The Shaped Transformer: Attention Models in the Infinite Depth-and-Width Limit at Initialization

Lorenzo Noci+,1, Chuning Li+,2, Mufan (Bill) Li+1, Bobby He†, Thomas Hofmann†, Chris Maddison‡, Daniel Roy‡

1 ETH Zurich, 2 Vector Institute, University of Toronto, 3 Oxford University

Contributions
- We study the feature covariance $V_{\text{lin}}^\beta = \frac{1}{n} (x_\beta, x_\beta)^T$ of residual networks and attention models at initialization in the proportional infinite depth-and-width limit $(d, n) \to \infty$ with constant ratio $d/n \to T$.
- For ResNet with shaped ReLU, we show that the feature covariance $V_{\text{lin}}^\beta$ can be described by an SDE in the proportional limit.
- We propose shaped attention, where we modify the Softmax attention to prevent the degeneracy of $V_{\text{lin}}^\beta$, in contrast to standard Transformers.
- We derive an SDE for $V_{\text{lin}}^\beta$ for shaped attention in the proportional limit.

Background: Transformer Architecture

- Attention Block, where $X_t \in \mathbb{R}^{m \times x}$, $m$ is sequence length, and $n$ is width:
  $$X_{t+1} = \mathcal{A} X_t + \gamma A_t X_t,$$
  $$A_t = \text{Softmax} \left( \frac{1}{\sqrt{n}} W^{\alpha \beta} \right) \mathcal{W}_{n \times n},$$
- Residual Block with Shaped ReLU:
  $$X_{t+1} = \mathcal{A} X_t + \gamma A_t X_t,$$
  $\mathcal{A}$ where all the weight matrices have iid $\mathcal{N}(0,1)$ entries, and $\sigma(x) = s_{c} \max(x,0) + s_{m} \min(x,0)$ with $s_{c} = s_{m} = \frac{1}{\sqrt{d}}$, $c^{-1} = E \sigma(g)^2$ for $g \sim \mathcal{N}(0,1)$. We assume that $\lambda^2 + \gamma^2 = 1$ for $\lambda$ independent of $d, n$.
- Background: Neural Covariance for MLPs
  Consider a multilayer perceptron (MLP):
  $$X_{t+1} = \sigma (X_t) \sqrt{n},$$
  Then as $(d, n) \to \infty$ with $d/n \to T$, the covariance $V_{\text{lin}}^\beta$ converges to the SDE
  $$dV_t = b_{\text{lin}}(V_t) dt + \Sigma_{\text{lin}}(V_t)^{1/2} dB_t, \quad V_0 = \frac{1}{n} X_0 X_0^T,$$
  where $b_{\text{lin}}$ and $\Sigma_{\text{lin}}$ have explicit formulas.

Neural Covariance SDE for ResNets

**Theorem**

Let $X_t$ be the hidden layers of a ResNet with $\lambda^2 + \gamma^2 = 1$. Then the feature covariance $V_t$ converges to the solution of the following SDE:
$$dV_t = b(V_t) dt + \Sigma(V_t)^{1/2} dB_t, \quad V_0 = \frac{1}{n} X_0 X_0^T,$$
where $b(V_t) = \gamma^2 \sigma_{\text{lin}}(V_t)$, $\Sigma(V_t) = 2\gamma^2 \Sigma_{\text{lin}}(V_t)$.

Neural Covariance SDEs for Shaped Attention Models

**Theorem**

Let $X_t$ be the hidden layers of a residual attention network with shaped attention, parameters $\lambda^2 + \gamma^2 = 1$ and $\tau = \kappa \sqrt{m/n}$. Then the feature covariance $V_t$ converges locally to the solution of the following SDE:
$$dV_t = b(V_t) dt + \Sigma(V_t)^{1/2} dB_t, \quad V_0 = \frac{1}{n} X_0 X_0^T,$$
where the drift $b(V_t)$ has the following form
$$b(V_t) = \kappa^2 \left( \frac{1}{n} \sum_{i=1}^m V_{\text{lin}}(\text{Softmax}(V_t)) \right) \gamma^2 \sum_{\nu,\kappa} \left( V_{\text{lin}} \gamma^2 \sum_{\sigma,\omega} \Sigma_{\text{lin}} + V_{\text{lin}} \gamma^2 \right),$$
the diffusion coefficient is defined by
$$\Sigma(V_t) = \gamma^2 \sum_{\nu,\kappa} \left( V_{\text{lin}} \gamma^2 \sum_{\sigma,\omega} \Sigma_{\text{lin}} + V_{\text{lin}} \gamma^2 \right),$$
and $A_t = \kappa \sqrt{m/n}$ is equal to:
$$A_t = \kappa^2 \left( \frac{1}{n} \sum_{i=1}^m V_{\text{lin}} \right) \gamma^2 \sum_{\nu,\kappa} \left( V_{\text{lin}} \gamma^2 \sum_{\sigma,\omega} \Sigma_{\text{lin}} + V_{\text{lin}} \gamma^2 \right).$$

Shaped Attention

- We propose the following modifications to the standard softmax-based attention mechanism:
  $$A_t = I + \text{Softmax}(\gamma^2 V_t) - \frac{1}{n} I^T,$$
  $\tau = \kappa \sqrt{m/n}$.
- All the modifications are necessary to stabilize $V_{\text{lin}}^\beta$.

Improving Stability of the SDE

- Notice that the drift term of the SDE for shaped attention is cubic in the entries of $V_t$, which may not have a global solution in time.
- Choosing an adversarial set of parameters, we are able to replicate numerical instabilities, and remedy them by tuning $\gamma$.
- Stopping time is defined as $t = d^*/n$ with $d^*$ the depth beyond which one of the eigenvalues of the covariance matrix exceeds $10^8$ or drops below $10^{-4}$. 