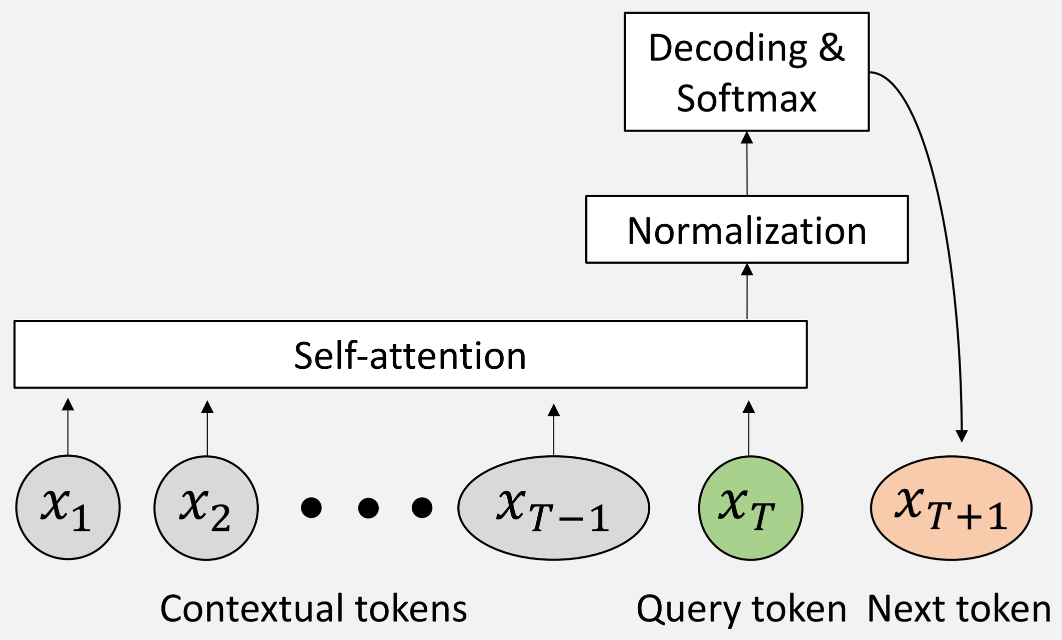
Scan and Snap: Understanding Training Dynamics and Token Composition in 1-layer Transformer

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

Problem Setting



Notation

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

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

Self-attention

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

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

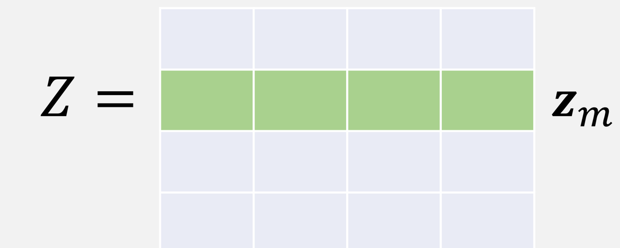
Objective:

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

Reparameterization

Reparametrize the problem with independent variable Y and Z

- Decoder $Y = UW_V^T U^T$
- Self-attention $Z = UW_Q W_K^T U^T$



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

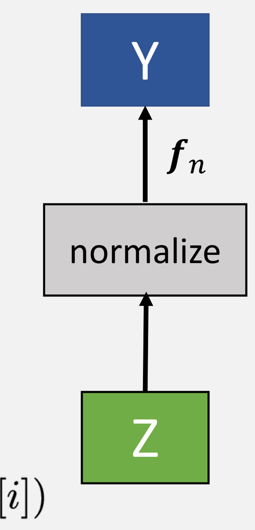
Then the SGD dynamics can be simplified:

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

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

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

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

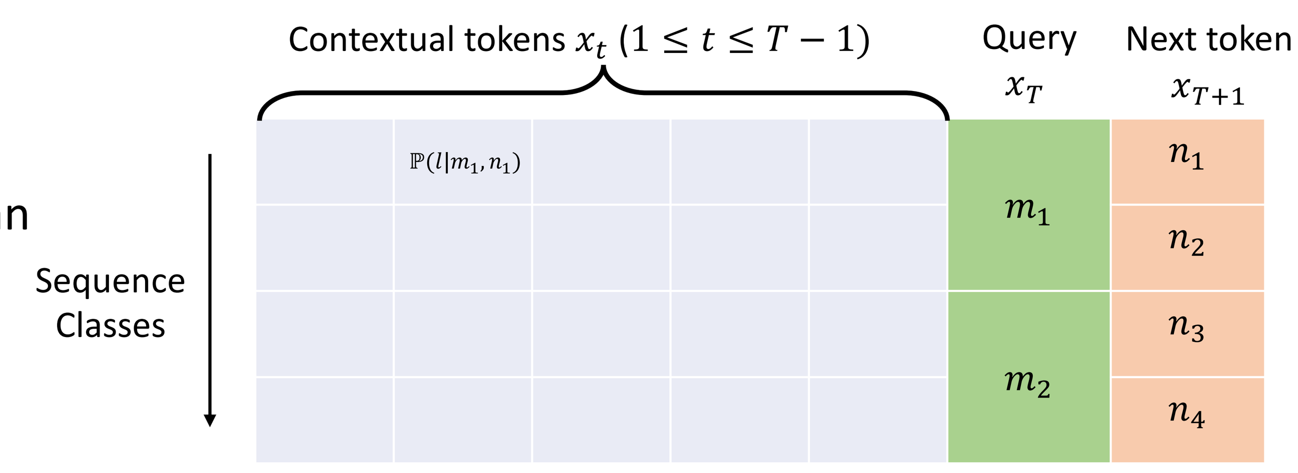


Major Assumptions

1. No positional encoding
2. Sequence length $T \to +\infty$
3. Learning rate of decoder Y is larger than self-attention layer Z ($\eta_Y \gg \eta_Z$)

For other technical assumptions, please check the paper.

Data Distribution



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

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

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

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

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

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

Dynamics of Decoder Y

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

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

K: number of possible next tokens to be predicted

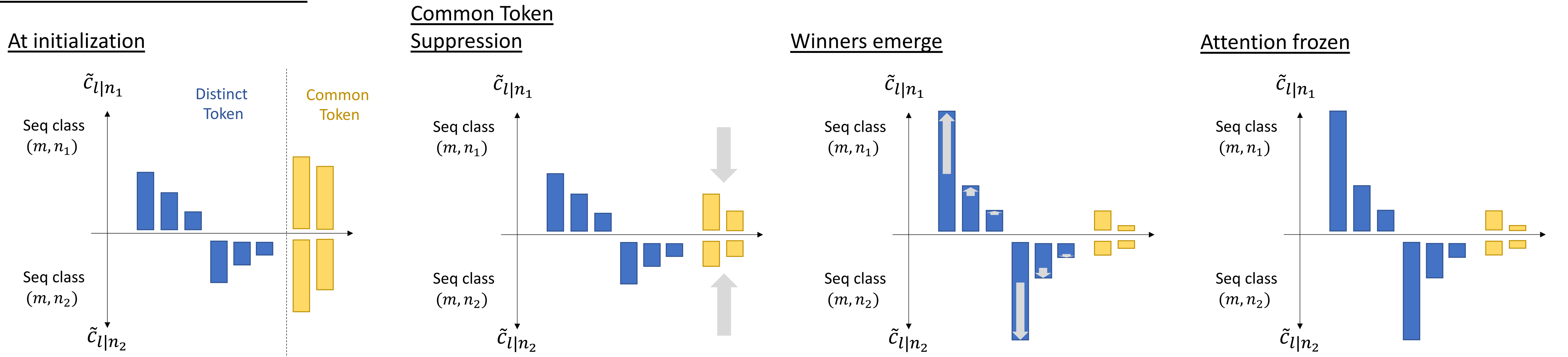
Theorem 1. *If Assumption 2 holds, the initial condition $Y(0) = 0$, $M \gg 100$, η_Y satisfies $M^{-0.99} \ll \eta_Y < 1$, and each sequence class appears uniformly during training, then after $t \gg K^2$ steps of batch size 1 update, given event $x_{T+1}[i] = n$, the backpropagated gradient $\boldsymbol{g}[i]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

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

Here the coefficients $\iota_n(t)$, $\beta_{nn'}(t)$ and $\gamma(t)$ are defined in Appendix with the following properties:

- ***(a)** $\xi_n(t) := \gamma(t)\sum_{n\neq n'}\beta_{nn'}(t)\boldsymbol{f}_n^\top(t)\boldsymbol{f}_{n'}(t) > 0$ for any $n \in [K]$ and any t;*
- ***(b)** The speed control coefficient $\gamma(t) > 0$ satisfies $\gamma(t) = O(\eta_Y t/K)$ when $t \le \frac{\ln(M)\cdot K}{\eta_Y}$ and $\gamma(t) = O\left(\frac{K\ln(\eta_Y t/K)}{\eta_Y t}\right)$ when $t \ge \frac{2(1+\delta')\ln(M)\cdot K}{\eta_Y}$ with $\delta' = \Theta(\frac{\ln\ln M}{\ln M})$.*

Dynamics of Self-attention Z



(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 $\mathbb{P}(l|m,n)$

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

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

where $B_n(t) \ge 0$ is a monotonously increasing function with $B_n(0) = 0$.

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

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

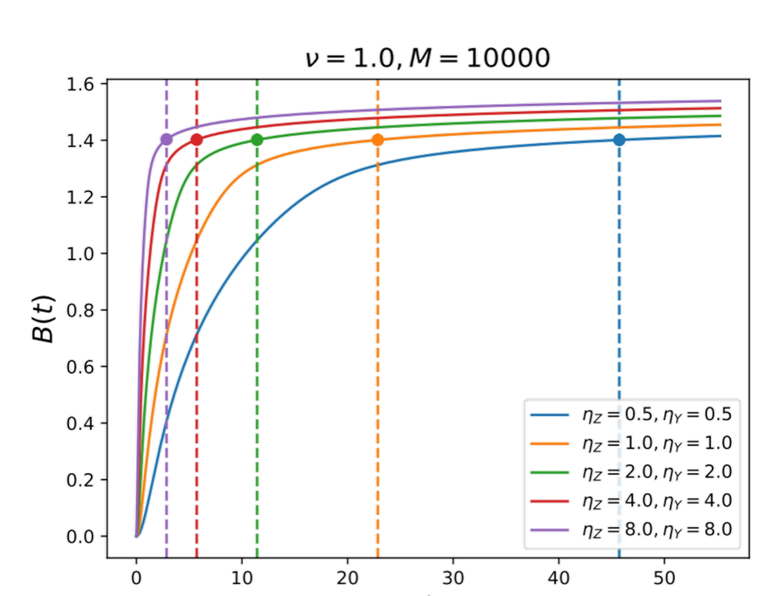
Attention scanning:

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

Attention snapping:

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

(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



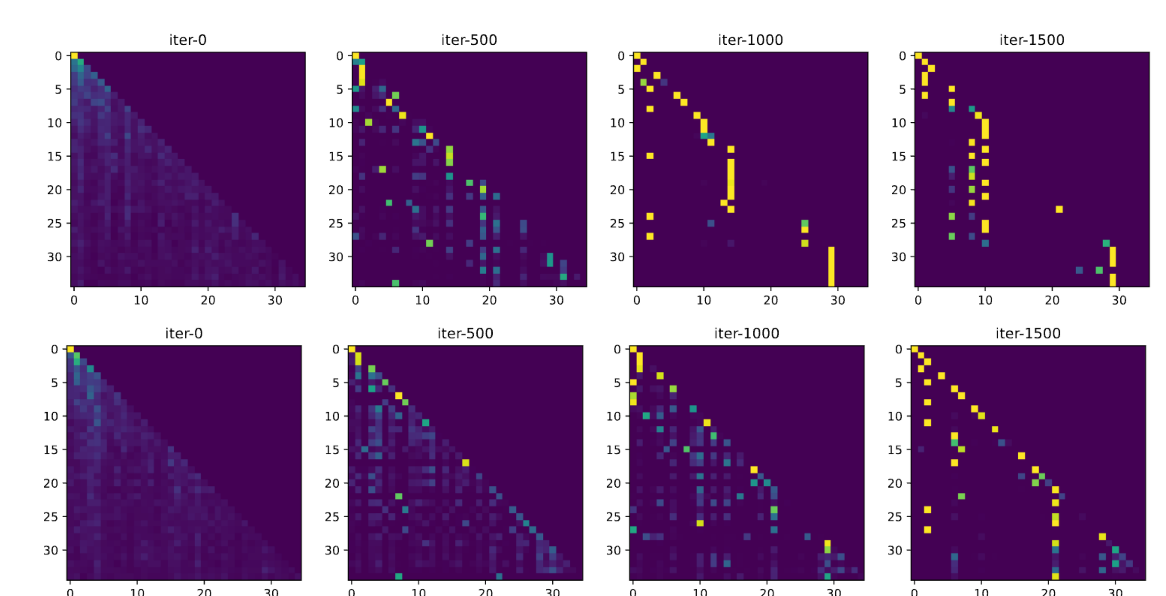


Figure 7: Attention patterns in the lowest self-attention layer for 1-layer (top) and 3-layer (bottom) Transformer trained on WikiText2 using SGD (learning rate is 5). Attention becomes sparse over training.

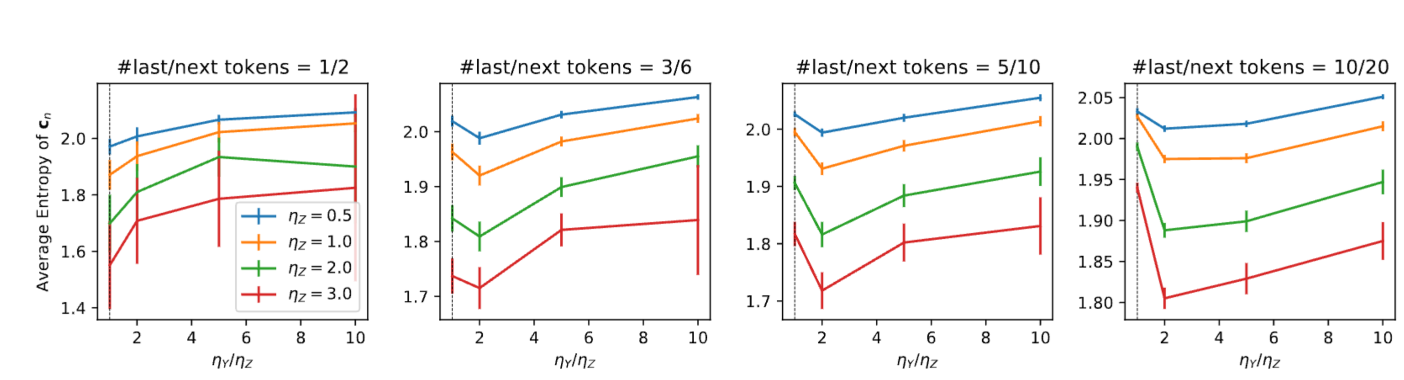


Figure 6: Average entropy of c_n (Eqn. 3) on distinct tokens versus learning rate ratio η_Y/η_Z with more last tokens M/next tokens K. We report mean values over 10 seeds and standard derivation of the mean.