Consensus is all you get: the role of attention in transformers

Á. Rodríguez Abella, J.P. Silvestre and P. Tabuada

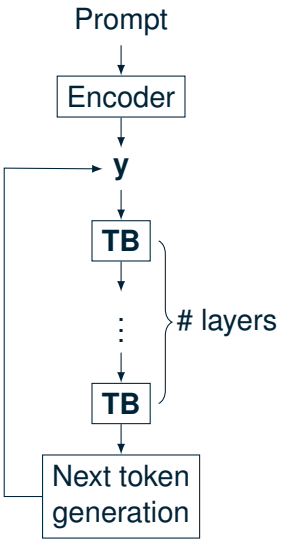
UCLA Samueli Electrical & Computer Engineering

CyPhy Lab

42nd International Conference on Machine Learning

July 13–19, 2025

- Prompt encoded: $\mathbf{y} = (\mathrm{y}_1, \ldots, \mathrm{y}_\ell) \in (\mathbb{R}^{n+1})^\ell$.
- Propagation through the layers of the network.
- **TB**: Transformer block.
- Next token generation based on current tokens and appended at the end of the original sequence or string.
- Process finishes when an 'end of sentence' token is generated.



TRANSFORMER BLOCK

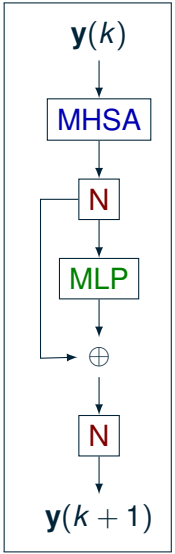
- MHSA: multi-head self-attention.
- N: layer normalization.
- MLP: multi-layer perceptron.

Transformer dynamics

$$\mathrm{y}_i(k+1) = N\big(N(f_{\mathrm{sa},i}(\mathbf{y}(k))) + f_{\mathrm{ff}}(N(f_{\mathrm{sa},i}(\mathbf{y}(k))))\big)$$

Self-attention dynamics

$$\mathrm{y}_i(k+1) = N\big(f_{\mathrm{sa},i}(\mathbf{y}(k))\big).$$



- Self-attention: $f_{\mathrm{sa},i}(\mathbf{y}) = \mathrm{y}_i + \tau \sum_{\eta=1}^{h} \sum_{j=1}^{k} \alpha_{ij}^{\eta}(\mathbf{y})\, U_\eta\, \mathrm{y}_j$.
- Value matrix: $U_\eta \in \mathbb{R}^{(n+1)\times(n+1)}$.

Self-attention matrix, $A_\eta(\mathbf{y}) = \left(\alpha_{ij}^{\eta}(\mathbf{y})\right)_{1\le i,j\le \ell} \in \mathbb{R}^{\ell\times\ell}$

$$\alpha_{ij}^{\eta}(\mathbf{y}) = \frac{1}{Z_i^{\eta}(\mathbf{y})} \exp(\mathrm{y}_i^\top P_\eta\, \mathrm{y}_j), \quad Z_i^{\eta}(\mathbf{y}) = \sqrt{n+1} \sum_{j=1}^{\ell} \exp(\mathrm{y}_i^\top P_\eta\, \mathrm{y}_j).$$

- Key and query matrices: $K_\eta, Q_\eta \in \mathbb{R}^{w\times(n+1)} \rightsquigarrow P_\eta = K_\eta^\top Q_\eta \in \mathbb{R}^{(n+1)\times(n+1)}$.

- Normalization: $N(\mathrm{y}) = \dfrac{\mathrm{y}}{|\mathrm{y}|_W}$, where:
  - $W \in \mathbb{R}^{(n+1)\times(n+1)}$ symmetric, positive definite.
  - $|\mathrm{y}|_W = \mathrm{y}^\top W \mathrm{y} \quad \rightsquigarrow \quad N(\mathbf{y}) \in \mathcal{E}_W^n = \{\mathrm{y} \in \mathbb{R}^{n+1} \mid |\mathrm{y}|_W = 1\}$.
- When $0 < |\tau| \ll 1$, the discrete dynamics can be approximated by:

Continuous dynamics

$$\dot{\mathrm{y}}_i = T_{\mathrm{y}_i} N \cdot \left(\sum_{\eta=1}^{h} \sum_{j=1}^{\ell} \alpha_{ij}^{\eta}(t, \mathbf{y})\, U_\eta(t)\, \mathrm{y}_j \right), \qquad t \geq 0,\ 1 \leq i \leq \ell.$$

- The solution of this equation models the evolution of the tokens along the consecutive attention layers.

	Full attention	**Causal attention (auto-regressive)**	
# of heads	$h \geq 1$	$h \geq 1$	$h = 1$
$P(t) = Q(t)^\top K(t)$	Time varying, uniformly continuous, bounded	Time varying, bounded	Time varying, bounded
$U(t)$	Identity	Identity	Time invariant, symmetric
Result	Convergence to consensus	Asympt. stability of consensus	Asympt. stability of consensus
Domain of attraction	Some hemisphere	Conull (complement of zero measure)	Fixed hemisphere

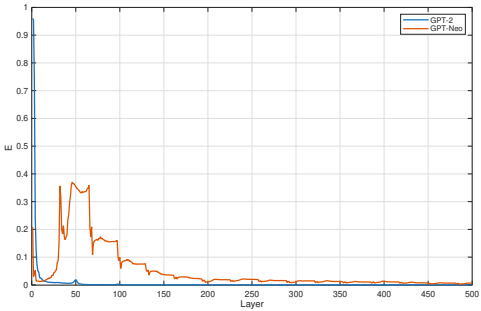
- Closest results in the literature are based on stronger assumptions.
- Time-varying case has no analogue in the literature.

- Experiments performed on GPT-2 XL and GPT-Neo 2.7B.
- Depth increased by looping the transformers, i.e., the output after each pass is fed as an input for a new pass of the model.
- 100 random prompts with 200 tokens each.

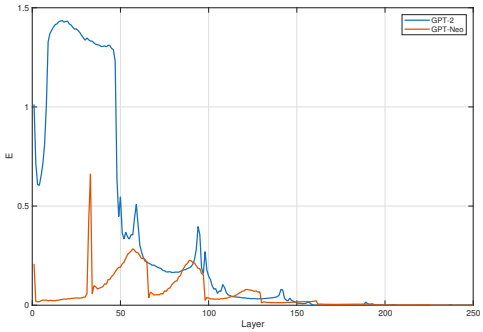
Experiments

- Looped with trained weights.
- Looped with random weights.
- Random weights reinitialized after each pass of the model.

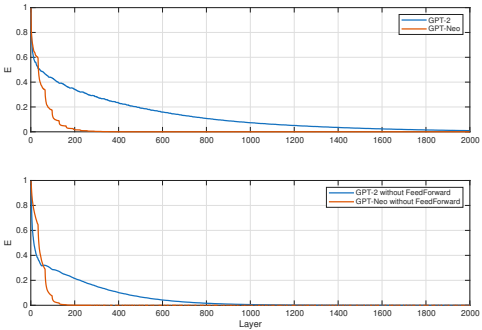
- Misalignment measured by $E(\mathbf{y}) = 1 - \frac{1}{\ell} \sum_{i=1}^{\ell} \frac{\mathrm{y}_1^\top \mathrm{y}_i}{|\mathrm{y}_1| \, |\mathrm{y}_i|}$.



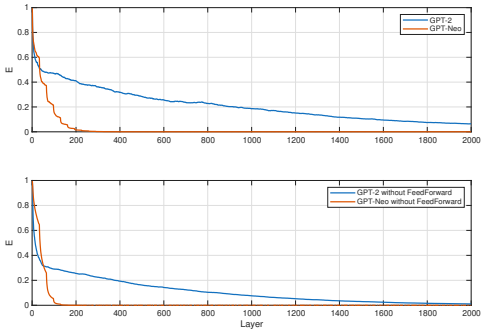
Pure self-attention.



Full model.



Looped with fixed random weights.
Top: full model.
Bottom: pure self-attention.



Random weights after each pass.
Top: full model.
Bottom: pure self-attention.

Conclusions

- Real Transformer dynamics approximated by continuous model (ResNet).
- Asymptotic analysis of the self-attention mechanism.
- Theoretical results show convergence to consensus: model collapse.
- Experiments confirm collapse, even for the full Transformer (self-attention + feed-forward layer).