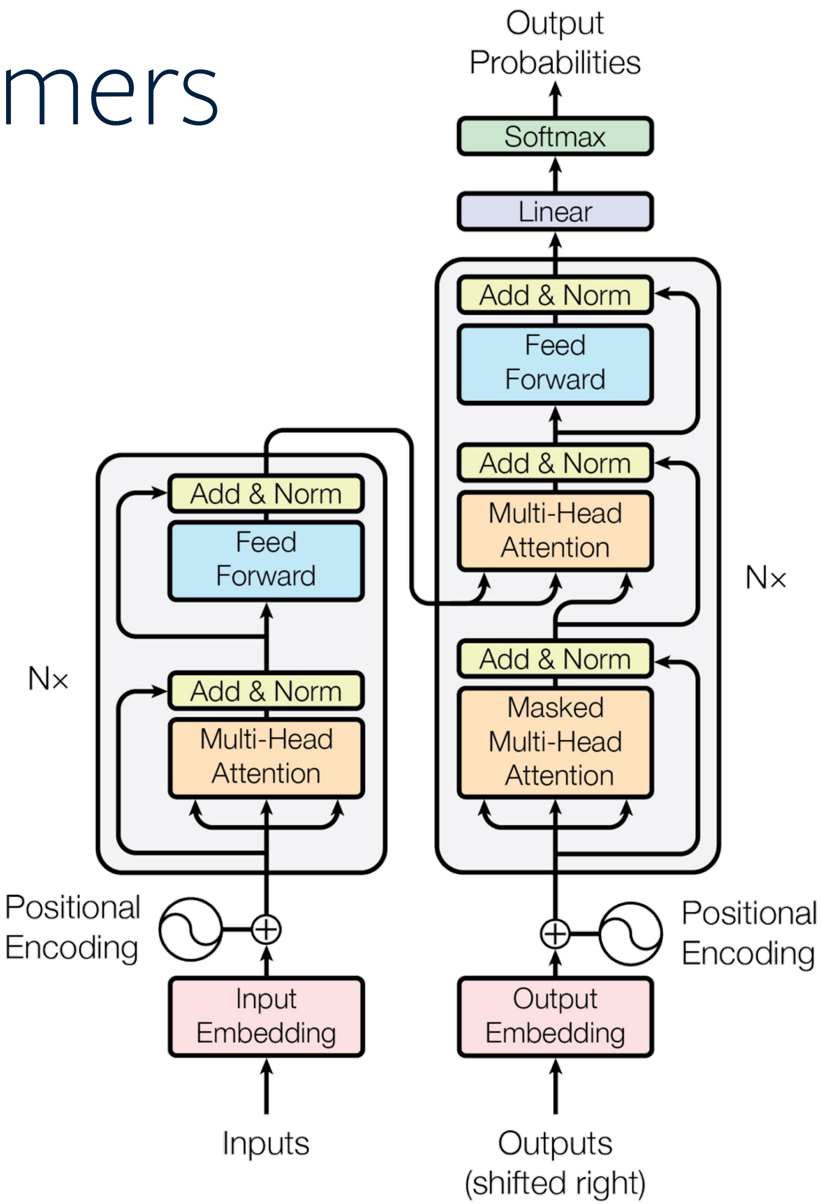
Scan and Snap: Understanding Training Dynamics and Token Composition in 1-layer Transformer

Yuandong Tian[1] Yiping Wang[2,4] Beidi Chen[1,3] Simon Du[2]

[1]Meta AI (FAIR) [2]University of Washington [3]Carnegie Mellon University [4]Zhejiang University

facebook Artificial Intelligence

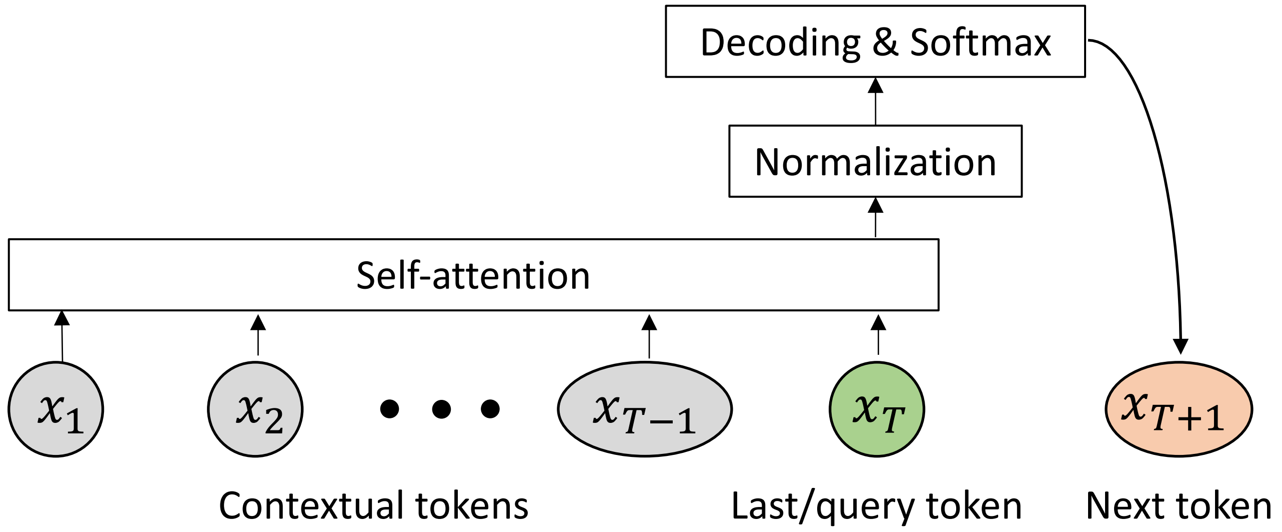
Transformers



Why it works?

[A. Vaswani et al, Attention is all you need, NeurIPS'17]

Problem Setting



$U = [\boldsymbol{u}_1, \boldsymbol{u}_2, \ldots \boldsymbol{u}_M]^T$: token embedding matrix

Self-attention

$$\hat{\boldsymbol{u}}_T = \sum_{t=1}^{T-1} b_{tT} \boldsymbol{u}_{x_t} = U^T X^T \boldsymbol{b}_T$$

$$b_{tT} := \frac{\exp(\boldsymbol{u}_{x_T}^\top W_Q W_K^\top \boldsymbol{u}_{x_t} / \sqrt{d})}{\sum_{t=1}^{T-1} \exp(\boldsymbol{u}_{x_T}^\top W_Q W_K^\top \boldsymbol{u}_{x_t} / \sqrt{d})}$$

Normalized version $\tilde{\boldsymbol{u}}_T = U^T \mathrm{LN}(X^T \boldsymbol{b}_T)$

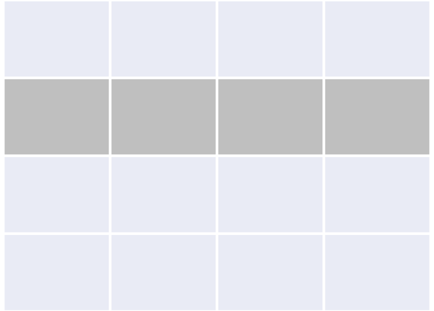
<u>Objective:</u>

$$\max_{W_K, W_Q, W_V, U} J = \mathbb{E}_D \left[\boldsymbol{u}_{x_{T+1}}^T W_V \tilde{\boldsymbol{u}}_T - \log \sum_l \exp(\boldsymbol{u}_l^T W_V \tilde{\boldsymbol{u}}_T) \right]$$

Reparameterization

- Parameters W_K, W_Q, W_V, U makes the dynamics complicated.
- Reparameterize the problem with independent variable Y and Z
  - $Y = UW_V^T U^T$
  - $Z = UW_Q W_K^T U^T$ (pairwise logits of self-attention matrix)
- Then the dynamics becomes easier to analyze

Training dynamics of Y and Z



$\boldsymbol{z}_m$: All logits of the contextual tokens when attending to last token $x_T = m$

Training Dynamics:

$$\dot{Y} = \eta_Y \mathrm{LN}(X^T \boldsymbol{b}_T)(\boldsymbol{x}_{T+1} - \boldsymbol{\alpha})^T$$

$$\dot{Z} = \eta_Z \boldsymbol{x}_T (\boldsymbol{x}_{T+1} - \boldsymbol{\alpha})^T Y^T \frac{P^{\perp}_{X^T \boldsymbol{b}_T}}{\|X^T \boldsymbol{b}_T\|_2} X^T \mathrm{diag}(\boldsymbol{b}_T) X$$

Here $Z = [\boldsymbol{z}_1, \boldsymbol{z}_2, \ldots, \boldsymbol{z}_M]^T$, each $\boldsymbol{z}_m \in \mathbb{R}^M$ is the attention score for query/last token m:

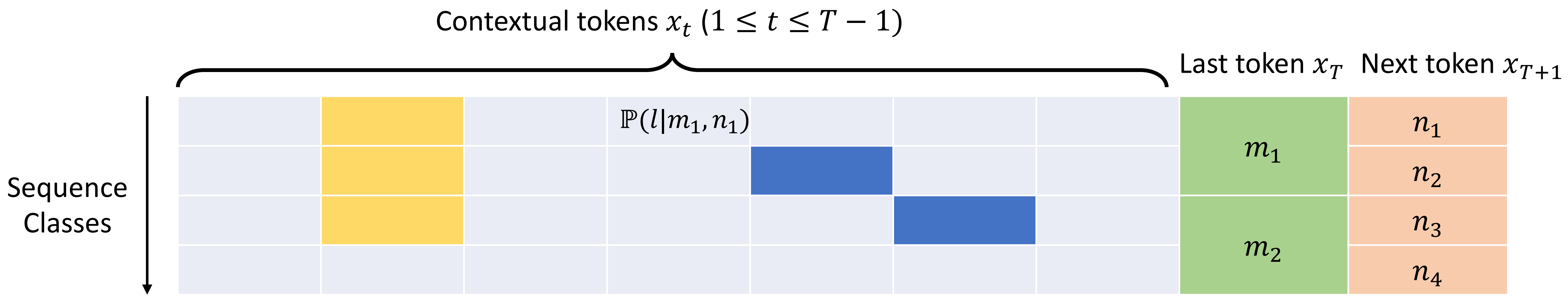
$$\dot{\boldsymbol{z}}_m = \eta_Z X^\top[i] \mathrm{diag}(\boldsymbol{b}_T[i]) X[i] \frac{P^{\perp}_{X^\top[i]\boldsymbol{b}_T[i]}}{\|X^\top[i]\boldsymbol{b}_T[i]\|_2} Y(\boldsymbol{x}_{T+1}[i] - \boldsymbol{\alpha}[i])$$

Major Assumptions

- No positional encoding
- Sequence length $T \rightarrow +\infty$
- Learning rate of decoder Y larger than self-attention layer Z ($\eta_Y \gg \eta_Z$)
- Other technical assumptions

Data Distribution

$x_t \in [M]$ for $1 \leq t \leq T$
$x_{T+1} \in [K]$
$K \ll M$



Distinct tokens: There exists unique n so that $\mathbb{P}(l|n) > 0$

Common tokens: There exists multiple n so that $\mathbb{P}(l|n) > 0$

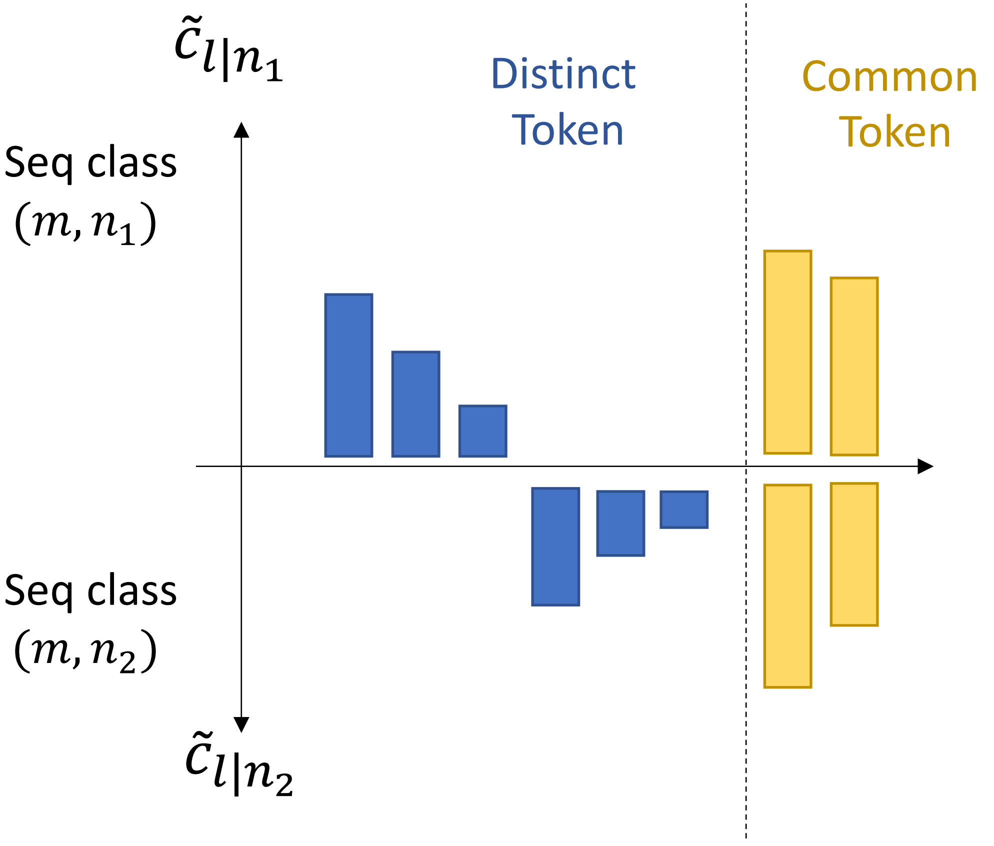
$\mathbb{P}(l|m,n) = \mathbb{P}(l|n)$ is the conditional probability of token l given last token $x_T = m$ and $x_{T+1} = n$

Assumption: $m = \psi(n)$, i.e., no next token shared among different last tokens

Question: Given the data distribution, how does the self-attention layer behave?

Overall Picture of the Training Dynamics

At initialization

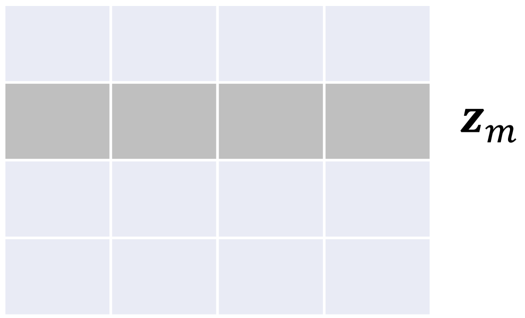


Co-occurrence probability

$$\tilde{c}_{l|n_1} := \mathbb{P}(l|m, n_1) \exp(z_{ml})$$

Initial condition: $z_{ml}(0) = 0$

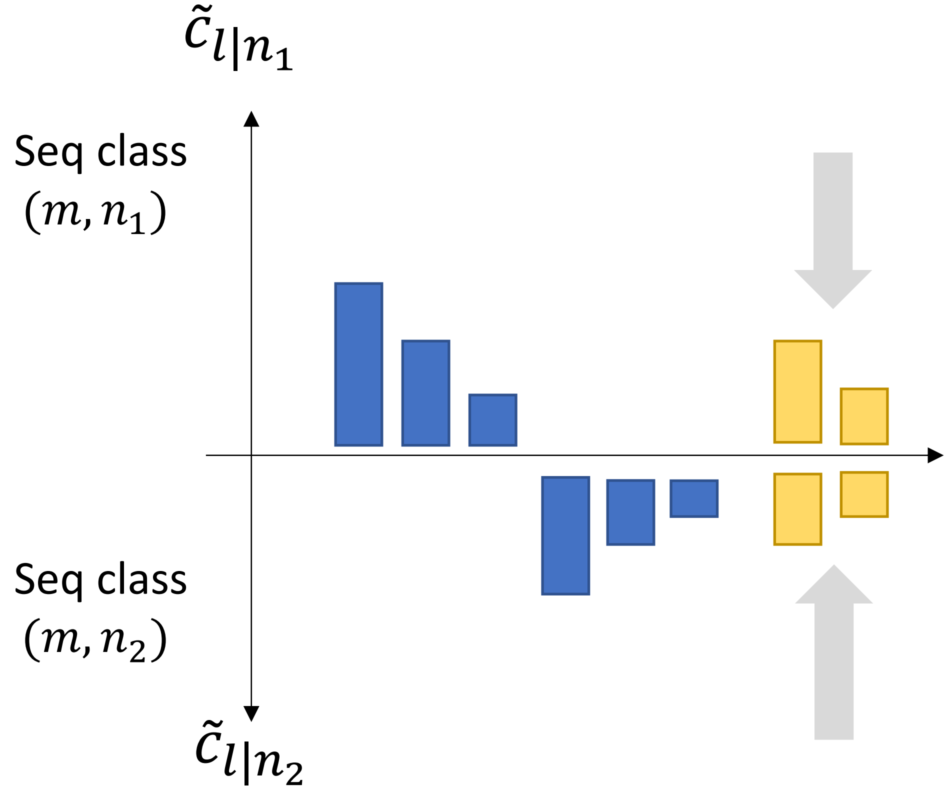
$Z =$



$\boldsymbol{z}_m$: All logits of the contextual tokens when attending to last token $x_T = m$

Overall Picture of the Training Dynamics

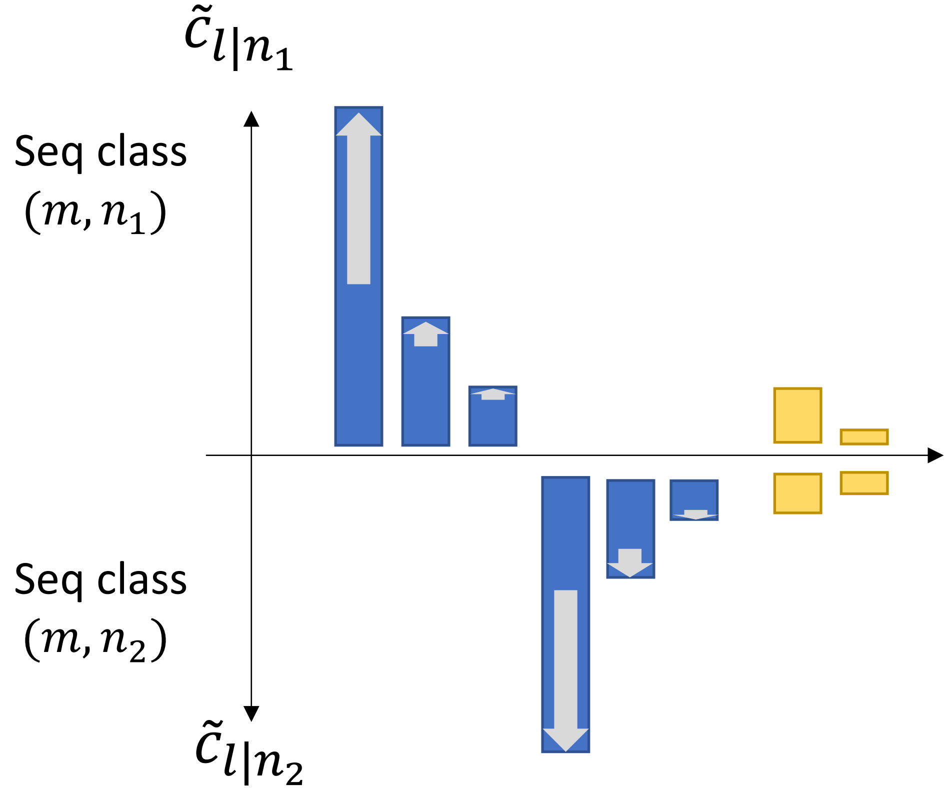
<u>Common Token Suppression</u>



(a) $\dot{z}_{ml} < 0$, for common token l

Overall Picture of the Training Dynamics

Winners-emergence



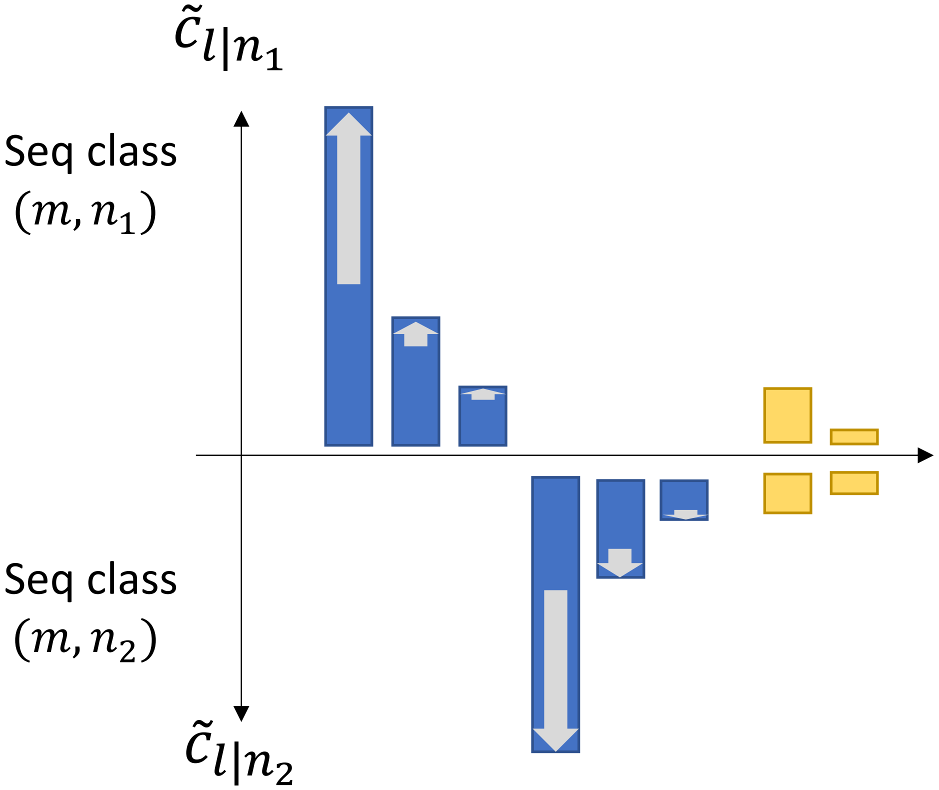
(a) $\dot{z_{ml}} < 0$, for common token l

(b) $\dot{z_{ml}} > 0$, for distinct token l

Learnable TF-IDF (Term Frequency, Inverse Document Frequency)

Overall Picture of the Training Dynamics

Winners-emergence



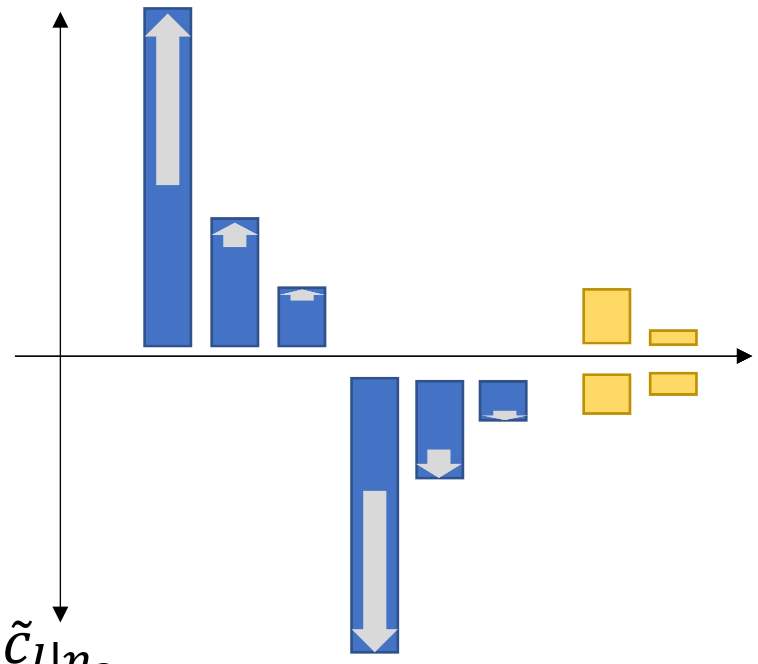
(a) $\dot{z}_{ml} < 0$, for common token l

(b) $\dot{z}_{ml} > 0$, for distinct token l

(c) $z_{ml}(t)$ grows faster with larger $\mathbb{P}(l|m,n)$

Overall Picture of the Training Dynamics

Winners-emergence



(c) $z_{ml}(t)$ grows faster with larger $\mathbb{P}(l|m,n)$

Theorem 3 Relative gain $r_{l/l'|n}(t) := \frac{\tilde{c}^2_{l|n}(t)}{\tilde{c}^2_{l'|n}(t)} - 1$ has a close form:

$$r_{l/l'|n}(t) = r_{l/l'|n}(0)\chi_l(t)$$

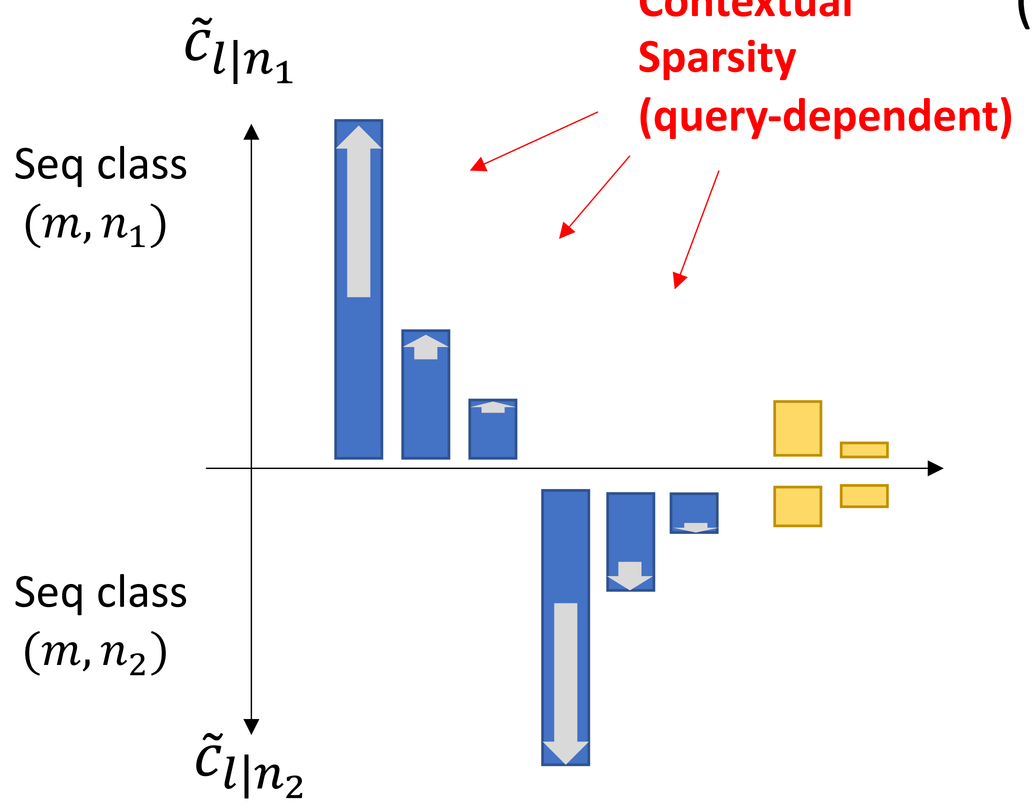
If l_0 is the dominant token: $r_{l_0/l|n}(0) > 0$ for all $l \neq l_0$ then

$$e^{2f^2_{nl_0}(0)B_n(t)} \leq \chi_{l_0}(t) \leq e^{2B_n(t)}$$

where $B_n(t) \geq 0$ monotonously increases, $B_n(0) = 0$

Overall Picture of the Training Dynamics

Winners-emergence



(c) $z_{ml}(t)$ grows faster with larger $\mathbb{P}(l|m,n)$

Theorem 3 Relative gain $r_{l/l'|n}(t) := \frac{\tilde{c}^2_{l|n}(t)}{\tilde{c}^2_{l'|n}(t)} - 1$ has a close form:

$$r_{l/l'|n}(t) = r_{l/l'|n}(0)\chi_l(t)$$

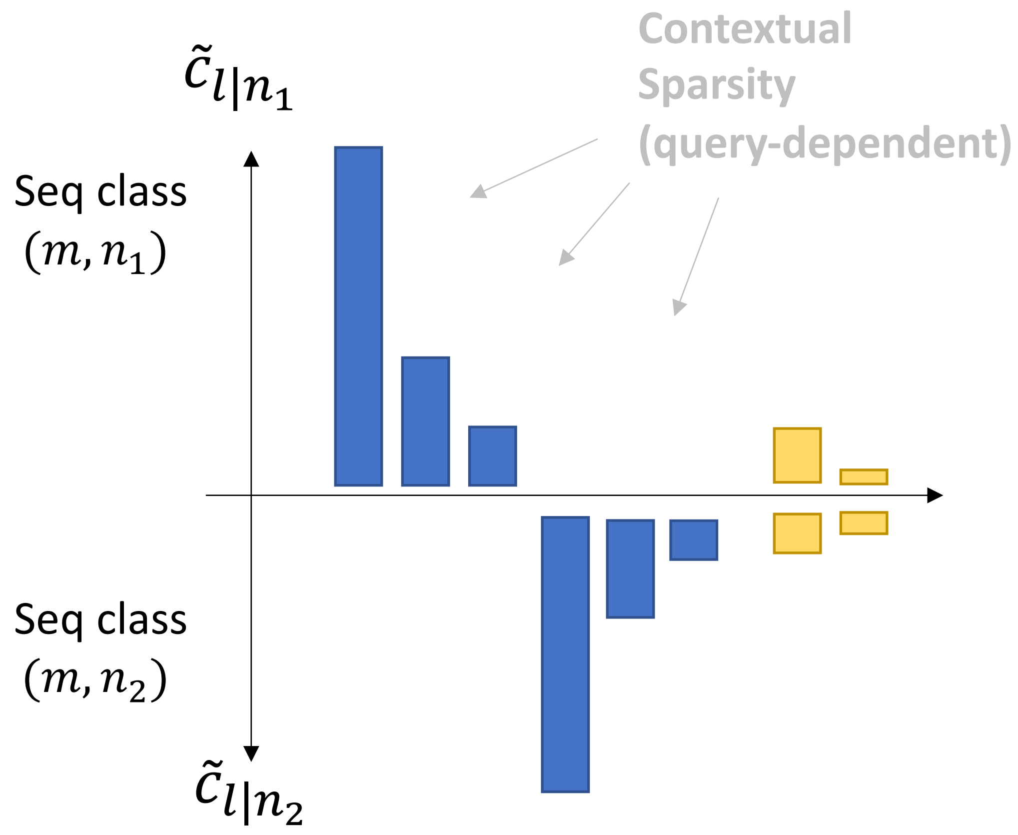
If l_0 is the dominant token: $r_{l_0/l|n}(0) > 0$ for all $l \neq l_0$ then

$$e^{2f^2_{nl_0}(0)B_n(t)} \leq \chi_{l_0}(t) \leq e^{2B_n(t)}$$

where $B_n(t) \geq 0$ monotonously increases, $B_n(0) = 0$

Overall Picture of the Training Dynamics

<u>Attention frozen</u>



Theorem 4 When $t \to +\infty$,

$$B_n(t) \sim \ln\left(C_0 + 2K\frac{\eta_z}{\eta_Y}\ln^2\left(\frac{M\eta_Y t}{K}\right)\right)$$

Attention **scanning**:

When training starts, $B_n(t) = O(\ln t)$

Attention **snapping**:

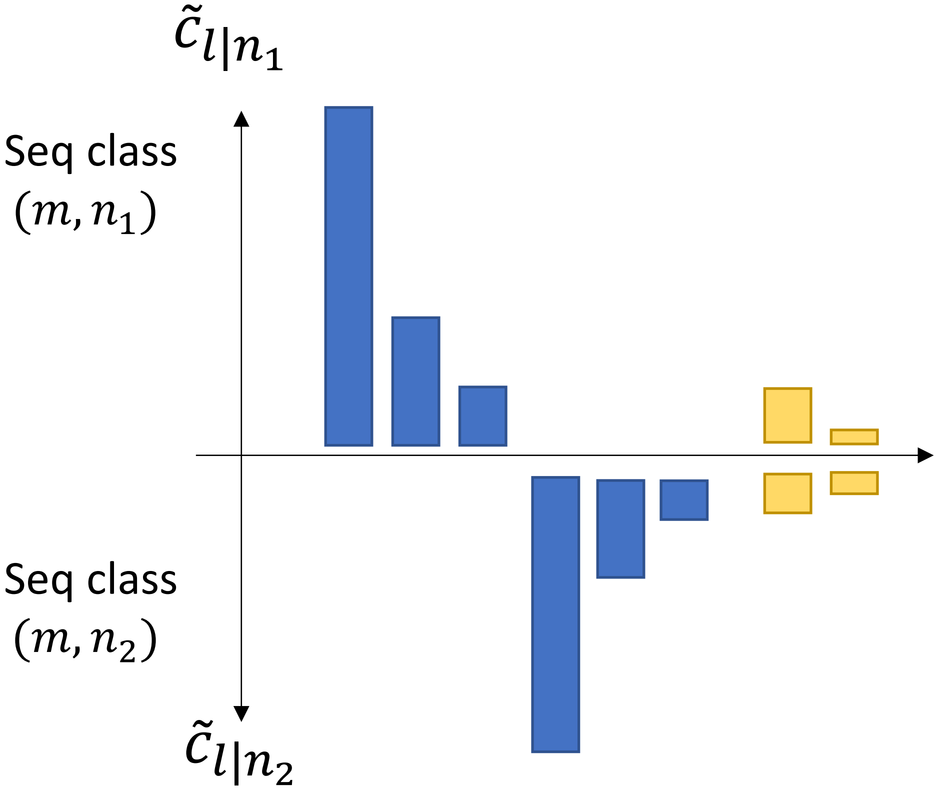
When $t \geq t_0 = O\left(\frac{2K\ln M}{\eta_Y}\right)$, $B_n(t) = O(\ln\ln t)$

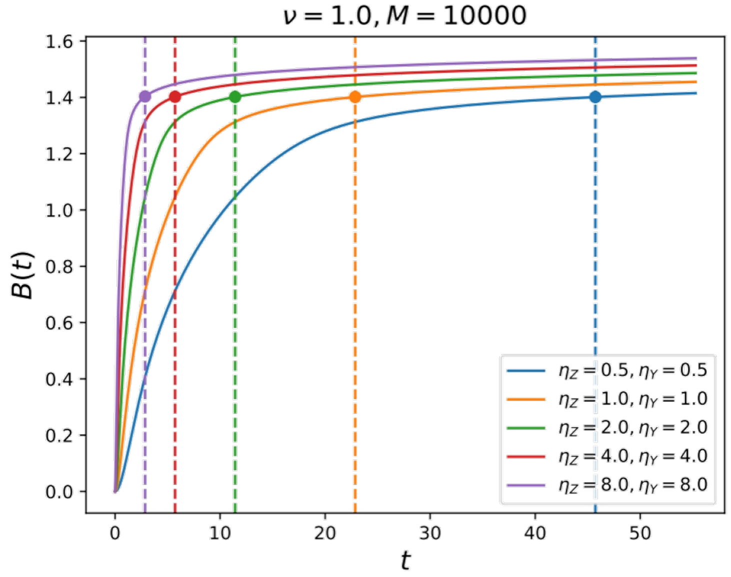
(1) η_z and η_Y are large, $B_n(t)$ is large and attention is sparse

(2) Fixing η_z, large η_Y leads to slightly small $B_n(t)$ and denser attention

Overall Picture of the Training Dynamics

Attention frozen





Larger learning rate η_z leads to faster phase transition

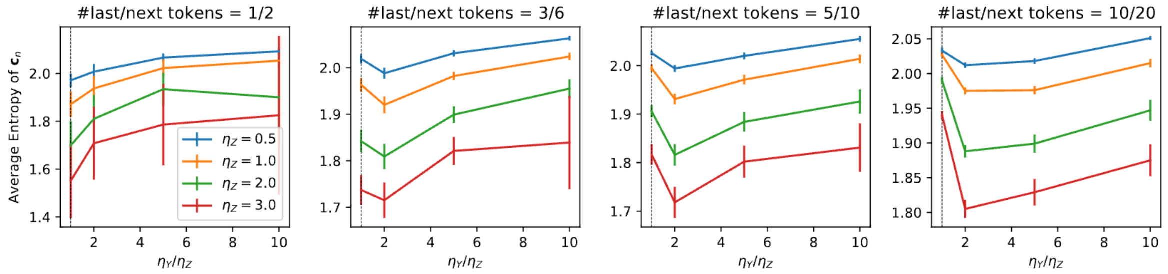


Figure 6: Average entropy of c_n (Eqn. 5) on distinct tokens versus learning rate ratio η_Y/η_Z with more last tokens M/next tokens K. We report mean values over 10 seeds and standard derivation of the mean.

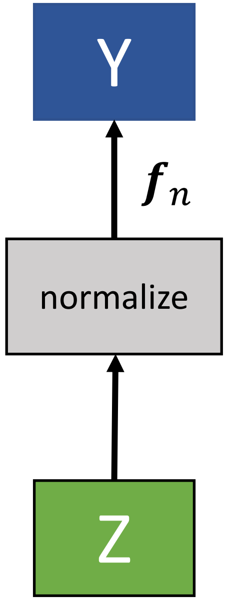
Overall strategy of the theoretical analysis

- The power of infinite sequence length $T \to +\infty$

Lemma 2. *Given the event* $\{x_T = m, x_{T+1} = n\}$*, when* $T \to +\infty$*, we have*

$$X^\top \boldsymbol{b}_T \to \boldsymbol{c}_{m,n}, \qquad X^\top \mathrm{diag}(\boldsymbol{b}_T) X \to \mathrm{diag}(\boldsymbol{c}_{m,n})$$

where $\boldsymbol{c}_{m,n} = [c_{1|m,n}, c_{2|m,n}, \ldots, c_{M|m,n}]^\top \in \mathbb{R}^M$. *Note that* $\boldsymbol{c}_{m,n}^\top \mathbf{1} = 1$.



Here $c_{l|m,n} := \dfrac{T\mathbb{P}(l|m,n)\exp(z_{ml})}{\sum_{l'} T\mathbb{P}(l'|m,n)\exp(z_{ml'})} = \dfrac{\mathbb{P}(l|m,n)\exp(z_{ml})}{\sum_{l'} \mathbb{P}(l'|m,n)\exp(z_{ml'})} =: \dfrac{\tilde{c}_{l|m,n}}{\sum_{l'} \tilde{c}_{l'|m,n}}$

Define $\boldsymbol{f}_n := \boldsymbol{f}_{m,n} := \boldsymbol{c}_{m,n} / \|\boldsymbol{c}_{m,n}\|_2$ a ℓ_2-normalized version of $\boldsymbol{c}_{m,n}$.

Overall strategy of the theoretical analysis

- Since $\eta_Y \gg \eta_Z$, we analyze the dynamics of decoder *Y* first, treating the output of *Z* as constant.

$$\dot{Y} = \eta_Y \boldsymbol{f}_n(\boldsymbol{e}_n - \boldsymbol{\alpha}_n)^\top, \quad \boldsymbol{\alpha}_n = \frac{\exp(Y^\top \boldsymbol{f}_n)}{\mathbf{1}^\top \exp(Y^\top \boldsymbol{f}_n)}$$

- The analysis gives backpropagated gradient:

Theorem 1. *If Assumption 2 holds, the initial condition* $Y(0) = 0$, $M \gg 100$, η_Y *satisfies* $M^{-0.99} \ll \eta_Y < 1$, *and each sequence class appears uniformly during training, then after* $t \gg K^2$ *steps of batch size* 1 *update, given event* $x_{T+1}[i] = n$, *the backpropagated gradient* $\boldsymbol{g}[i] := Y(\boldsymbol{x}_{T+1}[i] - \boldsymbol{\alpha}[i])$ *takes the following form:*

$$\boldsymbol{g}[i] = \gamma \left(\iota_n \boldsymbol{f}_n - \sum_{n' \neq n} \beta_{nn'} \boldsymbol{f}_{n'} \right) \tag{9}$$

Overall strategy of the theoretical analysis

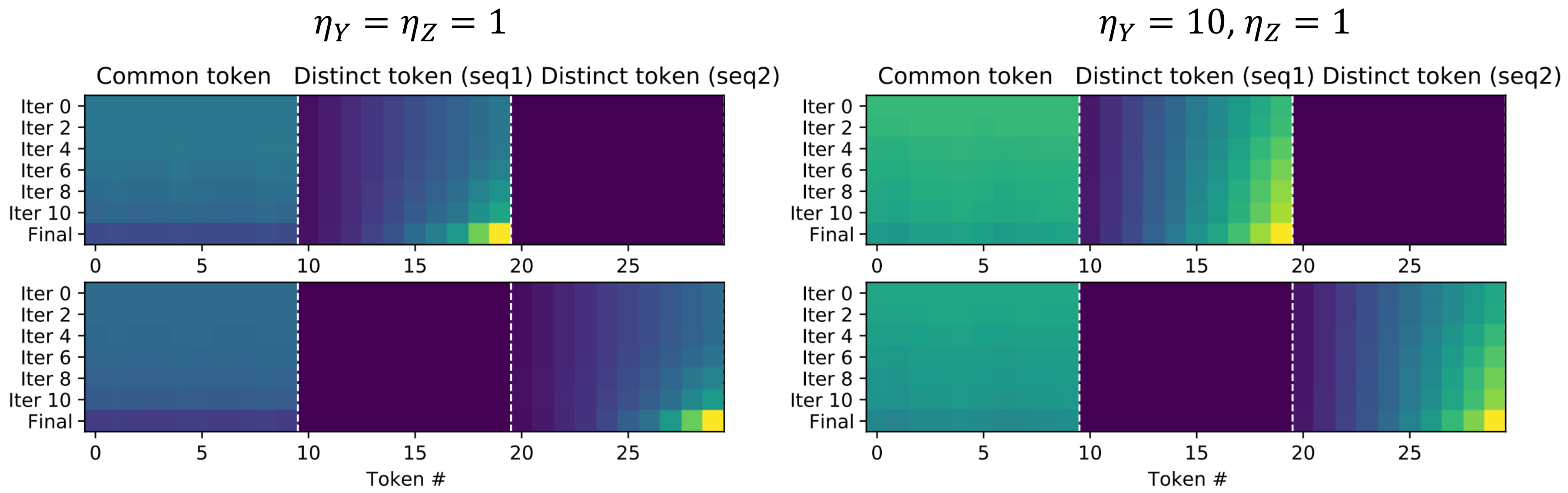
- Given the backpropagated gradient, we can analyze the behavior of the self-attention layer.

Theorem 2 (Fates of contextual tokens). *Let G_{CT} be the set of common tokens (CT), and $G_{DT}(n)$ be the set of distinct tokens (DT) that belong to next token n. Then if Assumption 2 holds, under the self-attention dynamics (Eqn. 10), we have:*

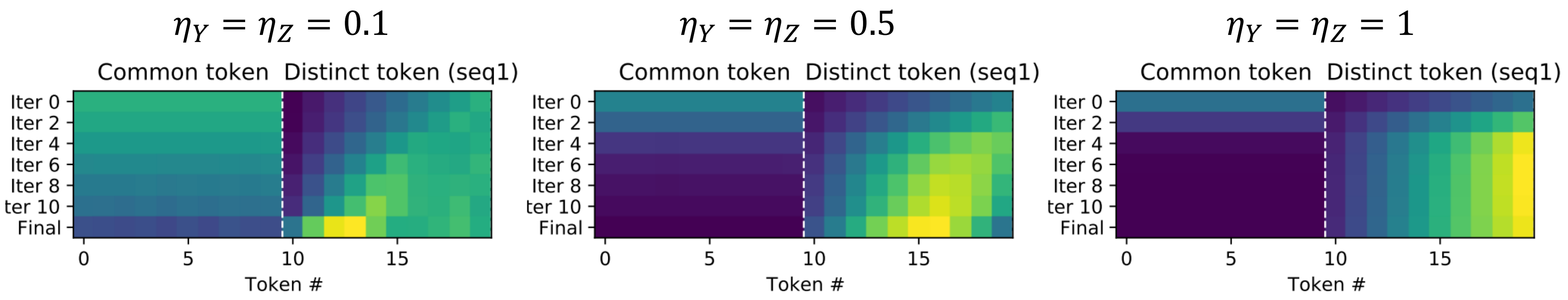
- ***(a)** for any distinct token $l \in G_{DT}(n)$, $\dot{z}_{ml} > 0$ where $m = \psi(n)$;*
- ***(b)** if $|G_{CT}| = 1$ and at least one next token $n \in \psi^{-1}(m)$ has at least one distinct token, then for the single common token $l \in G_{CT}$, $\dot{z}_{ml} < 0$.*

Visualization of c_n

SGD dynamics



Adam dynamics



Simple Real-world Experiments

WikiText2 (original parameterization)

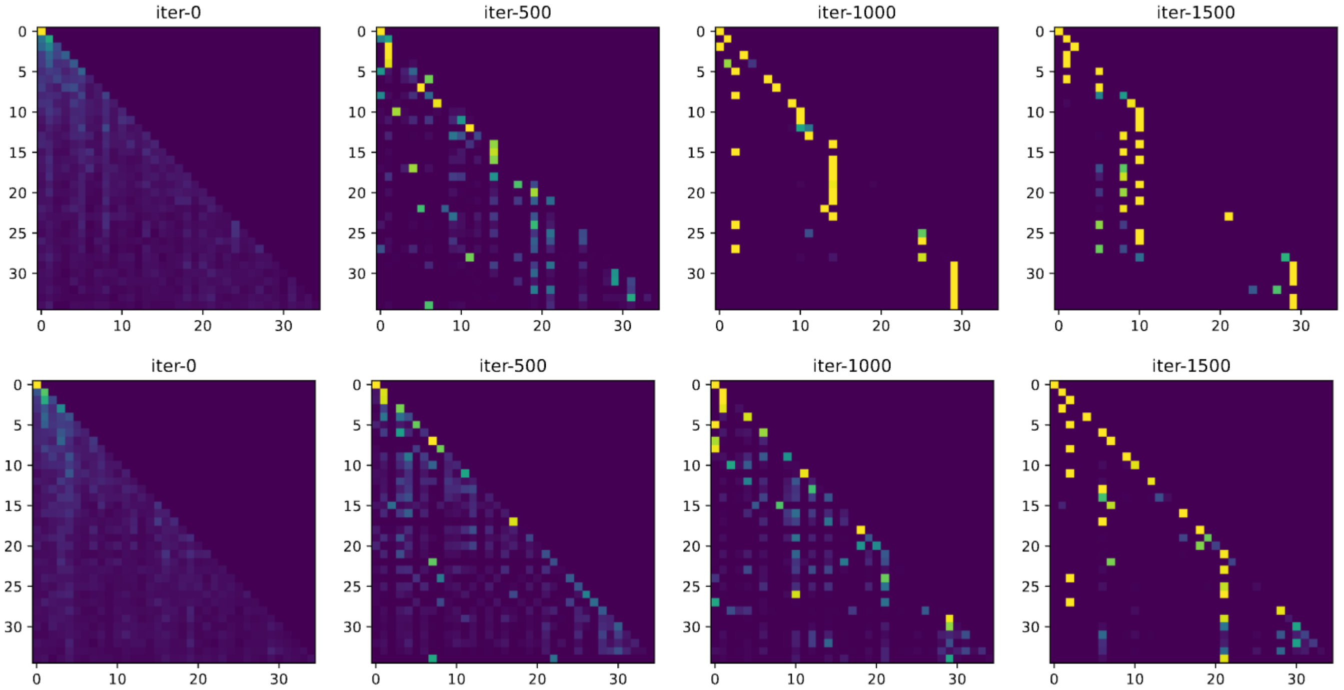


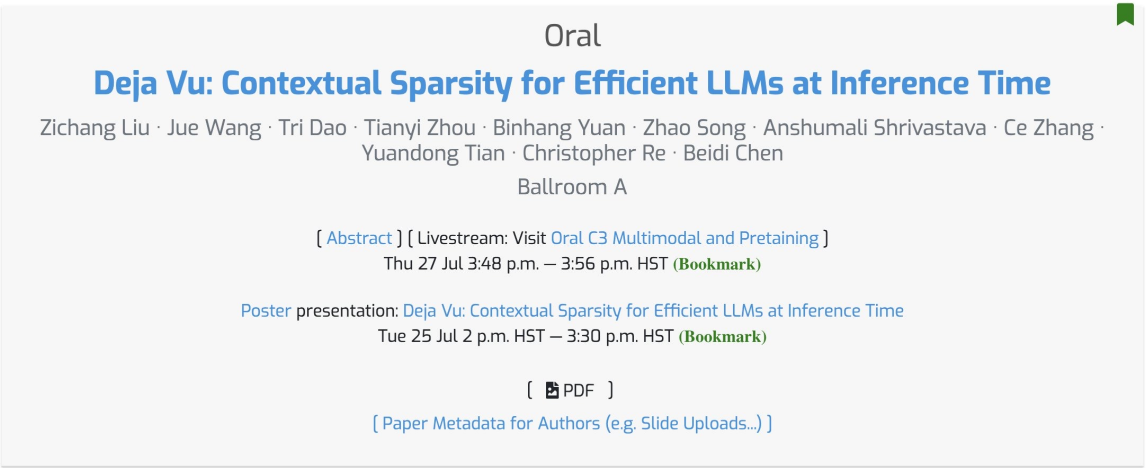
Figure 7: Attention patterns in the lowest self-attention layer for 1-layer (top) and 3-layer (bottom) Transformer trained on WikiText2 using SGD (learning rate is 5). Attention becomes sparse over training.

More ongoing experiments

- YZ parameterization works in WikiText2
  - Work even in multi-layer setting
  - Performance drops if stacking >3 layers
  - Higher perplexity than vanilla Transformer (embedding plays important role)

- Residual connection is important
  - Local distinct / common tokens

Conclusions


Oral

Deja Vu: Contextual Sparsity for Efficient LLMs at Inference Time

Zichang Liu · Jue Wang · Tri Dao · Tianyi Zhou · Binhang Yuan · Zhao Song · Anshumali Shrivastava · Ce Zhang · Yuandong Tian · Christopher Re · Beidi Chen

Ballroom A

[Abstract] [Livestream: Visit Oral C3 Multimodal and Pretaining]
Thu 27 Jul 3:48 p.m. — 3:56 p.m. HST (Bookmark)

Poster presentation: Deja Vu: Contextual Sparsity for Efficient LLMs at Inference Time
Tue 25 Jul 2 p.m. HST — 3:30 p.m. HST (Bookmark)

[PDF]

[Paper Metadata for Authors (e.g. Slide Uploads...)]

- Take home message
  - Dynamics of self-attention leads to *contextual sparsity*
  - Key tokens that do not co-occur a lot with the query token are suppressed.

- Future works
  - Why such sparsity is important for learning?
  - How to add embedding back?
  - Does understanding the dynamics of Transformer require understanding the dynamics of MLPs?

Thanks!