Breaking the Softmax Bottleneck: a High-Rank RNN Lnguage Model

Ирина Понамарева

Higher School of Economics

3 октября 2019 г.

Обзор

Как моделировать язык?

- $\mathbf{X} = (X_1, ..., X_T)$ — корпус токенов
- $P(\mathbf{X}) = \prod_t P(X_t|X_{<t}) = \prod_t P(X_t|C_t)$ — совместное распределение
- $C_t = X_{<t}$ — контекст
- Стандартный подход: закодировать контекст вектором фиксированной длины, умножить на предтавления (эмбеддинги) слов, получить логиты
- Логиты $\rightarrow$ Софтмакс
- Но способен ли такой подход хорошо моделировать вероятность?
- Авторы утверждают, что нет

Модель естественного языка

- Моделируем язык как набор пар (контекст, условное распределение вероятностей следующего токена):
- $L = \{(c_1, P^*(X|c_1)), ..., (c_N, P^*(X|c_N))\}$
- Предполагаем, что $P^* > 0$ для любого слова (невырожденность)
- Цель: $P_\theta(X|C) = P^*(X|C)$, θ — параметр

Софтмакс: стандартный подход

$$\boxed{\mathrm{P}_\theta(x|c) = \frac{\exp \mathbf{h}_c^T \mathbf{w}_x}{\sum_{x'} \exp \mathbf{h}_c^T \mathbf{w}_x}}$$

- $\mathbf{h}_c$ — некоторая функция от контекста c
- $\mathbf{w}_x$ — некоторая функция от слова x
- Обе параметризованы при помощи θ, имеют размерность d
- $\mathbf{h}_c^T \mathbf{w}_x$ называется логитом

Софтмакс как факторизация матриц

$$\mathbf{H}_\theta = \begin{bmatrix} \mathbf{h}_{c_1}^T \\ \mathbf{h}_{c_2}^T \\ \dots \\ \mathbf{h}_{c_N}^T \end{bmatrix} ; \mathbf{W}_\theta = \begin{bmatrix} \mathbf{w}_{x_1}^T \\ \mathbf{w}_{c_2}^T \\ \dots \\ \mathbf{w}_{x_M}^T \end{bmatrix} ;$$

$$\mathbf{A} = \begin{bmatrix} \log P^*(x_1|c_1), & \log P^*(x_2|c_1), & \dots & \log P^*(x_M|c_1) \\ \log P^*(x_1|c_1), & \log P^*(x_2|c_2), & \dots & \log P^*(x_2|c_1) \\ \vdots & \vdots & \ddots & \vdots \\ \log P^*(x_1|c_N), & \log P^*(x_1|c_N), & \dots & \log P^*(x_M|c_N) \end{bmatrix}$$

- $\mathbf{H}_\theta \in \mathbb{R}^{N*d}$, $\mathbf{W}_{theta} \in \mathbb{R}^{M*d}$, $\mathbf{A} \in \mathbb{R}^{N*M}$
- $\mathbf{H}_\theta$ может быть реализовано как выход нейросети, $\mathbf{W}_{theta}$ — некоторый эмбеддинг входного слова

Семейство $F(\mathbf{A})$

Введем следующее семейство матриц:

- $F(\mathbf{A}) = \{\mathbf{A} + \mathbf{\Lambda J}_{N,M} | \mathbf{\Lambda} \text{ is diagonal and } \mathbf{\Lambda} \in \mathbb{R}^{N,N}\}$
- $\mathbf{J}_{N,M}$ — матрица, заполненнная единицами
- По сути, добавляем произвольное целое число к каждому ряду матрицы $\mathbf{A}$.
- В $F(\mathbf{A})$ бесконечное количество элементов
- У $F(\mathbf{A})$ есть два полезных свойства:

Свойства $F(\mathbf{A})$

Свойство 1

Для любой матрицы $\mathbf{A}'$, $\mathbf{A}' \in F(\mathbf{A})$ если и только если $Softmax(\mathbf{A}') = P^*$. То есть $F(\mathbf{A})$ описывает множество всех возможных логитов, которые соответствуют истинному распределению

Свойство 2

Для любых $\mathbf{A}_1 \neq \mathbf{A}_2 \in F(\mathbf{A}), |rank(\mathbf{A}_1) - rank(\mathbf{A}_2)| < 1$. То есть у матриц из $F(\mathbf{A})$ ранг отличается максимум на единицу.

Лемма 1

При фиксированных параметрах θ, $\mathbf{H}_\theta \mathbf{W}_\theta^T \in F(\mathbf{A}')$ тогда и только когда, когда $P_\theta(X|c) = P^*(X|c)$ для всех c из L.

Свойства $F(\mathbf{A})$

Доказательство свойства 1

Возьмем $\mathbf{A}' \in F(\mathbf{A})$, $P_{\mathbf{A}'}(X|C)$ Пусть i — номер ряда, j — номер столбца, тогда $A'_{ij} = A_{ij} + \Lambda_{ii}$. Тогда

$$P_{\mathbf{A}'}(x_j|c_i) = \frac{\exp A'_{ij}}{\sum_k \exp A'_{ik}} = \frac{\exp(A_{ij}+\Lambda_{ii})}{\sum_k \exp(A_{ik}+\Lambda_{ii})} = \frac{\exp A_{ij}}{\sum_k \exp A_{ik}} = P^*(x_j|c_i)$$

Тогда для $\mathbf{A}'' \in \{\mathbf{A}''|Softmax(\mathbf{A}'') = P^*\}$, для любых i, j мы имеем

$$P_{\mathbf{A}''}(x_j|c_i) = P_{\mathbf{A}}(x_j|c_i)$$

Следовательно, для любых i, j, k

$$\frac{P_{\mathbf{A}''}(x_j|c_i)}{P_{\mathbf{A}''}(x_k|c_i)} = \frac{\exp A''_{ij}}{\exp A''_{ik}} = \frac{\exp A_{ij}}{\exp A_{ik}} = \frac{P_{\mathbf{A}}(x_j|c_i)}{P_{\mathbf{A}}(x_k|c_i)} \rightarrow A''_{ij} - A_{ij} = A''_{ik} - A_{ik}$$

Значит, существует диагональная матрица $\mathbf{\Lambda} \in \mathbb{R}^{N*N}$ такая что

$$\mathbf{A}'' = \mathbf{A} + \mathbf{\Lambda}\mathbf{J}_{N,M} \rightarrow \mathbf{A}'' \in F(\mathbf{A})$$

Доказательство свойства 2

Для любых $\mathbf{A}_1, \mathbf{A}_2 \in F(\mathbf{A})$ по определению $\mathbf{A}_1 = \mathbf{A} + \mathbf{\Lambda_1}\mathbf{J}_{N,M}$, $\mathbf{A}_2 = \mathbf{A} + \mathbf{\Lambda}_2\mathbf{J}_{N,M}$, где $\mathbf{\Lambda_1}, \mathbf{\Lambda_1}$ — диагональные матрицы

$$\mathbf{A}_1 = \mathbf{A}_2 + (\mathbf{\Lambda}_1 - \mathbf{\Lambda}_2)\mathbf{J}_{N,M}$$

Пусть S — максимальное множество линейно независимых столбцов A_2. Пусть $\mathbf{e}_N$ — вектор единиц. Тогда

$$\mathbf{a}_{1,i} = \mathbf{a}_{2,i} + (\Lambda_{1,ii} - \Lambda_{2,ii})\mathbf{e}_N$$

Так как $\mathbf{a}_{2,i}$ есть линейная комбинация элементов из S, то $\mathbf{a}_{1,i}$ — из $S \cup \mathbf{e}_N$. Поэтому ранги матриц $\mathbf{A_1}$ и $\mathbf{A_2}$ отличаются не больше чем на 1.

Вопрос об экспрессивности софтмакса

Существует ли такой параметр θ и $\mathbf{A}' \in F(\mathbf{A})$, что $\mathbf{H}_\theta \mathbf{W}_\theta^T = \mathbf{A}$?

- По сути, это проблема факторизации матрицы. Заметим, что ранг матрицы $\mathbf{H}_\theta \mathbf{W}_\theta^T$ должен быть хотя бы равен рангу $\mathbf{A}'$.
- Ранг $\mathbf{H}_\theta \mathbf{W}_\theta^T$ строго ограничен сверху размерностью эмбеддинга d.
- Таким образом, если $d < rank(\mathbf{A}')$, никакая пара $(\mathbf{H}_\theta, \mathbf{W}_\theta^T)$ не может восстановить истинное распределение логитов.

Предложение 1

Параметр θ, такой что $P_\theta X|c = P^*(X|c)$ для всех c из L существует тогда и только тогда, когда $d \geq \min_{\mathbf{A}' \in F(\mathbf{A})} rank(\mathbf{A}')$.

Следствие об экспрессивности софтмакса

Следствие 1 (Softmax Bottleneck)

Если $d < rank(\mathbf{A}) - 1$ для любого семейства функций U и любого параметра θ, то существует контекст c такой, что $P_\theta(X) \neq P^*(X|c)$

Это означает, что когда размерность d слишком мала, софтмакс не обладает достаточной "экспрессивностью чтобы описать истинное распределение данных. Можно ли легко починить?

- Повысить размерность?
- Использовать Н-грамы?

Гипотеза о естественном языке

Предполагается, что распределение $P^*(X|c)$ для естественного языка представлено высокоранговой матрицей. Почему?

- Следующее слово сильно зависит от контекста
- Если бы матрица была низкоранговой, это значитло бы, что все семантические значения могут быть представлены как взвешенные суммы и отрицания маленького количества *оснований*.
- Эмпирически, продемонстрированная высокоранговая модель показывает себя лучше низкоранговых

Смесь софмаксов (MoS)

$$\boxed{P_\theta(x|c) = \sum_{k=1}^{K} \pi_{c,k} \frac{\exp \mathbf{h}_{c,k}^T \mathbf{w}_x}{\sum_{x'} \exp \mathbf{h}_{c,k}^T \mathbf{w}_x}; \;\; s.t. \sum_{k=1}^{K} \pi_{c,k} = 1}$$

- Улучшенная экспрессивность: MoS аппроксимирует $\mathbf{A}$ при помощи

$$\hat{\mathbf{A}}_{MoS} = \log \sum_{k=1}^{K} \mathbf{\Pi}_k \exp(\mathbf{H}_{\theta,k} \mathbf{W}_\theta^T),$$

  где $\mathbf{\Pi}_k$ это $(N * N)$ дмиагональная матрица с элементами $\pi_{c,k}$.
- $\hat{\mathbf{A}}_{MoS}$ это нелинейная функция от контекстных векторов, а значит она может быть скольки угодно размерной
- Улучшенная генерализация по сравнению с Ngrams и высокоразмерным софтмаксом

Смесь контекстов (MoC)

Смесь контекстов: другой способ взвешивать представления контекстов

$$\boxed{P_\theta(x|c) = \frac{\exp(\sum_k^K \pi_{c,k}\mathbf{h}_{c,k})^T \mathbf{w}_x}{\sum_{x'} \exp(\sum_k^K \pi_{c,k}\mathbf{h}_{c,k})^T \mathbf{w}'_x} = \frac{\exp(\sum_k^K \pi_{c,k}\mathbf{h}_{c,k}^T \mathbf{w}_x)}{\sum_{x'} \exp(\sum_k^K \pi_{c,k}\mathbf{h}_{c,k}^T \mathbf{w}'_x)}}$$

- Несмотря на схожесть с MoS, страдает от того же ограничения, что и обычный софтмакс
- Заметим, что $\mathbf{h}'_c = \sum_{k=1}^K \pi_{c,k}\mathbf{h}_{c,k}$

Эксперименты (Penn Treebank)

- Perplexity — степень "неуверенности"модели

Model	#Param	Validation	Test
Mikolov & Zweig (2012) – RNN-LDA + KN-5 + cache	9M‡	-	92.0
Zaremba et al. (2014) – LSTM	20M	86.2	82.7
Gal & Ghahramani (2016) – Variational LSTM (MC)	20M	-	78.6
Kim et al. (2016) – CharCNN	19M	-	78.9
Merity et al. (2016) – Pointer Sentinel-LSTM	21M	72.4	70.9
Grave et al. (2016) – LSTM + continuous cache pointer†	-	-	72.1
Inan et al. (2016) – Tied Variational LSTM + augmented loss	24M	75.7	73.2
Zilly et al. (2016) – Variational RHN	23M	67.9	65.4
Zoph & Le (2016) – NAS Cell	25M	-	64.0
Melis et al. (2017) – 2-layer skip connection LSTM	24M	60.9	58.3
Merity et al. (2017) – AWD-LSTM w/o finetune	24M	60.7	58.8
Merity et al. (2017) – AWD-LSTM	24M	60.0	57.3
Ours – AWD-LSTM-MoS w/o finetune	22M	58.08	55.97
Ours – AWD-LSTM-MoS	22M	**56.54**	**54.44**
Merity et al. (2017) – AWD-LSTM + continuous cache pointer†	24M	53.9	52.8
Krause et al. (2017) – AWD-LSTM + dynamic evaluation†	24M	51.6	51.1
Ours – AWD-LSTM-MoS + dynamic evaluation†	22M	**48.33**	**47.69**

Table 1: Single model perplexity on validation and test sets on Penn Treebank. Baseline results are obtained from Merity et al. (2017) and Krause et al. (2017). † indicates using dynamic evaluation.

Эксперименты (WikiText-2, Switchboard)

Model	#Param	Validation	Test
Inan et al. (2016) – Variational LSTM + augmented loss	28M	91.5	87.0
Grave et al. (2016) – LSTM + continuous cache pointer†	-	-	68.9
Melis et al. (2017) – 2-layer skip connection LSTM	24M	69.1	65.9
Merity et al. (2017) – AWD-LSTM w/o finetune	33M	69.1	66.0
Merity et al. (2017) – AWD-LSTM	33M	68.6	65.8
Ours – AWD-LSTM-MoS w/o finetune	35M	66.01	63.33
Ours – AWD-LSTM-MoS	35M	**63.88**	**61.45**
Merity et al. (2017) – AWD-LSTM + continuous cache pointer †	33M	53.8	52.0
Krause et al. (2017) – AWD-LSTM + dynamic evaluation†	33M	46.4	44.3
Ours – AWD-LSTM-MoS + dynamical evaluation†	35M	**42.41**	**40.68**

Table 2: Single model perplexity over WikiText-2. Baseline results are obtained from Merity et al. (2017) and Krause et al. (2017). † indicates using dynamic evaluation.

Model	Perplexity	BLEU-1		BLEU-2		BLEU-3		BLEU-4	
		prec	recall	prec	recall	prec	recall	prec	recall
Seq2Seq-Softmax	34.657	0.249	0.188	0.193	0.151	0.168	0.133	0.141	0.111
Seq2Seq-MoC	33.291	0.259	0.198	0.202	0.159	0.176	0.140	0.148	0.117
Seq2Seq-MoS	**32.727**	**0.272**	**0.206**	**0.213**	**0.166**	**0.185**	**0.146**	**0.157**	**0.123**

Table 4: Evaluation scores on Switchboard.

Ablation Study

Model	PTB Validation	PTB Test	WT2 Validation	WT2 Test
AWD-LSTM-MoS	**58.08**	**55.97**	**66.01**	**63.33**
AWD-LSTM-MoC	59.82	57.55	68.76	65.98
AWD-LSTM (Merity et al. (2017) hyper-parameters)	61.49	58.95	68.73	65.40
AWD-LSTM (MoS hyper-parameters)	78.86	74.86	72.73	69.18

Table 5: Ablation study on Penn Treebank and WikiText-2 without finetuning or dynamical evaluation.

- Просто взять параметры MoS не помогает
- MoC все же может быть лучше, чем Softmax
- Можно показать, как MoS более внимательно, чем MoC, использует контекст

Важна ли высокоранговость

- Существуют способы оценки ранга матриц $\hat{\mathbf{A}}_{MoS}, \hat{\mathbf{A}}_{MoC}, \hat{\mathbf{A}}_{Softmax}$
- Увелические компонент в *MoS* влечет увеличение размерности матрицы
- Когда у обычного софтмакса нет проблемы с ограничением ранга, *MoS* не улучшает результат результаты

Model	Validation	Test
Softmax	400	400
MoC	280	280
MoS	**9981**	**9981**

Table 6: Rank comparison on PTB. To ensure comparable model sizes, the embedding sizes of Softmax, MoC and MoS are 400, 280, 280 respectively. The vocabulary size, i.e., M, is 10,000 for all models.

#Softmax	Rank	Perplexity
3	6467	58.62
5	8930	57.36
10	9973	56.33
15	9981	55.97
20	9981	56.17

Table 7: Empirical rank and test perplexity on PTB with different number of Softmaxes.

Важна ли высокоранговость

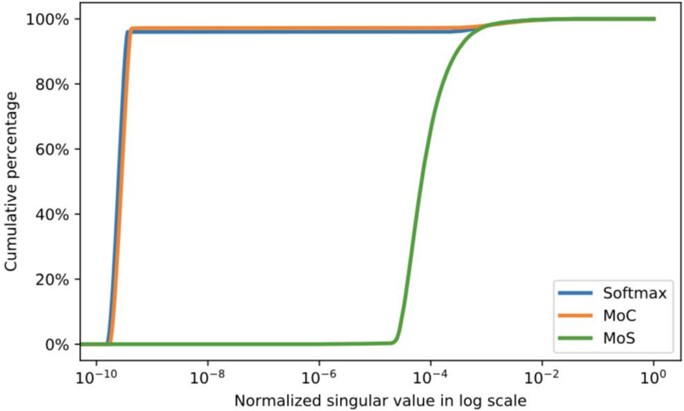

Figure 1: Cumulative percentage of normalized singulars given a value in $[0, 1]$.

Источники

- Yang, Zhilin, et al. "Breaking the softmax bottleneck: A high-rank RNN language model."arXiv preprint arXiv:1711.03953 (2017).

Вопросы

- Опишите суть проблемы Softmax Bottleneck. Отвечая на вопрос, необходимо упомянуть свзь матричных рангов
- Напишите формулу для смеси софтмаксов (MoS)
- В чем различие смеси софтмаксов (MoS) и смеси контекстов (MoC)? Почему смесь контекстов показывает худшее качество?