Vocabulary Learning via Optimal Transport for Neural Machine Translation (VOLT)

Семерова Елена (докладчик)
Болотин Арсений (рецензент)
Шапкин Антон (исследователь)
Медведев Антон (хакер)

План выступления

Доклад (Лена Семерова):

MT -> метрика MUV -> Optimal Transport Problem -> VOLT

Рецензия (Арсений Болотин):

Вклад –> Сильные и слабые стороны –> Оценка статьи

Исследование (Антон Шапкин):

Все интересное про статью

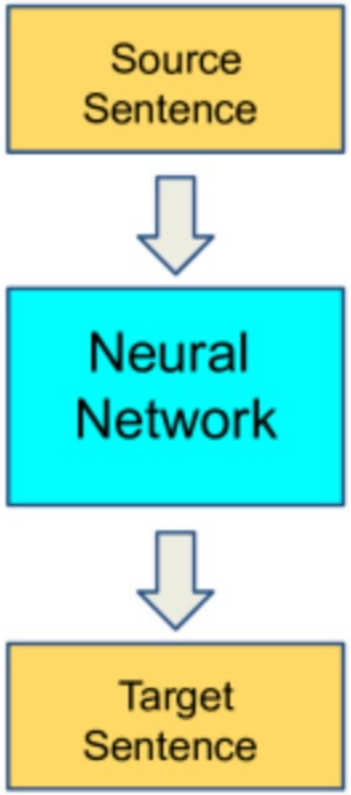
Эксперименты (Антон Медведев):

Реализация алгоритма

Вспомним MT (Machine Translation)

Есть предложение на исходном языке (Source Sentence)

Хотим это же предложение на другом языке (Target Sentence)



Вспомним MT (Machine Translation)

Какие есть способы решения Machine Translation:

- Классические методы:
  - Rule-Based
  - Statistical

- Нейросетевые методы:
  - RNN
  - Attention

Словари в NMT

Обычно главный критерий: частота подслов или энтропия

Хотим учитывать размер словаря

⬇

Получаем MUV - Marginal Utility of Vocabularization

MUV (Marginal Utility of Vocabularization)

Основа: предельная полезность (MU из экономики)

Учитывает: энтропию и размер словаря

v(k) – словарь размера k, $H_{v(k)}$ – энтропия для словаря v(k)

$$\mathcal{M}_{v(k+m)} = \frac{-(\mathcal{H}_{v(k+m)} - \mathcal{H}_{v(k)})}{m}$$

$$\mathcal{H}_v = -\frac{1}{l_v} \sum_{j \in v} P(j) \log P(j)$$

OTP (Optimal Transport Problem)

Есть m поставщиков: $A_1, \ldots, A_m$

Есть n потребителей: $B_1, \ldots, B_n$

Расстояния между поставщиками и потребителями: c_{ij}

Количество груза от A_i к B_j: x_{ij}

Цель: составить план перевозок, чтобы общее пройденное расстояние было минимальным

Суммарное расстояние:

$$z = \sum_{i=1}^{m} \sum_{j=1}^{n} c_{ij} x_{ij}$$

VOLT (VOcabulary Learning approach via optimal Transport)

$$\arg\max_{t} \frac{1}{k} \Big[\arg\max_{v(t) \in \mathbb{V}_{\boldsymbol{S}[t]}} \mathcal{H}_{v(t)} - \arg\max_{v(t-1) \in \mathbb{V}_{\boldsymbol{S}[t-1]}} \mathcal{H}_{v(t-1)} \Big]$$

VOLT (VOcabulary Learning approach via optimal Transport)

$$\arg\max_{t} \frac{1}{k} \Big[\arg\max_{v(t)\in \mathbb{V}_{S[t]}} \mathcal{H}_{v(t)} - \arg\max_{v(t-1)\in \mathbb{V}_{S[t-1]}} \mathcal{H}_{v(t-1)} \Big]$$

1 шаг:

$$\min_{v\in \mathbb{V}_{S[t]}} \frac{1}{l_v} \sum_{j\in v} P(j) \log P(j),$$

$$\text{s.t.} \quad P(j) = \frac{\text{Token}(j)}{\sum_{j\in v} \text{Token}(j)}, \; l_v = \frac{\sum_{j\in v} len(j)}{|v|}.$$

VOLT (VOcabulary Learning approach via optimal Transport)

1 шаг:

$$\min_{v \in \mathbb{V}_{S[t]}} \frac{1}{l_v} \sum_{j \in v} P(j) \log P(j),$$

$$\text{s.t.} \quad P(j) = \frac{\text{Token}(j)}{\sum_{j \in v} \text{Token}(j)}, \; l_v = \frac{\sum_{j \in v} len(j)}{|v|}.$$

$$= \underbrace{\sum_{j \in \mathbb{T}} \sum_{i \in \mathbb{C}} P(j,i) \log P(j,i)}_{\mathcal{L}_1}$$

$$+ \underbrace{\sum_{j \in \mathbb{T}} \sum_{i \in \mathbb{C}} P(j,i)(-\log P(i|j))}_{\mathcal{L}_2}$$

VOLT (VOcabulary Learning approach via optimal Transport)

1 шаг:

$$\min_{v \in \mathbb{V}_{S[t]}} \frac{1}{l_v} \sum_{j \in v} P(j) \log P(j),$$

$$\text{s.t.} \quad P(j) = \frac{\text{Token}(j)}{\sum_{j \in v} \text{Token}(j)}, \; l_v = \frac{\sum_{j \in v} len(j)}{|v|}.$$

$$= \underbrace{\sum_{j \in \mathbb{T}} \sum_{i \in \mathbb{C}} P(j,i) \log P(j,i)}_{\mathcal{L}_1}$$

$$+ \underbrace{\sum_{j \in \mathbb{T}} \sum_{i \in \mathbb{C}} P(j,i)(-\log P(i|j))}_{\mathcal{L}_2}$$

L_1 – отриц.энтропия совместного распределения (обозн. –H(P))

L_2 – скалярное произведение матрицы P и D

матрица P – матрица совместных распределений

матрица D: D_{ij} = $-\log P(i|j)$

$$\min_{\boldsymbol{P} \in \mathbb{R}^{m \times n}} \langle \boldsymbol{P}, \boldsymbol{D} \rangle - \gamma H(\boldsymbol{P})$$

VOLT (VOcabulary Learning approach via optimal Transport)

2 шаг:

На каждом временном шаге t – получаем оптимальный словарь v(t)

Собираем эти словари v(1), …, v(t)

Решаем

$$\arg\max_{t} \frac{1}{k}\left[\arg\max_{v(t)\in\mathbb{V}_{S[t]}} \mathcal{H}_{v(t)} - \arg\max_{v(t-1)\in\mathbb{V}_{S[t-1]}} \mathcal{H}_{v(t-1)}\right]$$

Получаем оптимальный словарь

Сильные стороны:

- Предварительно демонстрируется корреляция MUV с метрикой BLEU.
- Приведены необходимые математические выкладки для сведения задачи максимизации MUV к задаче Optimal Transport
- Эксперименты на задачах двуязычного и многоязычного перевода с использованием двух разных архитектур - Transformer-big и Convolutional Seq2Seq. Сравнение проведено на разных наборах данных и при разных размерах используемого словаря.
- Подход требует намного меньше ресурсов, чем подбор размера словаря для BPE как гиперпараметра
- Предложена новая идея для данной области

Слабые стороны:

- Приводятся результаты только для задач машинного перевода
- Улучшения BLEU сравнительно небольшие, например, в WMT-14 получается добиться намного более значительных улучшений

Другие рецензии

Основные замечания:

- Необоснованное предположение о корреляции оптимизируемой величины с качеством модели
- Использование маленьких наборов данных в экспериментах
- Непонятна состоятельность алгоритма отдельно от BPE