NLP

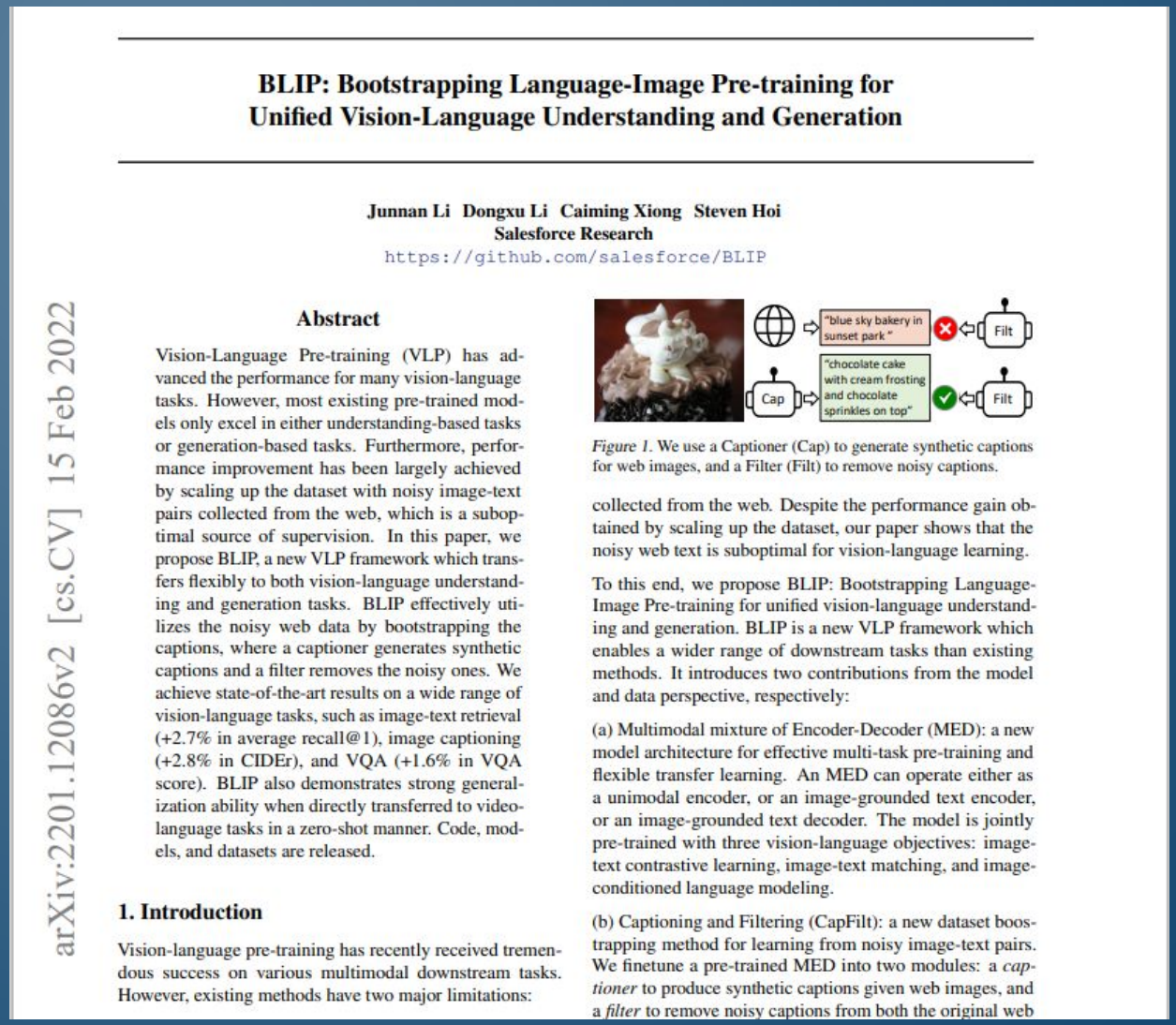
ABSTRACTIVE AND EXTRACTIVE SUMMARIZATION

ТЕРМИН И ПРИМЕНЕНИЕ

Суммаризации(реферирование) текста – процесс создания краткого пересказа длинных документов

Применение:

1. Создание аннотаций научных статей
2. Новостные дайджесты
3. Сжатие и анализ текстовых данных



BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation

Junnan Li Dongxu Li Caiming Xiong Steven Hoi
Salesforce Research
https://github.com/salesforce/BLIP

Abstract

Vision-Language Pre-training (VLP) has advanced the performance for many vision-language tasks. However, most existing pre-trained models only excel in either understanding-based tasks or generation-based tasks. Furthermore, performance improvement has been largely achieved by scaling up the dataset with noisy image-text pairs collected from the web, which is a suboptimal source of supervision. In this paper, we propose BLIP, a new VLP framework which transfers flexibly to both vision-language understanding and generation tasks. BLIP effectively utilizes the noisy web data by bootstrapping the captions, where a captioner generates synthetic captions and a filter removes the noisy ones. We achieve state-of-the-art results on a wide range of vision-language tasks, such as image-text retrieval (+2.7% in average recall@1), image captioning (+2.8% in CIDEr), and VQA (+1.6% in VQA score). BLIP also demonstrates strong generalization ability when directly transferred to video-language tasks in a zero-shot manner. Code, models, and datasets are released.

1. Introduction

Vision-language pre-training has recently received tremendous success on various multimodal downstream tasks. However, existing methods have two major limitations:

Figure 1. We use a Captioner (Cap) to generate synthetic captions for web images, and a Filter (Filt) to remove noisy captions.

collected from the web. Despite the performance gain obtained by scaling up the dataset, our paper shows that the noisy web text is suboptimal for vision-language learning.

To this end, we propose BLIP: Bootstrapping Language-Image Pre-training for unified vision-language understanding and generation. BLIP is a new VLP framework which enables a wider range of downstream tasks than existing methods. It introduces two contributions from the model and data perspective, respectively:

(a) Multimodal mixture of Encoder-Decoder (MED): a new model architecture for effective multi-task pre-training and flexible transfer learning. An MED can operate either as a unimodal encoder, or an image-grounded text encoder, or an image-grounded text decoder. The model is jointly pre-trained with three vision-language objectives: image-text contrastive learning, image-text matching, and image-conditioned language modeling.

(b) Captioning and Filtering (CapFilt): a new dataset bootstrapping method for learning from noisy image-text pairs. We finetune a pre-trained MED into two modules: a *captioner* to produce synthetic captions given web images, and a *filter* to remove noisy captions from both the original web

ЭКСТРАКТНАЯ

1. Алгоритм анализируют текст чтобы определить ключевые предложения. Анализ чаще всего содержит в себе: частота слов, позиция предложения, сходство предложений между собой, кластеризация предложений
2. На основе оценки значимости предложений выбираются лишь самые релевантные предложения.
3. Релевантные предложения соединяются в той же последовательности, в которой они были в тексте, создавая новый текст (реферат)

1. Частота слов – предложения с бОльшим кол-вом частых слов наиболее релевантно.
2. Позиция предложения – первое и последнее предложения наиболее релевантны.
3. Сходство предложений – если одно предложение схоже с 5ью предложениями, то лучше выбрать его, а другие 5 выкинуть.
4. Кластеризация предложений – центральный объект в кластере наиболее репрезентативен

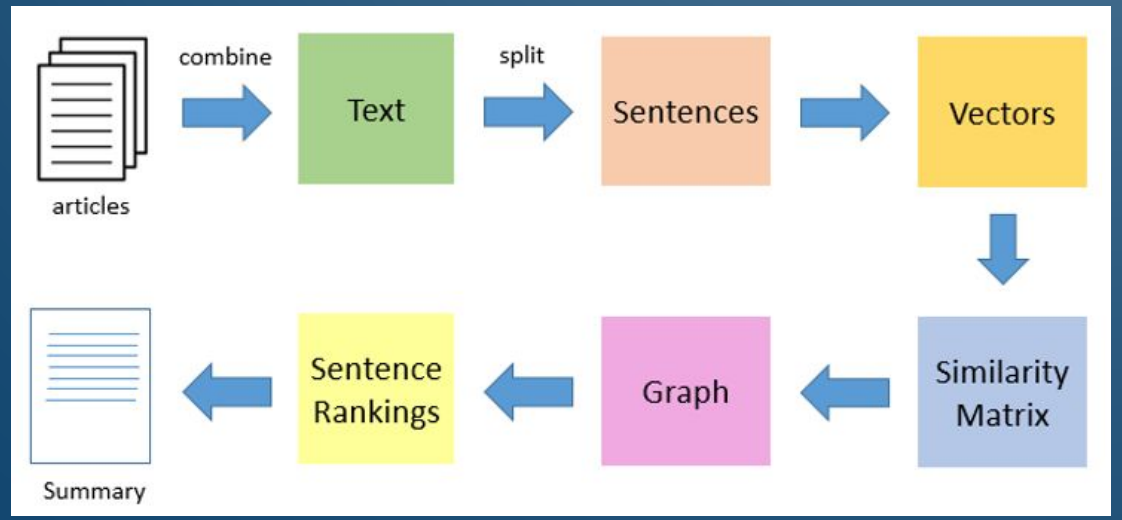
АБСТРАКТНАЯ

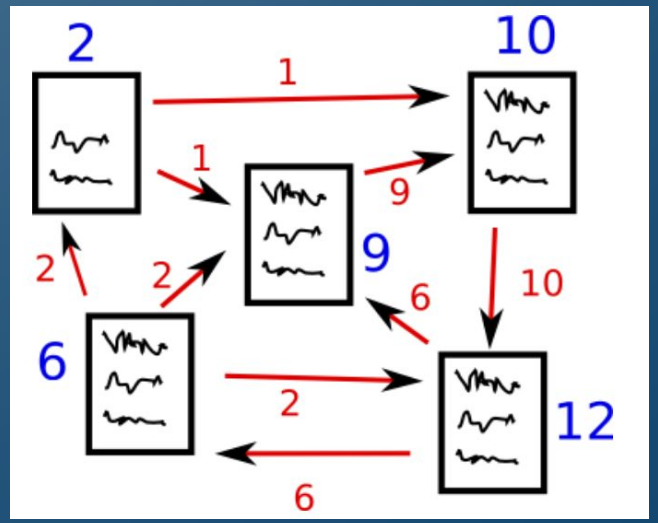
1. Алгоритмы (трансформеры) анализируют исходный текст, определяя его контекст, тему и структуру.
2. Алгоритмы соединяют между собой несколько важных предложений и переформулируют их в одно предложение, оставляя лишь важную суть.
3. Согласуется логическая связность и целостность
4. Генерируется новая последовательность (реферат)

По факту – мы не знаем, как всё происходит внутри трансформеров и LLM, все эти «шаги» расписаны на основе человеческого анализа входного текста и сгенерированного моделью реферата

АЛГОРИТМЫ ЭКСТРАКТНОЙ СУММАРИЗАЦИИ

1. **TextRank** - строит "граф", где каждое предложение — это узел. Узлы соединяются линиями, если предложения похожи друг на друга. Предложения с большим количеством соединений считаются более важными.
2. **LSA (Latent Semantic Analysis)** – создаёт матрицу, где строка – это слово, а столбец – это предложение. В ячейке (на пересечении) могут быть различные значения (к примеру, обычная частота слова в предложении). После чего применяется уменьшение размерности (упрощает данные и увеличивает семантическую связь), на выходе получается новая матрица, которая хорошо передаёт «важность» каждого предложения.
3. **Кластеризация** – предложения (эмбединги) кластеризуются, после чего из каждого кластера берутся самые центральные предложения (они считаются самыми репрезентативными)
4. **LexRank** – такой же как TextRank, но работает исключительно с предложениям (TextRank может еще и со словами). ~~Что говорит интернет:~~ TextRank использует косинусное расстояния для сравнения предложений, а LexRank в этом плане более кастомизируемый и может использовать другие (более продвинутые) метрики сходства, который будут учитывать особую семантику.





АЛГОРИТМЫ АБСТРАКТНОЙ СУММАРИЗАЦИИ

1. **Трансформеры**

МЕТРИКИ

ROUGE – это одна из наиболее распространенных метрик для оценки реферирования текста.

ROUGE обычно используется для оценки полноты (recall), точности (precision) и их гармонического среднего (F1-score). Точность показывает, насколько много информации из сгенерированного реферата соответствует эталонному, а полнота наоборот - насколько много информации из эталонного реферата было в сгенерированном. Гармоническое среднее - сбалансированная оценка точности и полноты, учитывая оба параметра одновременно.

ROUGE включает в себя несколько метрик

1. ROUGE-L: основана на наибольшей общей подпоследовательности между генерируемым рефератом и эталонным рефератом;
2. ROUGE-N: оценивает совпадение n-грамм между генерируемым рефератом и эталонным рефератом.
3. ROUGE-S: дает представление о семантической близости между генерируемым и эталонным рефератами, учитывая не только совпадение слов, но и их относительный порядок.

Используемые метрики

1. **ROUGE-L**
   Input text: The cat is on the mat. (6 words)
   Summary: The cat and the dog. (5 words)
   $$\text{ROUGE-L} = \frac{2(\frac{3}{6} * \frac{3}{5})}{(\frac{3}{6} + \frac{3}{5})} = 0.36$$
2. **ROUGE-N**
   Например, ROUGE-1 соответствует совпадению униграмм (отдельных слов), ROUGE-2 — биграмм (словосочетаний) и так далее
3. **ROUGE-S**
   Например, «The cat» и «The gray cat» – это одно и то же в рамках данной метрики

МЕТРИКИ

BLEU — основывается на точности совпадения n-грамм между сгенерированным и эталонным текстами. BLEU рассчитывает точность для каждого n-грамма (например, 1-грамм, 2-грамм), а затем усредняет эти показатели, применяя геометрическое среднее (иногда с возможным штрафом). Пример:

- Эталонный: "Кот сидит на коврике"
- Сгенерированный: "На коврике сидит кот"

Подсчёт:

- 1-граммы: {кот, сидит, на, коврике} — все 4 слова совпадают, точность = 4/4 = 1.0
- 2-граммы: {кот сидит, сидит на, на коврике} — совпадают 2 из 3, точность = 2/3
- BLEU с учетом 1-грамм и 2-грамм: sqrt(1.0×(2/3)) ~ sqrt(0.66) ~ 0.816

METEOR разработан для более сбалансированной оценки, учитывая не только точность, но и полноту (recall), а также синонимию и стемминг слов. Пример:

- Эталонный: "Кот сидит на коврике"
- Сгенерированный: "Пушистый кот находится на ковре"

Подсчёт:

- Совпадения: "кот" и "ковре" (синоним "коврик" - "ковер")
- Точность (Precision): 2 из 5 слов сгенерированного текста совпадают или синонимичны = 2/5 = 0.4
- Полнота (Recall): 2 из 4 слов эталонного текста найдены = 2/4 = 0.5
- F-score: $2\frac{0.4 \text{ x } 0.5}{0.4+0.5}$ ~ 0.4(4)
- *METEOR с учетом возможного штрафа за стемминг и синонимы будет чуть выше 0.4(4) (в зависимости от того, как реализованы штрафы)