Вариационная модель внимания в сетях глубокого обучения

Камиль Сафин

Научный руководитель:

к. ф.-м. н. Чехович Юрий Викторович

Московский физико-технический институт

5 июня 2019 г.

Постановка задачи

Дано:

$$\mathfrak{D} = \{(\boldsymbol{x}_i, \boldsymbol{y}_i)\},\ i = \overline{1, N},$$

множество пар из разных доменов $\boldsymbol{x}_i \in \boldsymbol{X}$, $\boldsymbol{y}_i \in \boldsymbol{Y}$.
Пара включает в себя два объекта, представляющие разные интерпретации друг друга, например:

- фраза и ее перевод на другой язык;
- текст и его краткое содержание.

Необходимо:
Построить алгоритм F, который по $\boldsymbol{x} \in \boldsymbol{X}$ получает $\hat{\boldsymbol{y}} \in \boldsymbol{Y}$ наиболее похожий на $\boldsymbol{y}$:

$$F:\ \boldsymbol{x} \longrightarrow \hat{\boldsymbol{y}},\ S(\hat{\boldsymbol{y}}, \boldsymbol{y}) \to \max,\ \boldsymbol{x} \in \boldsymbol{X}, \hat{\boldsymbol{y}} \in \boldsymbol{Y},$$

где мера сходства объектов S выбирается исходя из конкретной задачи.

Подходы

Существующие методы решения задач:

- Вариационный кодировщик
  - Повышение «разнообразия» финальных объектов;
  - Эффективное сжатие информации об объекте;
- Кодировщик с механизмом внимания
  - Учет контекстной информации в объектах;
  - Интерпретируемость.

Основная литература

H. Bahuleyan, L. Mou, O. Vechtomova, P. Poupart

Variational Attention for Sequence-to-Sequence Models.

ICLM 2018.

Y. Deng, Y. Kim, J. Chiu, D. Guo, A. M. Rush

Latent Alignment and Variational Attention.

NeurIPS, 2018.

E. Dupon

Learning Disentangled Joint Continuous and Discrete Representations.

NeurIPS, 2018.

E. Jang, S. Gu, B. Poole

Categorical Reparametrization With Gumbel-Softmax.

ICLR 2017.

Цель работы

Исследовать модель вариационного автокодировщика с добавленным механизмом внимания.

Мотивация:

- Добавление информации в процесс декодирования;
- Учет частной и общей информации;
- Интерпретируемость.

Проблема:

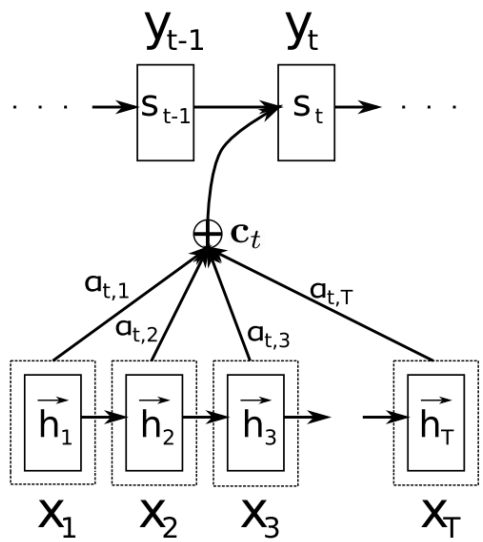
Добавление механизма внимания в стандартном виде ухудшает распределение скрытой переменной.

Механизм внимания

Суть: выделить часть данных для более детальной обработки.

$$\mathbf{c}_t = \sum_{j=1}^{|\boldsymbol{x}|} \alpha_{tj} \mathbf{h}_t, \;\; \mathbf{h}_t = f(\mathbf{x}_t, \mathbf{h}_{t-1}), \;\; t = \overline{1, |\boldsymbol{x}|}.$$

$$p\left(\mathbf{y}_t | \{\mathbf{y}_1, \ldots, \mathbf{y}_{t-1}, \mathbf{c}_t\}\right) = q(\mathbf{y}_{t-1}, \mathbf{c}_t).$$



Вариационная оценка

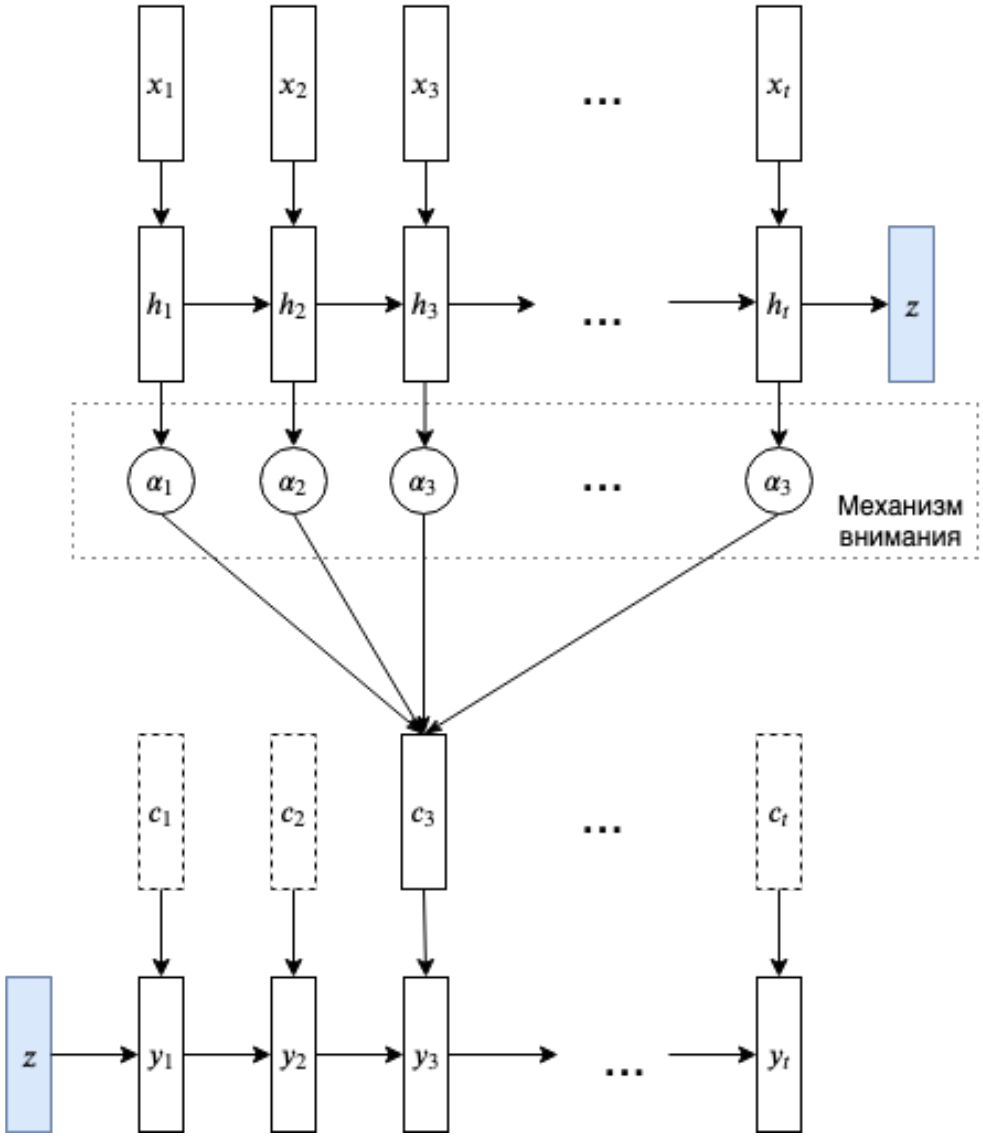
Предполагаем, что объекты получены при условии некоторой скрытой переменной $\mathbf{z}$.
Нижняя оценка функции правдоподобия:

$$\log P(X) \geqslant E_{\boldsymbol{z}\sim Q}[\log P(Y|\boldsymbol{z})] - \mathcal{D}[Q(\boldsymbol{z}|X)||P(\boldsymbol{z})],$$

$\mathbf{z}$ — скрытая переменная,
$p(\mathbf{z})$ — априорное распределение,
$q(\mathbf{z}|\mathbf{X})$ — апостериорное распределение,
$\mathcal{D}_{KL}$ — дивергенция Кульбака-Лейблера.

Предлагаемый подход

Суть: добавить в модель механизм внимания.



Предлагаемый подход

Проблема: добавление механизма внимания в неизменном виде приводит к тому, что скрытая переменная не хранит информацию о кодируемом объекте.

Теорема. Оптимизация нижней оценки правдоподобия в модели вариационного кодировщика с механизмом внимания:

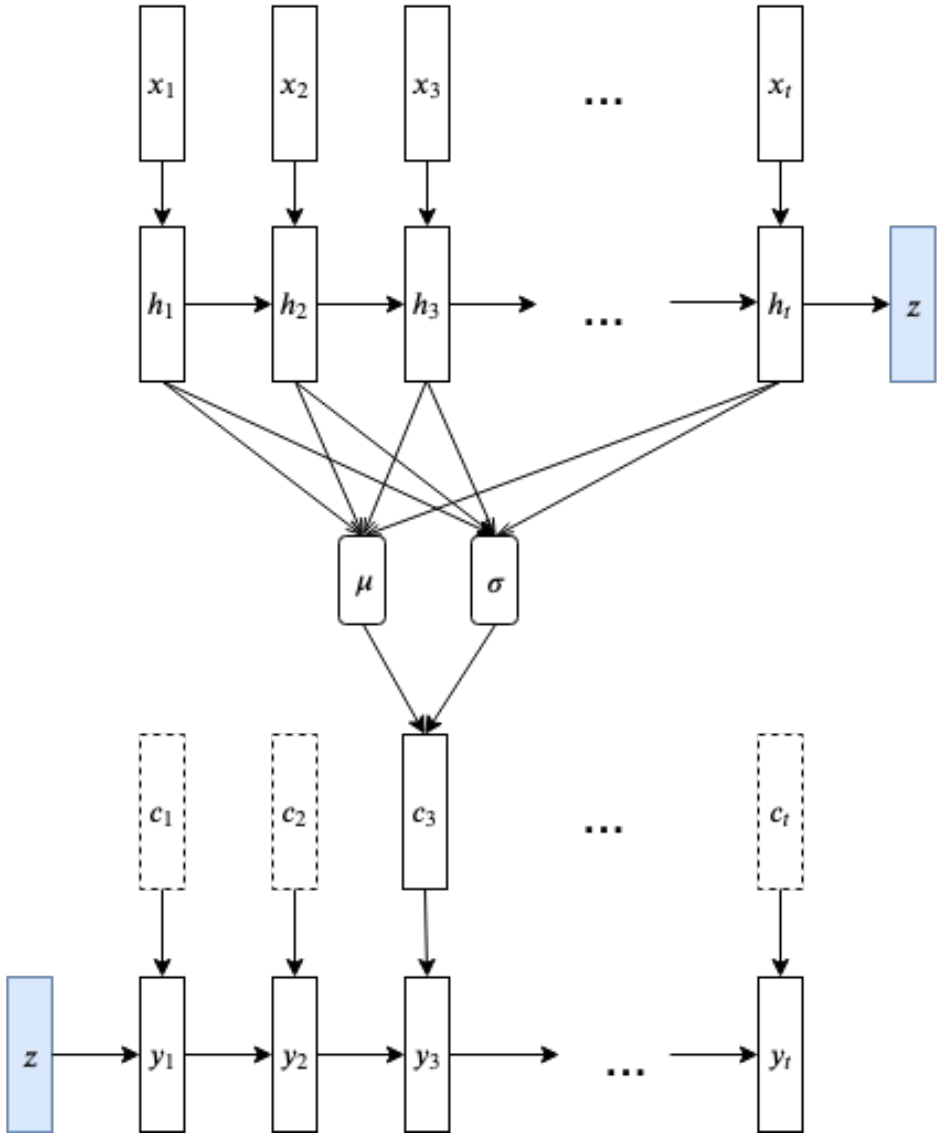
$$\mathcal{L} = E_{\boldsymbol{z}\sim Q}[\log P(Y|\boldsymbol{z})] - \mathcal{D}[Q(\boldsymbol{z}|X)||P(\boldsymbol{z})],$$

равносильна минимизации первого слагаемого, при

$$Q(\boldsymbol{z}|X) \equiv P(\boldsymbol{z})$$

Предлагаемый подход

Решение: введение доп. вариационного пространства.



Предлагаемый подход

Теорема. При введении дополнительного вариационного пространства скрытых векторов $\boldsymbol{a}$, нижняя оценка функции правдоподобия представима в виде:

$$\mathcal{L} = E[\log p(Y|\mathbf{z}, \mathbf{a})] - \mathcal{D}_{KL}[q^{(\mathbf{z})}(\mathbf{z}|X)||p(\mathbf{z})] - \mathcal{D}_{KL}[q^{(\mathbf{a})}(\mathbf{a}|X)||p(\mathbf{a})].$$

Возможно задавать разные структуры пространства векторов $\boldsymbol{a}$:

- Нормальное распределение — семплирование вектора контекста;
- Распределение по весам — семплирование весов входных элементов.

Распределения по весам

Распределение Гумбеля G позволяет семплировать веса для получения вектора контекста.
Свойства распределения:

- Связь с равномерным распределением:

$$G(0,1) = -\log(-\log U(0,1)),$$

где $U(0,1)$ — стандартное равномерное распределение;

- Получение весов:

$$\alpha_t = \frac{\exp((\log \pi_t + g_t)/\tau)}{\sum_{j=1}^{k} \exp((\log \pi_j + g_j)/\tau)}, \; g_i \in G(0,1),$$

где π_i — ненормализованные веса, τ — параметр семплирования.

Эксперимент

Задача: Машинный перевод с английского на французский;
Данные: Параллельный корпус диалогов EuroParl;

- train: 80000,
- test: 10000.

Качество перевода: BLEU метрика.
«Разнообразие» предложений: Entropy, Distinct.

$$\text{Entropy} = \sum_{\omega} p(\omega) \log p(\omega)$$

$$\text{Distinct-1} = \frac{|\text{set}(unigram_i)|}{|\bigcup_i unigram_i|}, \ \text{Distinct-2} = \frac{|\text{set}(bigram_i)|}{|\bigcup_i bigram_i|}$$

Эксперимент

Сравниваемые модели:

- Простая модель кодировщик-декодировщик (Simple AE);
- Модель кодировщик-декодировщик с классическим механизмом внимания (AE with Classic Attention);
- Кодировщик с вариационным механизмом внимания с семплированием по Гауссу (Var Attention, Normal dist);
- Кодировщик с вариационным механизмом внимания с семплированием по Гумбелю (Var Attention, Gumbel dist);

Эксперимент

Качество перевода.

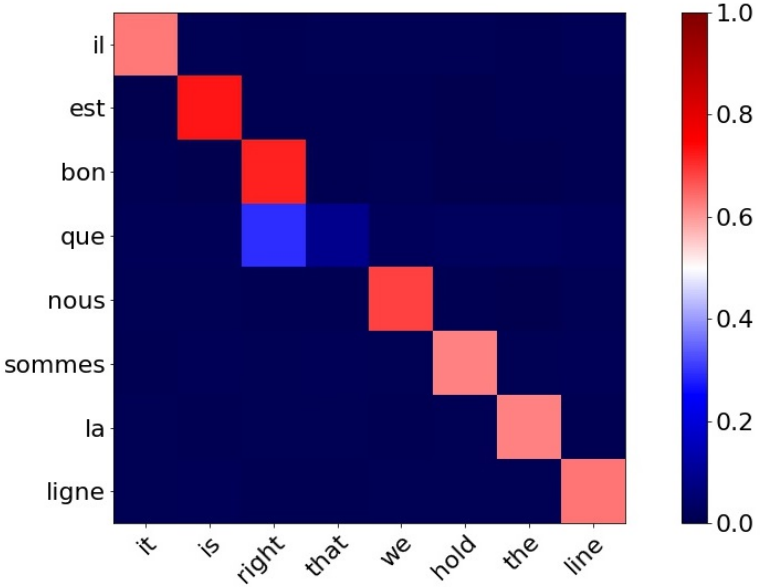
	BLEU-4	BLEU-3	BLEU-2	BLEU-1
Simple AE	24,3	13,5	8,1	4,9
AE with Classic Attention	34,2	22,3	15,14	10,17
Var Attention, Normal dist.	34,7	22,6	15,3	10,2
Var Attention, Gumbel dist	**35,3**	**23,4**	**16,1**	**11,0**

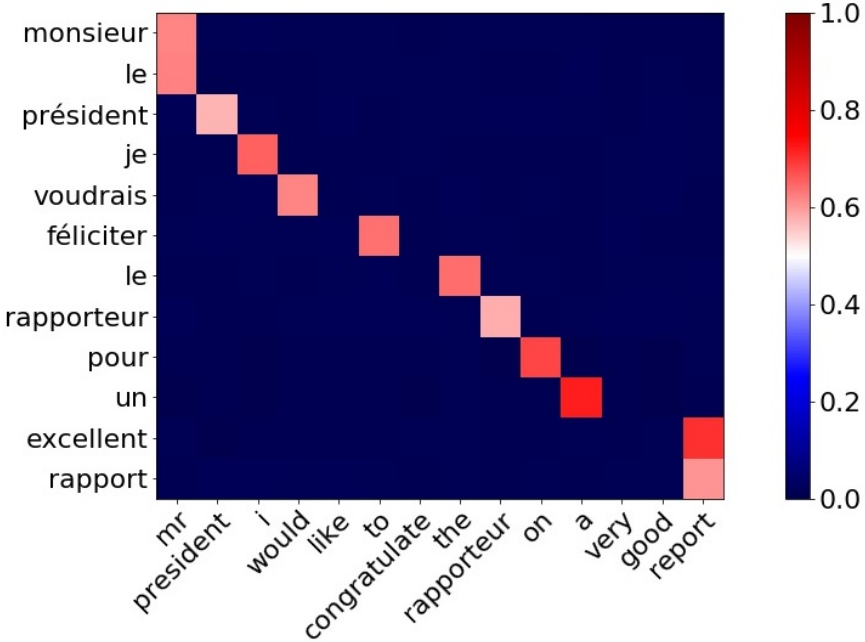
Показатели Entropy и Distinct.

	Entropy	Distinct-1	Distinct-2
Simple AE	2,2	0,073	0,085
AE with Classic Attention	2,4	0,083	0,093
Var Attention, Normal dist.	2,5	0,116	0,173
Var Attention, Gumbel dist	**2,6**	**0.126**	**0.195**

Интерпретируемость компонент

Визуализация семплированных весов из распределения Гумбеля.





Заключение

Основные положения, выносимые на защиту:

1. Предложен и протестирован вариационный механизм внимания с семплированием весов из распределения Гумбеля.
2. Произведено сравнение моделей, использующих разные вариационные распределения.
3. Доказана теорема о разложении правдоподобия модели при введении дополнительного пространства.

Публикации:

R. Kuznetsova, O. Bakhteev, A. Ogaltsov, K. Safin
VBTA: Variational Bi-domain Triplet Autoencoder for learning across domains with the relative constraints.
In review at NeurIPS 2019.

K. Safin, A. Ogaltsov
Detecting a change of style using text statistics.
Proceedings on CLEF, 2018.

K. Safin, R. Kuznetsova
Style Breach Detection with Neural Sentence Embeddings.
Proceedings on CLEF, 2017.