The Sparsely-Gated Mixture-of-Experts Layer

Мельников Артем

Зачем нам увеличивать размер сетей

Проблемы с которыми сталкиваемся при увеличении размера сети:

При росте сложности задачи резко растет кол-во параметров, которые сеть должна содержать

Что в свою очередь влияет на скорость обучения и скорость применения

Также в какой-то момент мы упираемся в ограничения по мощности вычислительных машин (GPU тоже имеют пределы в вычислительной мощности)

Зачем нам увеличивать размер сетей

Какие преимущества когда у нас больше размер сети:

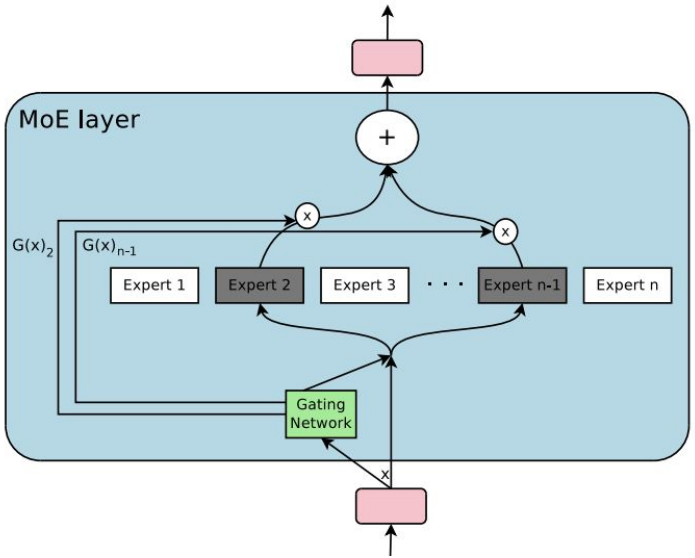
1) Большие батчи - лучше для вычисления тк уменьшают затраты на перемещение данных
2) Современные GPU лучше выполняют арифметику, а не в дроблении данных
3) В задачах распознавания текста/изображений датасеты имеют привычку иметь огромные размеры. Надо, чтобы модель могла поддерживать все тонкие различия/закономерности

Предложенная авторами архитектура

Mixture-of-Experts блок состоит из набора моделей-экспертов и gating нейронной сети.

$$y = \sum_{i=1}^{n} G(x)_i E_i(x)$$

Данные обрабатываются одновременно только несколькими из моделей экспертов.

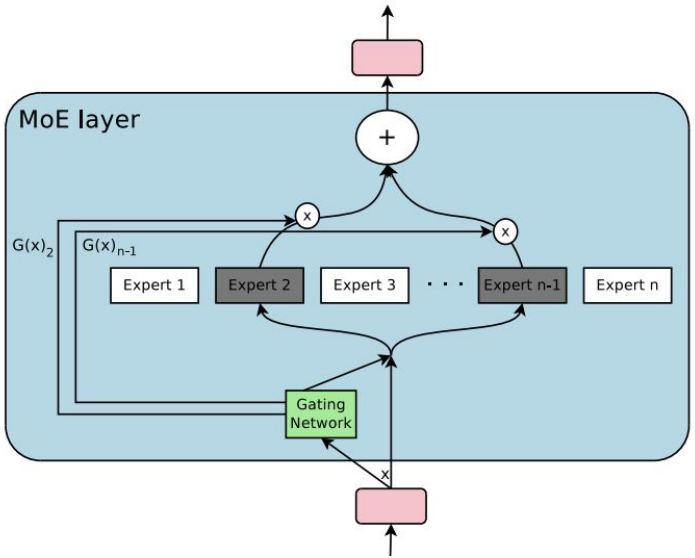


Предложенная авторами архитектура

G(x) - разряженный вектор, так что нам не надо вычислять все значения E_i(x)

$$y = \sum_{i=1}^{n} G(x)_i E_i(x)$$

Обучение всей системы - обычный backpropagation от финального предсказания блока



Предложенная авторами архитектура

$$G(x) = Softmax(KeepTopK(H(x), k))$$

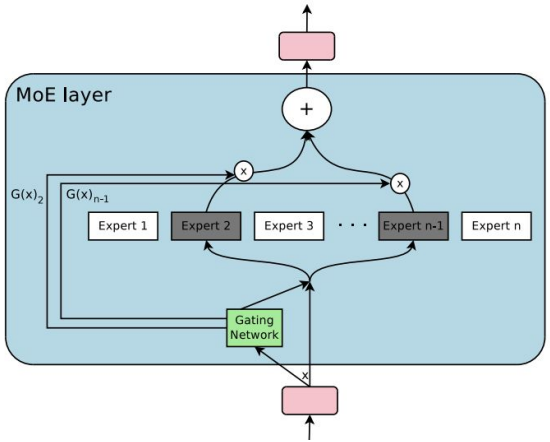
$$H(x)_i = (x \cdot W_g)_i + StandardNormal() \cdot Softplus((x \cdot W_{noise})_i)$$

$$KeepTopK(v, k)_i = \begin{cases} v_i & \text{if } v_i \text{ is in the top } k \text{ elements of } v. \\ -\infty & \text{otherwise.} \end{cases}$$

Gating network - модель определяющая какие k экспертов будут обрабатывать входные данные

k - гиперпараметр

Обучаем обычно с небольшим шумом



Предложенная авторами архитектура

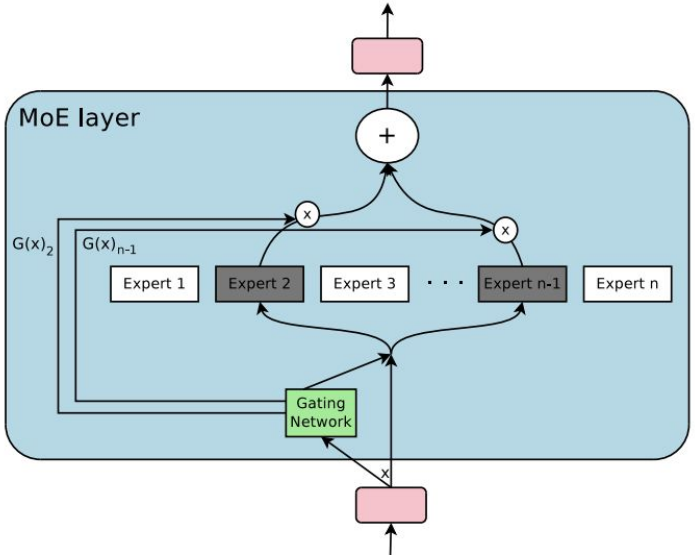
Эксперты - любая модель с определенным входным и выходным размером. Могут быть даже разные архитектуры!

Expert 381	Expert 752	Expert 2004
... with **researchers** , ...	... plays **a** core ...	... with **rapidly** growing ...
... to **innovation** .	... plays **a** critical ...	... under **static** conditions ...
... tics **researchers** .	... provides **a** legislative ...	... to **swift** ly ...
... the **generation** of ...	... play **a** leading ...	... to **dras** tically ...
... technology **innovations** is ...	... assume **a** leadership ...	... the **rapid** and ...
... technological **innovations** , ...	... plays **a** central ...	... the **fast** est ...
... support **innovation** throughout ...	... taken **a** leading ...	... the **Quick** Method ...
... role **innovation** will ...	... established **a** reconciliation ...	... rec **urrent**) ...
... research **scienti** st ...	... played **a** vital ...	... provides **quick** access ...
... promoting **innovation** where ...	... have **a** central ...	... of **volatile** organic ...
...	...	...

Предложенная авторами архитектура

Какие плюсы у блока

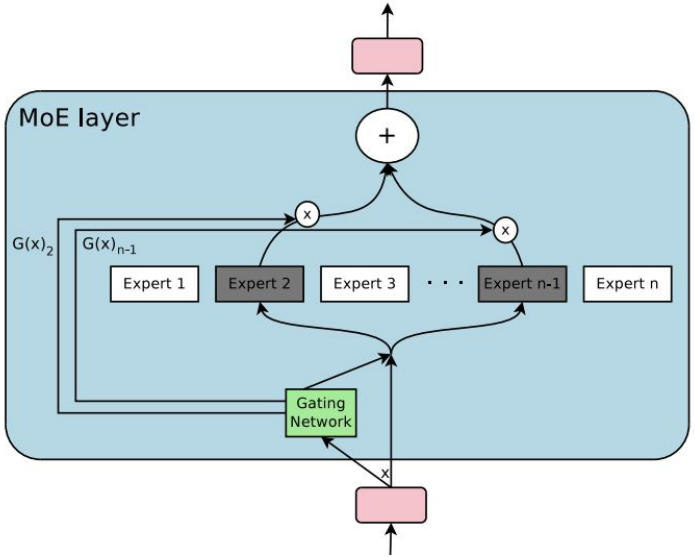
- Много параметров, но при обучении градиент надо считать не по всем тк отключаем большую часть модели
- Сравнительно быстрое применение + возможность распараллелить вычисления



Взвешиваем экспертов

Проблема - наша модель при обучении будет выбирать чаще уже более обученных экспертов, так как они будут давать меньший loss.

Что в свою очередь приведет к неравномерному обучению экспертов (что плохо)



Взвешиваем экспертов

$$Importance(X) = \sum_{x \in X} G(x)$$

$$L_{importance}(X) = w_{importance} \cdot CV(Importance(X))^2$$

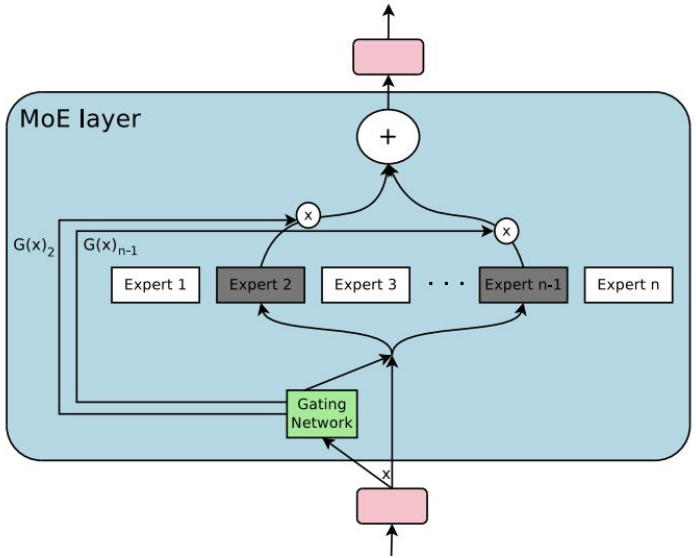
Решение - добавить дополнительный лосс, который будет ограничивать чрезмерное злоупотребление одним из экспертов

w_importance - гиперпараметр

Тонкости реализации на железе

Проблема - каждый эксперт получает только часть входных данных, как следствие размер батча для каждой модели-эксперта значительно уменьшается. Что сказывается на эффективности вычисления на GPU

$$\frac{kb}{n} \ll b$$



k - кол-во используемых экспертов
n - всего экспертов
b - размер батча

Тонкости реализации на железе

Возможные решения:

- Сначала запускаем параллельно на нескольких батчах gating сетку, затем каждому эксперту подаем релевантные данные

  Итого улучшение в размере батча в кол-во потоков раз

Тонкости реализации на железе

Возможные решения:

- Сначала запускаем параллельно на нескольких батчах gating сетку, затем каждому эксперту подаем релевантные данные

  Итого улучшение в размере батча в кол-во потоков раз

- Использовать трюки для рекурентных сетей (ломает предыдущий трюк)

Тонкости реализации на железе

Возможные решения:

- Сначала запускаем параллельно на нескольких батчах gating сетку, затем каждому эксперту подаем релевантные данные

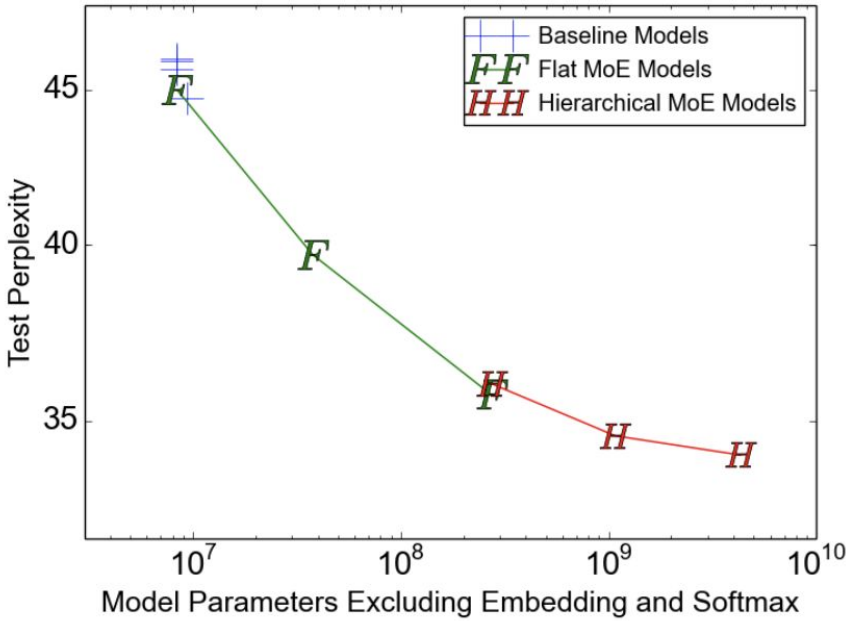
  Итого улучшение в размере батча в кол-во потоков раз

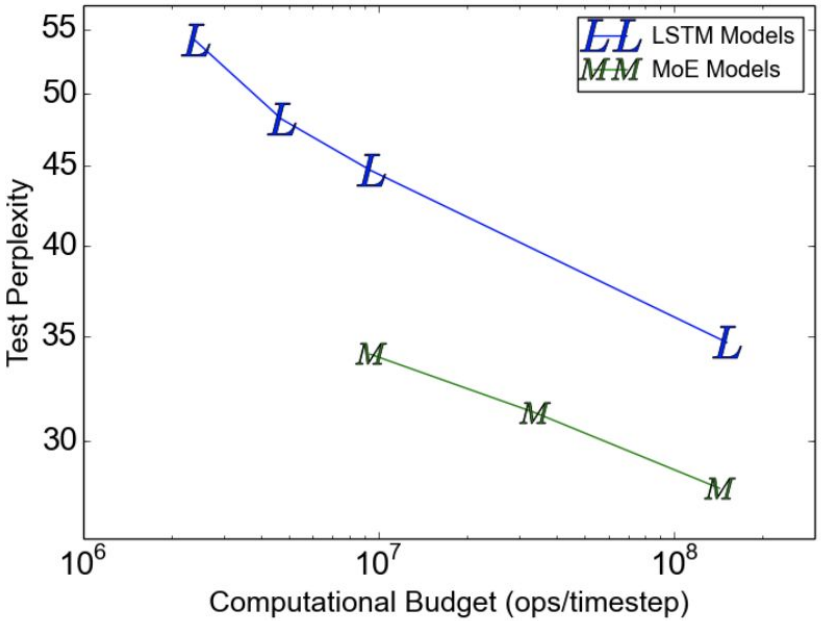
- Использовать трюки для рекурентных сетей (ломает предыдущий трюк)

- Дождаться выполнения свертки на нескольких батчах и подать их экспертам параллельно

Флекс показателями сети

Уничтожаем обычные модели





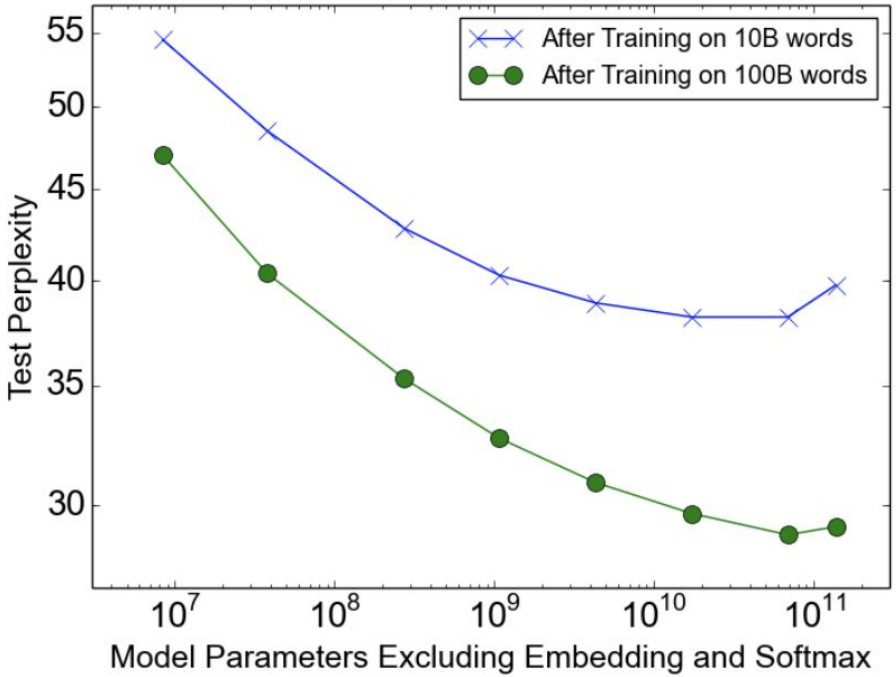
Флекс показателями сети

Гугловский корпус слов
100 миллиардов данных

Больше данных - хорошо
(на 39% лучше, если быть точнее)

65536 экспертов, 0.72 TFLOPS/GPU
(все еще эффективное вычисление)

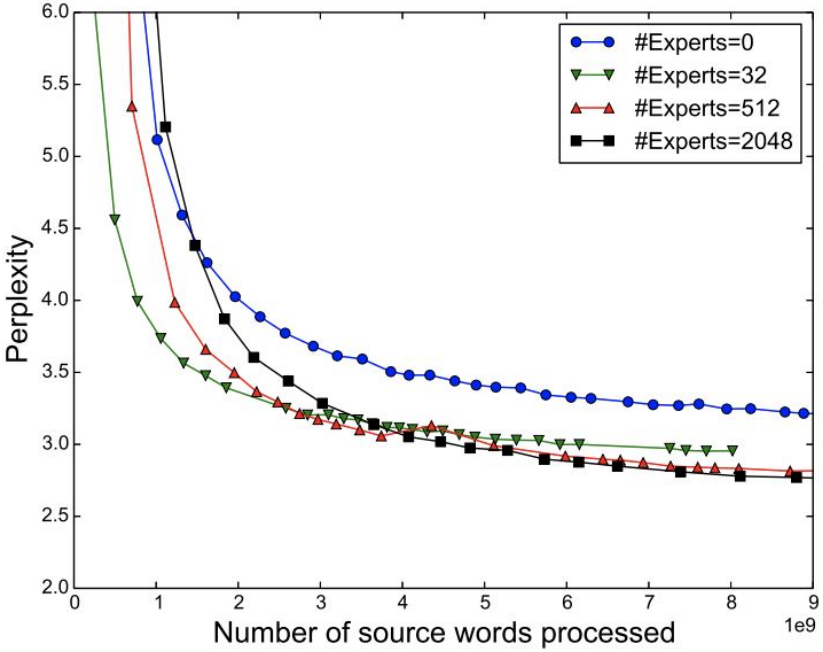
131072 экспертов - слишком
разряженное пространство

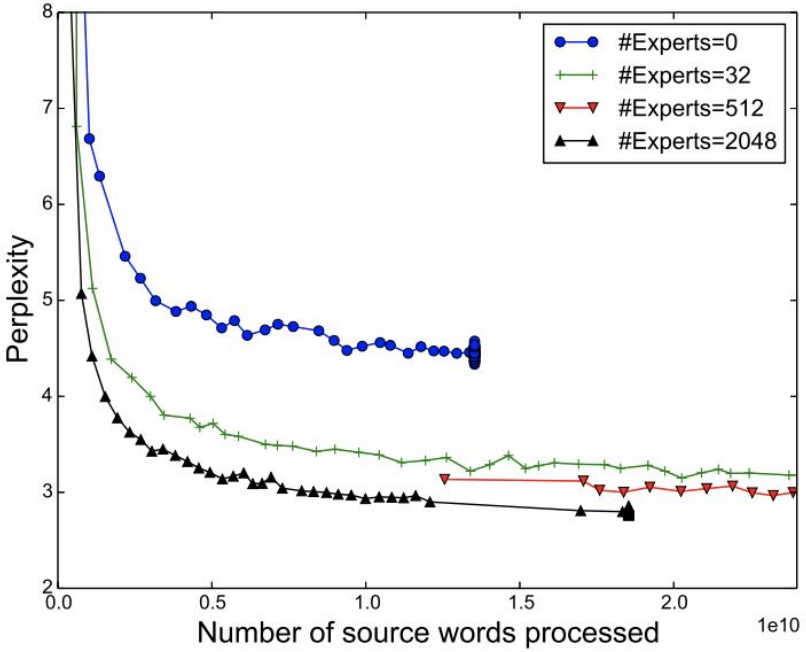


Флекс показателями сети

	GNMT-Mono	GNMT-Multi	MoE-Multi	MoE-Multi vs. GNMT-Multi
Parameters	278M / model	278M	8.7B	
ops/timestep	212M	212M	102M	
training time, hardware	various	21 days, 96 k20s	**12 days, 64 k40s**	
Perplexity (dev)		4.14	**3.35**	-19%
French → English Test BLEU	36.47	34.40	**37.46**	+3.06
German → English Test BLEU	31.77	31.17	**34.80**	+3.63
Japanese → English Test BLEU	23.41	21.62	**25.91**	+4.29
Korean → English Test BLEU	25.42	22.87	**28.71**	+5.84
Portuguese → English Test BLEU	44.40	42.53	**46.13**	+3.60
Spanish → English Test BLEU	38.00	36.04	**39.39**	+3.35
English → French Test BLEU	35.37	34.00	**36.59**	+2.59
English → German Test BLEU	**26.43**	23.15	24.53	+1.38
English → Japanese Test BLEU	**23.66**	21.10	22.78	+1.68
English → Korean Test BLEU	**19.75**	18.41	16.62	-1.79
English → Portuguese Test BLEU	**38.40**	37.35	37.90	+0.55
English → Spanish Test BLEU	34.50	34.25	**36.21**	+1.96

Флекс показателями сети





Собственно статья

- https://arxiv.org/abs/1701.06538