_i^t, x_j^t, x^t, \theta)$ - параметрическое семейство моделей для описания семантической связи между предложениями, x_i^t, x_j^t, x^t - фрагмент, мишень, полный текст
- параметр $\theta \in \Theta$, где Θ - пространство параметров модели
- $y_{ij} \in [0, 1]$ - связаны ли мишень j и фрагмент манипуляции i
- $L_2(y, \hat{y}) = \sum_{j \in J} \sum_{i \in I} y_{ij} \log(\hat{y}_{ij}) + (1 - y_{ij}) \log(1 - \hat{y}_{ij})$, где вектор $\hat{y}$ - вероятность семантической связи фрагмента и мишени.

Задача поиска оптимальных параметров модели сформулирована как:

$$\frac{1}{|I| + |J|} \sum_{i \in I} \sum_{j \in J} L_2(y_{ij}, p(x_i^t, x_j^t, x^t, \theta)) \to \min_{\theta \in \Theta}$$

Оценка качества Span Identification

Метрики качества

Для оценки качества и сравнения моделей необходимо ввести метрики качества решения задачи.

Пусть M - множество токенов, выделенных моделью, E - множество токенов, выделенных экспертом. Введем точность и полноту.

$$C(m, e, h) = \frac{|m \cap e|}{h},$$

$$P(M, E) = \frac{1}{|M|} \sum_{m \in M,\ e \in E} C(m, e, |m|) \tag{1}$$

$$R(M, E) = \frac{1}{|E|} \sum_{m \in M,\ e \in E} C(m, e, |e|)$$

XLM-RoBERTa + адаптеры

XLM-RoBERTa[a] - мультиязычная большая языковая модель.

Определение: *Адаптер* - оператор апроксимации векторного представления.

Для каждой из задач необходимо дообучение своего адаптера. В обеих задачах в качестве адаптера используется обучаемый линейный оператор $L : H \to C$, где H - пространство векторных представлений, C - пространство классов.

[a]Unsupervised Cross-lingual Representation Learning at Scale

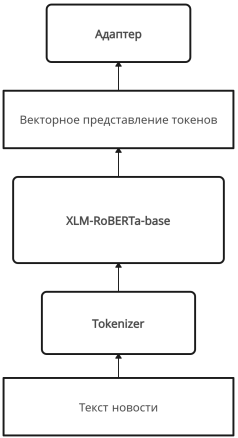


Рис.: Архитектура SI

Архитектура модели ST

Cross-Encoder - архитектура для решения задачи семантического сопоставления текстовых фрагментов.

Основная идея данного подхода заключается в том, что сопоставляемые фрагменты подаются на вход общему энкодеру В своей работе я расширяю данную архитектуру за счет использования контекста всей новости.

$$
\begin{aligned}
y_{span,target} = h_1 = \text{Encoder}_1(span, target)), \\
y_{full\ \ text} = h_2 = \text{Encoder}_2(full\ \ text)), \\
h_3 = \text{concat}(y_{span,target}, y_{full\ \ text}), \\
\text{output} = \text{adapter}(h_3)
\end{aligned} \tag{2}
$$

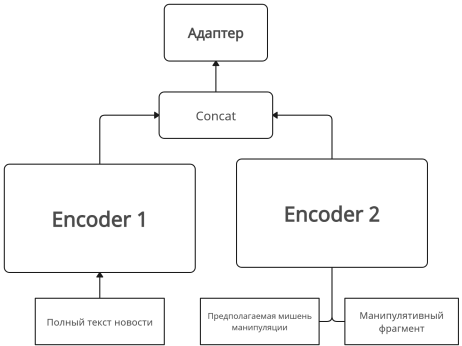


Рис.: Архитектура ST

Малоранговое приближение обновления весов

Малоранговое приближение матриц обновления весов в больших лингвистических моделях

Пусть $W_0 \in \mathbb{R}^{d \times k}$ - предобученные веса, $\Delta W \in \mathbb{R}^{d \times k}$ - изменение весов после дообучения под конкретную задачу. ΔW - будем апроксимировать произведением двух матриц малого ранга $B \in \mathbb{R}^{d \times r} = 0$, $A \in \mathbb{R}^{r \times k} \sim \mathcal{N}(0, \sigma^2)$, где $r << \min(d, k)$

$$Wh = W_0 x + \Delta W x = W_0 x + BAx \qquad (3)$$

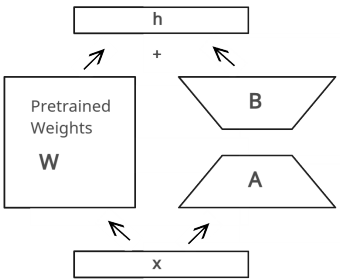


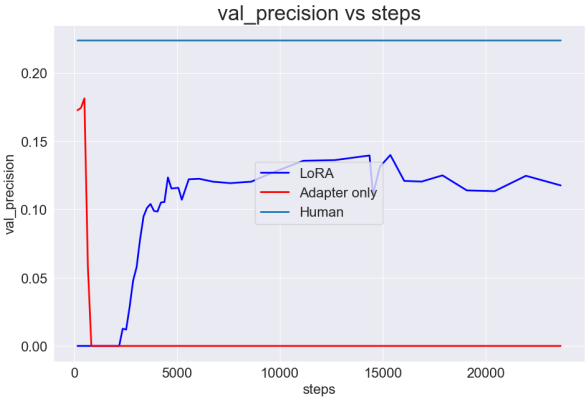
Рис.: LoRA

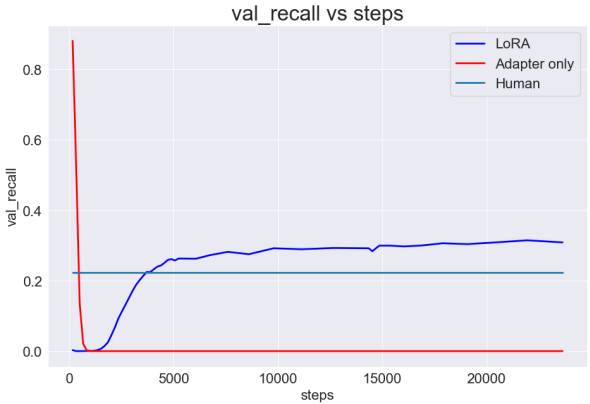
Эксперимент SI

Модель для задачи выделения фрагментов обучалась в двух режимах. Первый - заморозка слоев XLM-RoBERTa-base и обучения только адаптера. Второй - XLM-RoBERTa-base обучалась с использованием малорангового приближения весов и адаптер в стандартном режиме.
Для обоих задач использовался следующий набор гиперпараметров:

1. Оптимизатор AdamW
2. Темп обучения $1 \cdot 10^{-5}$
3. Шедулер: ExponentialLR с параметром затухания 0.99
4. Параметр Dropout-a: 0.2
5. 12 Encoder - слоев в XLM-RoBERTa-base

Визуализация эксперимента SI





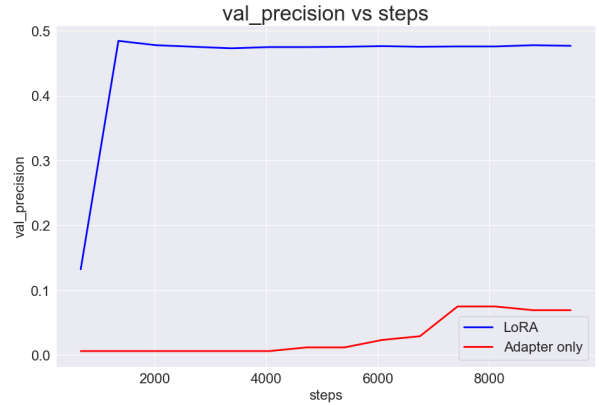
Визуализация эксперимента SI

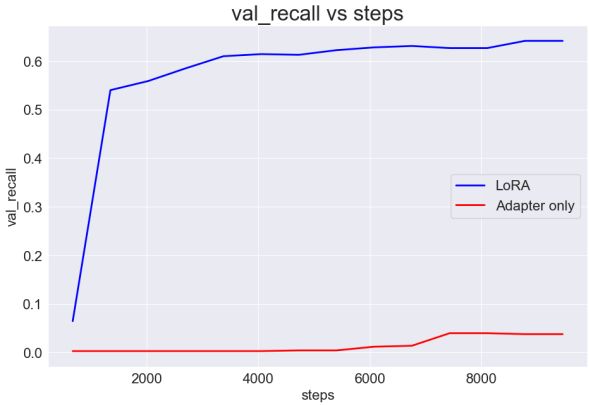
Эксперимент ST

Модель для задачи связи поиска мишени манипуляции обучалась в двух режимах. Первый - заморозка слоев XLM-RoBERTa-base и обучения только адаптера. Второй - XLM-RoBERTa-base обучалась с использованием малорангового приближения весов и адаптер в стандартном режиме.
Для обоих задач использовался следующий набор гиперпараметров:

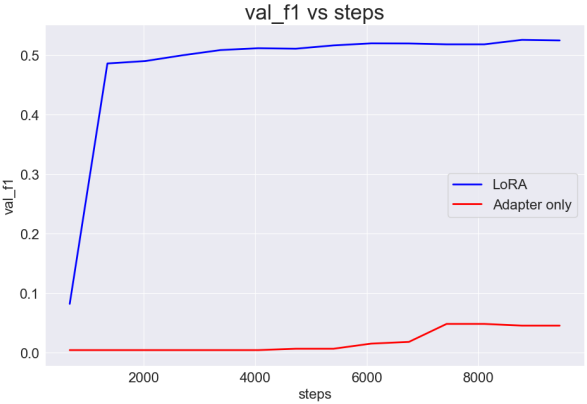
1. Оптимизатор AdamW
2. Темп обучения $1 \cdot 10^{-5}$
3. Шедулер: ExponentialLR с параметром затухания 0.99
4. Параметр Dropout-a: 0.2
5. 12 Encoder - слоев в XLM-RoBERTa-base

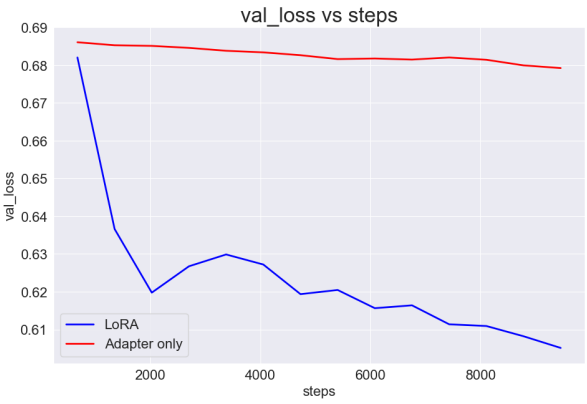
Визуализация эксперимента ST





Визуализация эксперимента ST





Анализ свойств полученных моделей

1. Модель с замороженными слоями после сходимости процесса обучения обладает низкой обобщающей способностью.
2. Модель, в которой происходило дообучение энкодера XLM-RoBERTa-base, после сходимости процесса обучения, обладает заметно лучшей способностью выделять фрагменты манипуляций. Достигнутая полнота выделения манипулятивных фрагментов превышает человеческие показатели, посчитанные между разметчиками обучающего корпуса.

Вывод:

Выделение манипулятивных фрагментов относится к классу задач NLP, в которых исследуемые тексты обладают сложным семантическим устройством языка. Дообучения языковых моделей является необходимым для решения задач связанных с выявлением манипуляций.

Выносится на защиту

Дипломная работа:

1. Разработаны и предложены архитектуры базовых моделей для задач выявления фрагментов манипуляций и поиска их мишеней.
2. Продемонстрирована эффективность обучения языковых моделей с использованием малорангового приближения.
3. Реализован и опубликован код для воспроизведения экспериментов из представленной работы[4].

Публикации на конференциях:

1. IWANN-2023: Long-Term Hail Risk Assessment with Deep Neural Networks[5]

[4] https://github.com/intsystems/Lukyanenko-BS-Thesis

[5] https://arxiv.org/abs/2209.01191