Пробинг: локализация грамматики в нейронных сетях

Научный контекст

Корнилов Альберт, Степанова Ангелина, Сухарева Мария, Шумакова Лада

Пробинг: основное

! модель не дообучается: работа происходит с существующей моделью

3 парадигмы пробинга [Lasri et. al 2022]:

- Диагностический пробинг
- Поведенческий пробинг
- Казуальный пробинг

Диагностический пробинг

Обучение классификаторов (как правило, линейных) на получаемых моделью эмбеддингах; оценка качества по accuracy

- не все данные могут быть закодированы линейно → нельзя отследить при помощи простых классификаторов [Belinkov et al. 2017; Conneau et al. 2018]
- Использование мощного классификатора: ? а информация точно не выучилась из пробинг-классификатора ? [Hewitt, Liang 2019] → control-task: перемешанные теги, если классификатор хорош все равно, ничего нельзя сказать о модели

Поведенческий пробинг

Прямое наблюдение за поведением модели; даем специальный датасет → смотрим, что модель делает

+ Не надо использовать дополнительный классификатор → мы уверены, что модель работает как она есть
+ Достоверные сведения о «знаниях» модели
- Ничего нельзя сказать о механизме этого знания (! важно для локализации!)

Причинно-следственный пробинг

Оценка того, как конкретные компоненты влияют на качество модели.

Ограничения пробинга

- Методы и датасеты ориентированы в основном на английский язык → не универсальны, для того же русского необходимо учитывать эти особенности [Eger at al. 2020]
- Количество данных, необходимых для достоверного эксперимента [Belinkov, Glass 2019; Eger at al. 2020; Zhu et al. 2022]
- Модель может содержать информацию, но при этом не использовать ее [Vanmassenhove et al. 2018]

Разреженные автоэнкодеры

Автоэнкодеры – это тип нейронных сетей, предназначенный для обучения без учителя, в первую очередь для уменьшения размерности или выделения признаков.

Они состоят из энкодера, который сжимает входные данные в представление меньшей размерности (скрытое пространство), и декодера.

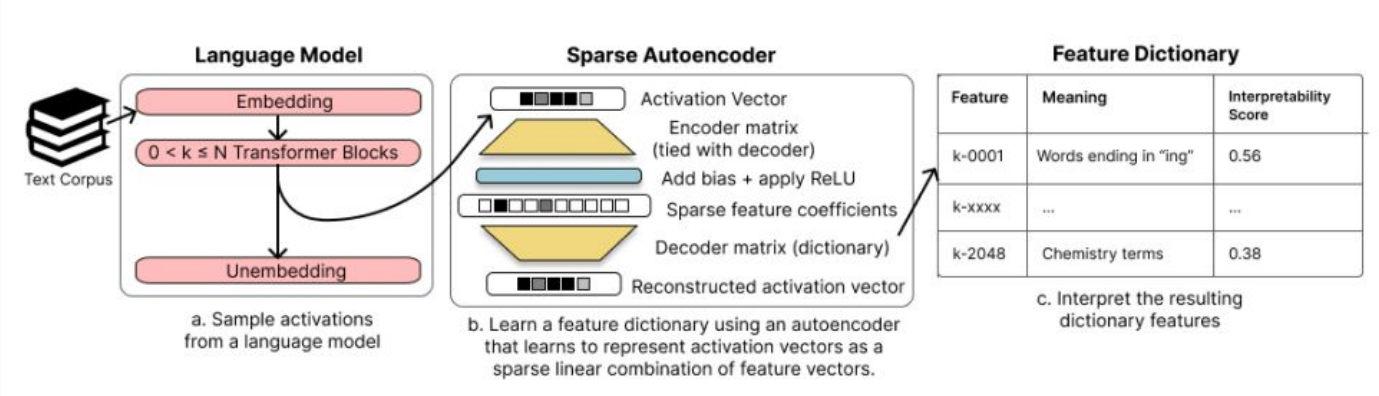
Разреженные автоэнкодеры включают ограничение разреженности во время обучения, которое побуждает модель изучать представления, которые являются более эффективными, гарантируя, что только небольшое количество нейронов активно в любой момент времени (например, **L1** или **KL**).

Разреженные автоэнкодеры

Sparse Autoencoders Find Highly Interpretable Features in Language Models [Cunningham et al. 2023]:

1. Отдельные нейроны активируются на множестве различных семантических признаков из-за **суперпозиции**
2. Результат - словарь редко активирующихся признаков
3. С помощью **activation patching**, демонстрируется, что полученные признаки могут существенно изменить выходные данные модели с меньшим количеством вмешательств, чем альтернативные методы

Разреженные автоэнкодеры



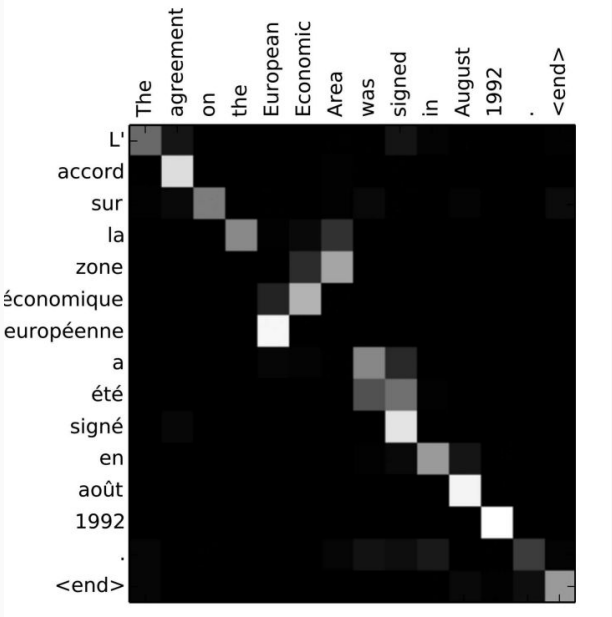
1. Отбираются внутренние активации языковой модели (residual stream, MLP, attention head)
2. Эти активации используются для обучения разреженного автоэнкодера, веса которого формируют словарь признаков
3. Результирующие признаки интерпретируются с помощью таких методов, как оценка автоинтерпретируемости OpenAI

Разреженные автоэнкодеры

Применение словарей признаков:

1. **Input:** какие токены активируют признак из словаря и в каких контекстах
2. **Output:** как удаление признака изменяет выходные логиты модели
3. **Intermediate features:** какие признаки на предыдущих слоях приводят к активации анализируемого признака

Еще методы: визуализируем внутрь [Bahdanau et al. 2014]

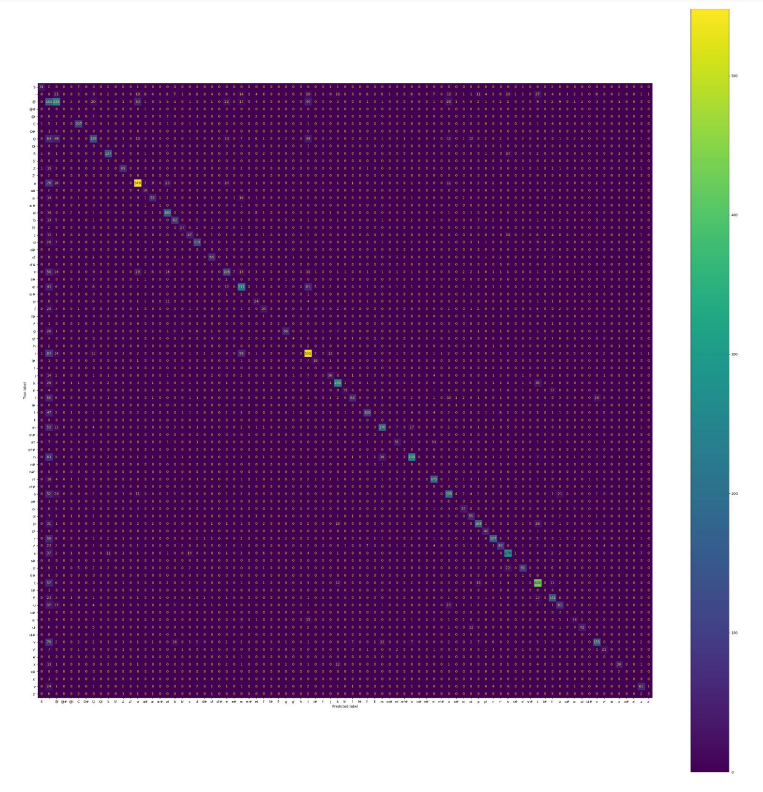


Визуализация весов слов между предложением и его переводом: соответствие.

Все понятно, но:

- А мы можем полагаться на свои глаза as it were?
- Как оценивать? Есть ли более объективные показатели?

Не совсем про нейросети, но...



Теория информации

1. A Bayesian Framework for Information-Theoretic Probing [Pimentel, Cotterell 2021]

- пробинг как аппроксимация Mutual Information
- Bayesian Mutual Information
- идея: работать не с косинусом и линейно-алгебраическими преобразованиями, а с теоретико-информационными функционалами в латентном пространстве, задавая их для различающихся токенов

Bayesian Mutual Information

Взаимная информация (Mutual Information):
$I(X, Y) = H(X) - H(X|Y)$.
(Claude E. Shannon, "A Mathematical Theory of Communication", 1948)

Свойства теории информации, которые не соотносятся с реальностью ML

1) новые данные не добавляют информацию: $H(X \mid D_N = d_N) = H(X)$
   (т. к. предположение, что они из того же распределения)
2) условное распределение уменьшает энтропию: $H(X \mid Y) \leq H(X)$
3) процессинг данных не увеличивает информативность:
   $I(X; f(Y)) \leq I(X; Y)$

Bayesian Mutual Information

Решение: байесовская взаимная информация

~~$I(X, Y) = H(X) - H(X|Y)$~~

$I_\theta(Y \to X \mid d_N) = H_\theta(X \mid d_N) - H_\theta(X \mid Y, d_N)$

главное отличие от исходной взаимной информации:

реальное распределение ≠ распределение модели

Bayesian Mutual Information

Решение: байесовская взаимная информация

~~$I(X, Y) = H(X) - H(X|Y)$~~

$I_\theta(Y \to X \mid d_N) = H_\theta(X \mid d_N) - H_\theta(X \mid Y, d_N)$

1) новые данные <u>добавляют</u> информацию: ~~$H(X \mid D_N = d_N) = H(X)$~~
2) условное распределение уменьшает энтропию: ~~$H(X \mid Y) \leq H(X)$~~
3) процессинг данных может увеличивать информативность:

   ~~$I(X; f(Y)) \leq I(X; Y)$~~

   $I_\theta(f(Y) \to X \mid d_N) \; ? \; I_\theta(Y \to X \mid d_N)$

“data can add information, processing can help, and information can hurt”

Теория информации

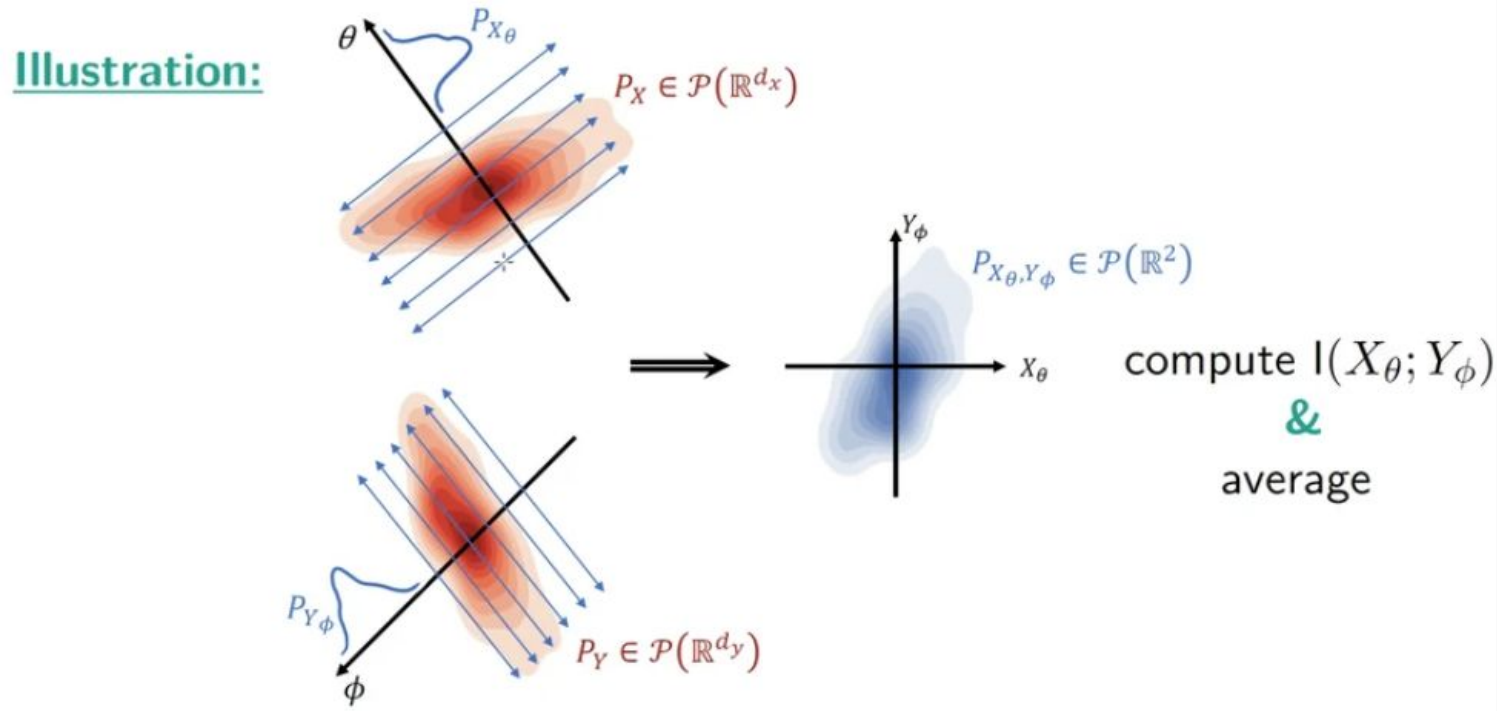
2. Sliced Mutual Information: A Scalable Measure of Statistical Dependence [Goldfeld, Greenewald 2021]

Definition (ZG-Greenewald'21)

The SMI between $(X, Y) \sim P_{X,Y} \in \mathcal{P}(\mathbb{R}^{d_x} \times \mathbb{R}^{d_y})$ is

$$\mathsf{SI}(X;Y) := \frac{1}{S_{d_x-1} S_{d_y-1}} \oint_{\mathbb{S}^{d_x-1}} \oint_{\mathbb{S}^{d_y-1}} \mathsf{I}(\theta^\mathsf{T} X; \phi^\mathsf{T} Y)\, \mathrm{d}\theta\, \mathrm{d}\phi.$$

Sliced Mutual Information



Sliced Mutual Information

Сохраняет многие свойства Mutual Information

Например:

- независимость: $SI(X; Y) = 0 \Leftrightarrow X$ и Y независимы
- определение через энтропию: $SI(X; Y) = SH(X) - SH(X|Y)$

Отличия от Mutual Information:

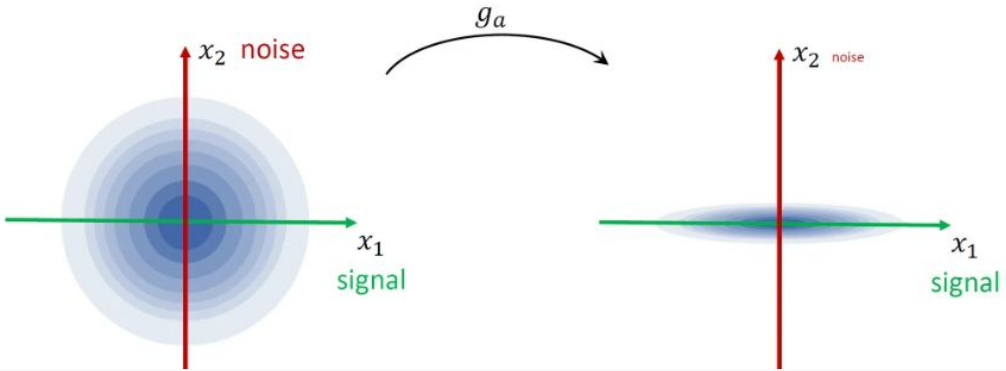
Sliced Mutual Information

processing can help! (как в Bayesian Mutual Information)

если у f(X) более информативные проекции, чем у самой X:

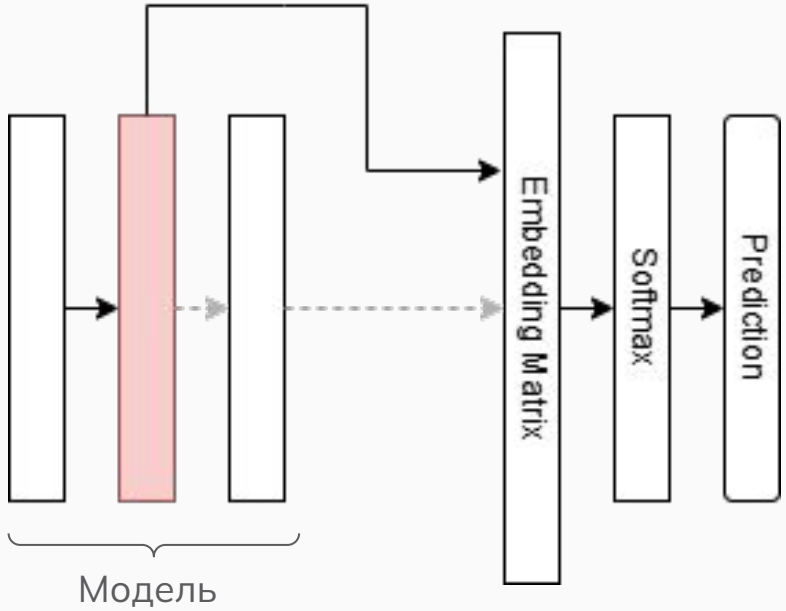
SI(X; Y) < SI(f(X); Y)

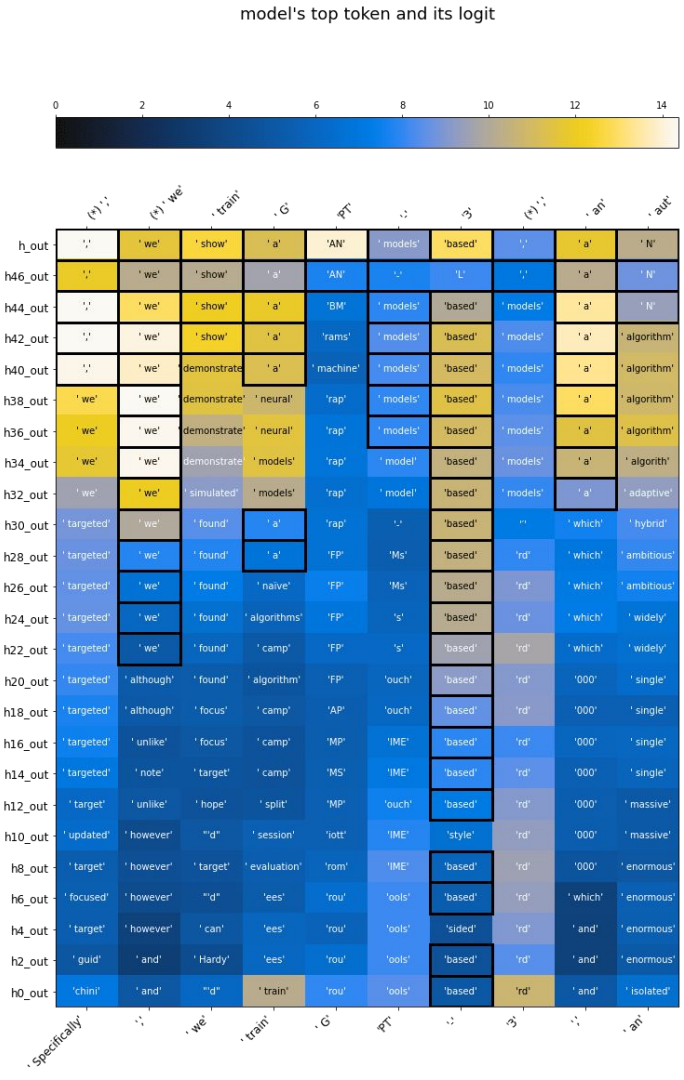
- **Example:** $X = (X_1\ X_2) \sim \mathcal{N}(0, \mathrm{I}_2)$, $Y = X_1$, $g_a(x_1, x_2) = [x_1\ ax_2]^\intercal$



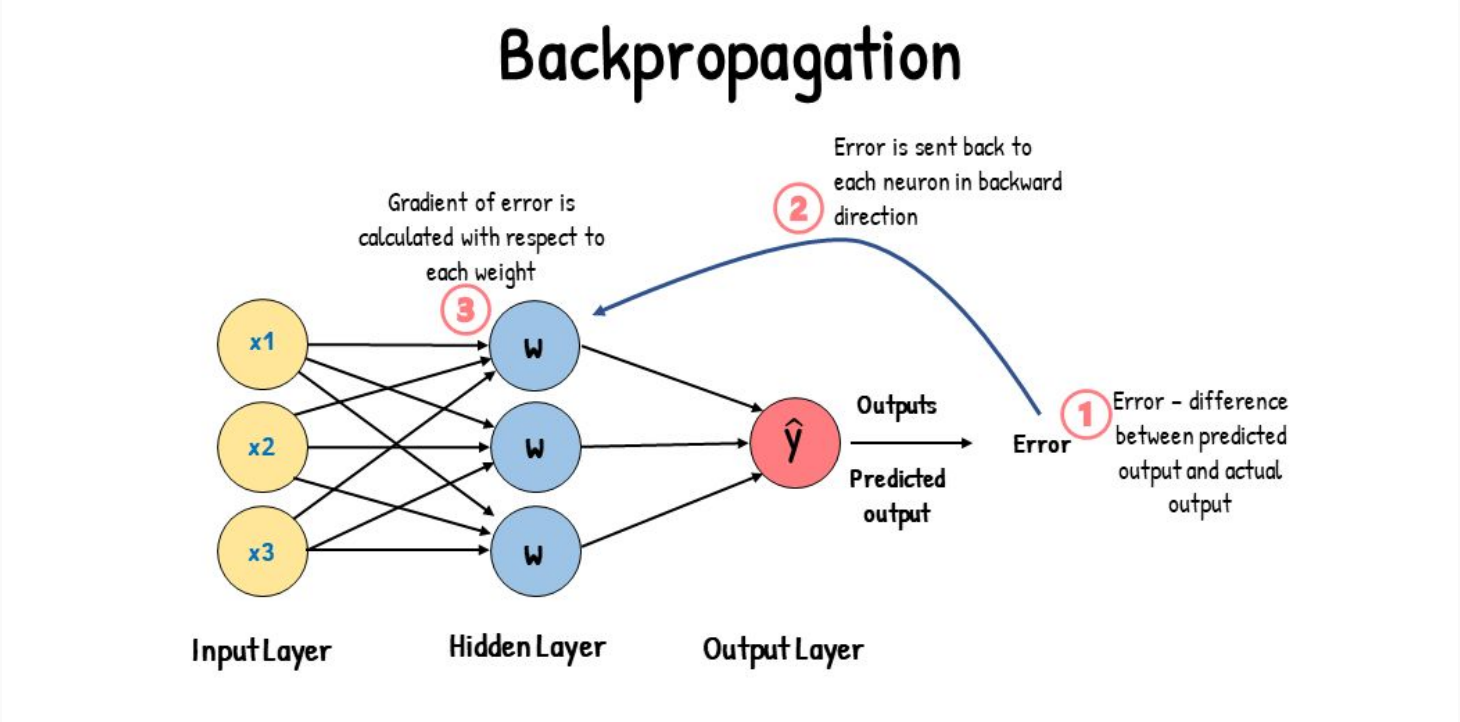
Интерпретация моделей в пространстве градиентов

- Преобразование скрытых состояний LMs в вероятности словаря через Logit Lens (LL) [nostalgebraist 2020]
- Проекция нейронов в токены с использованием LL [Geva et al. 2021; Geva et al. 2022].
- Анализ матриц внимания и их влияние на предсказания [Dar et al. 2022; Geva et al. 2023]

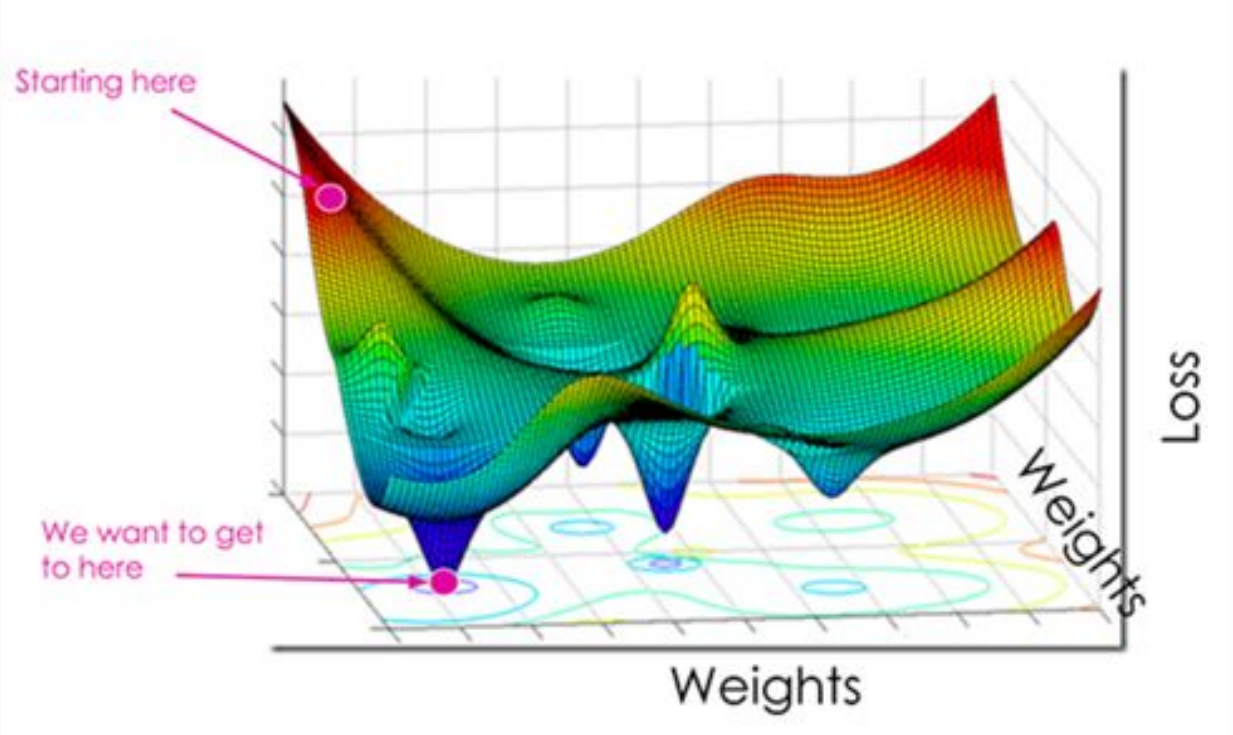



model's top token and its logit
0
2
4
6
8
10
12
14
(+) ',' (+) ' we' ' train' ' G' 'PT' ':' '3' (+) ',' ' an' ' aut'
h_out ',' ' we' ' show' ' a' 'AN' ' models' 'based' ',' ' a' ' N'
h46_out ',' ' we' ' show' ' a' 'AN' '-' 'L' ',' ' a' ' N'
h44_out ',' ' we' ' show' ' a' 'BM' ' models' 'based' ' models' ' a' ' N'
h42_out ',' ' we' ' show' ' a' 'rams' ' models' 'based' ' models' ' a' ' algorithm'
h40_out ',' ' we' ' demonstrate' ' a' ' machine' ' models' 'based' ' models' ' a' ' algorithm'
h38_out ' we' ' we' ' demonstrate' ' neural' 'rap' ' models' 'based' ' models' ' a' ' algorithm'
h36_out ' we' ' we' ' demonstrate' ' neural' 'rap' ' models' 'based' ' models' ' a' ' algorithm'
h34_out ' we' ' we' ' demonstrate' ' models' 'rap' ' model' 'based' ' models' ' a' ' algorith'
h32_out ' we' ' we' ' simulated' ' models' 'rap' ' model' 'based' ' models' ' a' ' adaptive'
h30_out ' targeted' ' we' ' found' ' a' 'rap' '-' 'based' '''' ' which' ' hybrid'
h28_out ' targeted' ' we' ' found' ' a' 'FP' 'Ms' 'based' 'rd' ' which' ' ambitious'
h26_out ' targeted' ' we' ' found' ' naive' 'FP' 'Ms' 'based' 'rd' ' which' ' ambitious'
h24_out ' targeted' ' we' ' found' ' algorithms' 'FP' 's' 'based' 'rd' ' which' ' widely'
h22_out ' targeted' ' we' ' found' ' camp' 'FP' 's' 'based' 'rd' ' which' ' widely'
h20_out ' targeted' ' although' ' found' ' algorithm' 'FP' 'ouch' 'based' 'rd' '000' ' single'
h18_out ' targeted' ' although' ' focus' ' camp' 'AP' 'ouch' 'based' 'rd' '000' ' single'
h16_out ' targeted' ' unlike' ' focus' ' camp' 'MP' 'IME' 'based' 'rd' '000' ' single'
h14_out ' targeted' ' note' ' target' ' camp' 'MS' 'IME' 'based' 'rd' '000' ' single'
h12_out ' target' ' unlike' ' hope' ' split' 'MP' 'ouch' 'based' 'rd' '000' ' massive'
h10_out ' updated' ' however' '"d' ' session' 'iott' 'IME' 'style' 'rd' '000' ' massive'
h8_out ' target' ' however' ' target' ' evaluation' 'rom' 'IME' 'based' 'rd' '000' ' enormous'
h6_out ' focused' ' however' '"d' 'ees' 'rou' 'ools' 'based' 'rd' ' which' ' enormous'
h4_out ' target' ' however' ' can' 'ees' 'rou' 'ools' 'sided' 'rd' ' and' ' enormous'
h2_out ' guid' ' and' ' Hardy' 'ees' 'rou' 'ools' 'based' 'rd' ' and' ' enormous'
h0_out 'chini' ' and' '"d' ' train' 'rou' 'ools' 'based' 'rd' ' and' ' isolated'
' Specifically' ',' ' we' ' train' ' G' 'PT' ':' '3' ',' ' an'

Backpropagation



Backpropagation



Backpropagation

Чтобы минимизировать ошибку, считаем градиент функции потерь относительно весов
Но это сложная функция! Дифференцируем по цепному правилу

$$\frac{\partial L}{\partial W} = \frac{\partial z}{\partial W}\frac{\partial L}{\partial z}$$

Интерпретация backpropagation

Katz et al. 2024

Магия математики (если слой линейный):

- частную производную z по весам можем подсчитать сразу;
- оставшаяся часть – Vector-Jacobian Product (VJP), о котором можно думать, как о скрытом представлении ошибки

$$\frac{\partial z}{\partial W} = \frac{\partial xW}{\partial W} = x^\top$$

$$\delta = \frac{\partial L}{\partial z} \in \mathbb{R}^n$$

$$\frac{\partial L}{\partial W} = \frac{\partial z}{\partial W}\frac{\partial L}{\partial z} = x^\top \cdot \delta \in \mathbb{R}^{n \times m}$$

Интерпретация backpropagation

С помощью такого разложения можно доказать, что

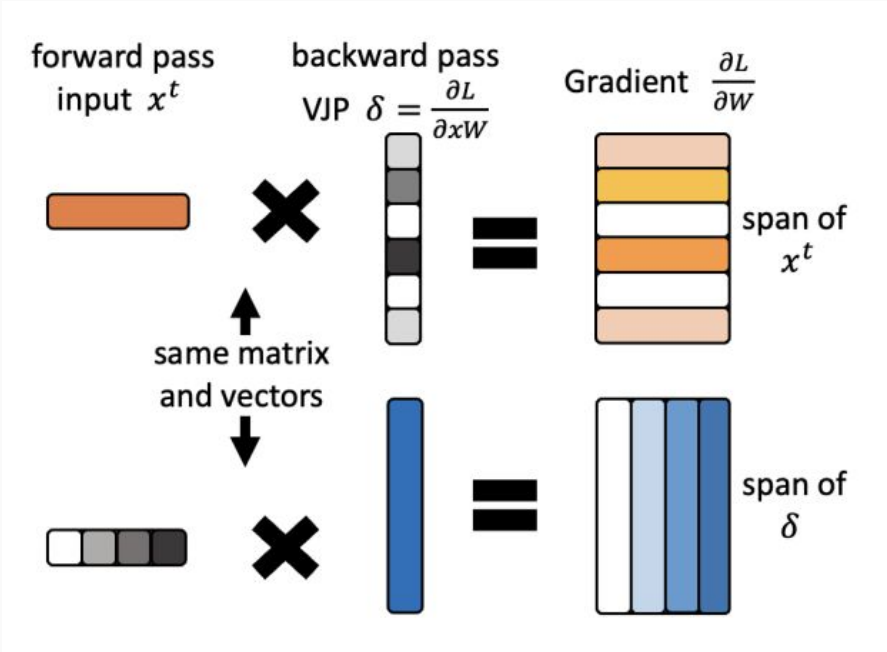
Lemma 4.1. *Given a sequence of inputs of length n, a parametric matrix W and a loss function L, the gradient $\frac{\partial L}{\partial W}$ produced by a backward pass is a matrix with a rank of n or lower.*

$$\frac{\partial L}{\partial W} = \sum_{i=1}^{n} \frac{\partial z_i}{\partial W} \frac{\partial L}{\partial z_i} = \sum_{i=1}^{n} x_i^\top \cdot \delta_i$$

Интерпретация backpropagation

Теперь мы можем смотреть на матрицу градиента как на

- линейную комбинацию векторов x_i
- линейную комбинацию векторов $\square_i$



Список литературы

Lasri et. al 2022 — K. Lasri, T. Pimentel, A. Lenci, T. Poibeau, R. Cotterell. Probing for the usage of grammatical number // arXiv preprint arXiv:2204.08831, 2022.

Belinkov et al. 2017 — Y. Belinkov, N. Durrani, F. Dalvi, H. Sajjad, J. Glass. What do neural machine translation models learn about morphology? // Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Vancouver, Canada, July 30 - August 4, 2017. P. 861-872.

Conneau et al. 2018 — A. Conneau, G. Kruszewski, G. Lample, L. Barrault, M. Baroni. What you can cram into a single $&!#* vector: Probing sentence embeddings for linguistic properties // The 56th Annual Meeting of the Association for Computational Linguistics, Melbourne, Australia, July 15-20, 2018. P. 2126–2136.

Hewitt, Liang 2019 — J. Hewitt, P. Liang. Designing and interpreting probes with control tasks // arXiv preprint arXiv:1909.03368, 2019.

Список литературы

Eger at al. 2020 — S. Eger, J. Daxenberger, I. Gurevych. How to probe sentence embeddings in low-resource languages: On structural design choices for probing task evaluation // arXiv preprint arXiv:2006.09109, 2020.

Belinkov, Glass 2019 — Y. Belinkov, J. Glass. Analysis Methods in Neural Language Processing: A Survey // arXiv preprint arXiv:1812.08951, 2019.

Zhu et al. 2022 — Z. Zhu, J. Wang, B. Li, F. Rudzicz. On the data requirements of probing // arXiv preprint arXiv:2202.12801, 2022.

Vanmassenhove et al. 2018 — E. Vanmassenhove, C. Hardmeier, A. Way. Getting Gender Right in Neural Machine Translation // Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. Brussels, Belgium, 2018. P. 3003–3008.

Список литературы

Cunningham et al. 2023 — H. Cunningham, A. Ewart, L. Riggs, R. Huben, L. Sharkey. Sparse Autoencoders Find Highly Interpretable Features in Language Models // arXiv preprint arXiv:2309.08600, 2023.

Bahdanau et al. 2014 — D. Bahdanau, K. Cho, Y. Bengio. Neural Machine Translation by Jointly Learning to Align and Translate // arXiv preprint arXiv:1409.0473, 2014.

nostalgebraist 2020 — nostalgebraist. interpreting GPT: the logit lens // URL: https://www.lesswrong.com/posts/AcKRB8wDpdaN6v6ru/interpreting-gpt-the-logit-lens. 2020.

Geva et al. 2021 — M. Geva, R. Schuster, J. Berant, O. Levy. Transformer Feed-Forward Layers Are Key-Value Memories // Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Punta Cana, 2021. P. 5484–5495.

Список литературы

Geva et al. 2022 — M. Geva, A. Caciularu, K. Ro Wang, Y. Goldberg. Transformer Feed-Forward Layers Build Predictions by Promoting Concepts in the Vocabulary Space. // arXiv preprint arXiv:2203.14680, 2022.

Dar et al. 2022 — G. Dar, // M. Geva, A. Gupta, Jonathan Berant. Analyzing Transformers in Embedding Space // arXiv preprint arXiv:2209.02535, 2022.

Geva et al. 2023 — M. Geva, J. Bastings, K. Filippova, A. Globerson. Dissecting Recall of Factual Associations in Auto-Regressive Language Models // arXiv preprint arXiv:2304.14767, 2023.

Pimentel, Cotterell 2021 — T. Pimentel, R. Cotterell. A Bayesian Framework for Information-Theoretic Probing // arXiv preprint arXiv:2109.03853, 2021.

Список литературы

Goldfeld, Greenewald 2021 — Z. Goldfeld, K. Greenewald. Sliced Mutual Information: A Scalable Measure of Statistical Dependence. // arXiv preprint arXiv:2110.05279, 2021.

Claude E. Shannon, A Mathematical Theory of Communication // Reprinted with corrections from The Bell System Technical Journal. Vol. 27. P. 379–423, 623–656. 1948

Место для ваших вопросов

Спасибо за внимание!