The Effect of Attention Head Count on Transformer Approximation

Penghao Yu[1], Haotian Jiang[1], Zeyu Bao[1], Ruoxi Yu[2], Qianxiao Li [1]

[1]National University of Singapore
[2] Peking University

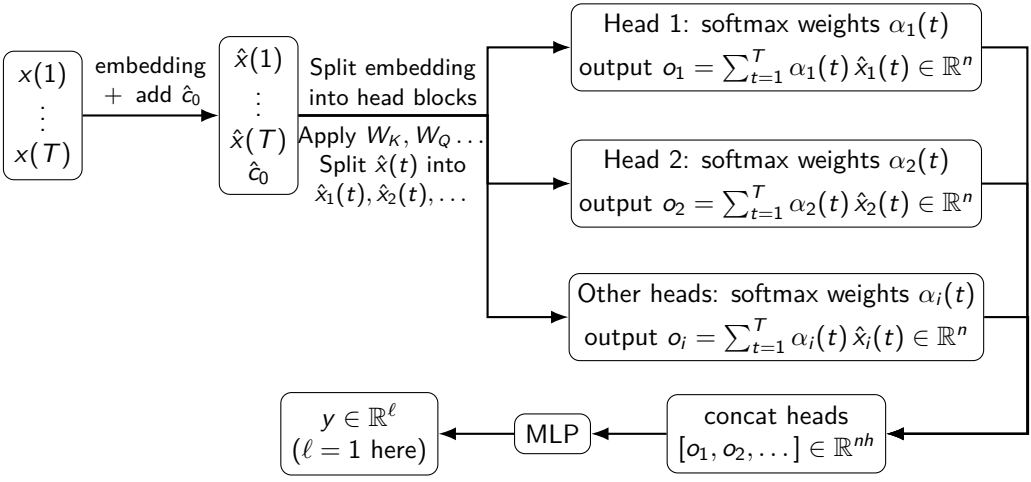
Contributions

- We establish the first rigorous lower bounds for transformers in nonlinear settings, showing that when $h < D$, parameter complexity grows exponentially with sequence length.
- We provide constructive upper bounds, proving that $h \geq D$ enables efficient approximation with parameter growth independent of sequence length T.
- In the memorization regime, single-head transformers with embedding dimension $n \geq Td$ approximate by memorizing sequences, with the complexity residing in the feed-forward block.

Background

- **Transformers** [Vaswani et al., 2017] are a dominant family of models for *sequence data*.
- They are popular in large-scale learning because they model long range dependencies and can be trained efficiently on modern hardware.
- A Transformer turns an input sequence into an output by selectively pooling information across positions.
- This leads to a basic theory question: which design choices make this pooling *expressive and efficient* as sequences get longer? We focus on one especially important choice: the **number of attention heads**.

Pipeline for Single-layer Transformer with h Heads and Per-head Embedding Dimension n (Informal)



Target class: generalized D-retrieval tasks

Build targets by (i) extracting D features from a sequence via retrieval (min over many positions), then (ii) applying a continuous post-processing map.

Definition (generalized D-retrieval tasks)

For $i = 1, \dots, D$, let $f_i : [0,1]^d \to [0,1]$ and choose subsets $S_i \subset [T]$ with $|S_i| \geq \frac{1}{4}T$. Define $\bar{z}_i(X_T) = \min_{t \in S_i} f_i(x(t)), \qquad \bar{z}(X_T) = (\bar{z}_1, \dots, \bar{z}_D) \in [0,1]^D$, and set $H(X_T) = F_0(\bar{z}(X_T)), \qquad F_0 : [0,1]^D \to \mathbb{R}$. Denote by $\mathcal{F}_D^{d,T}$ the class of all such targets.

Theorem 1 (density)

For fixed d, T, the union $\bigcup_{D \geq 1} \mathcal{F}_D^{d,T}$ is dense in $C(\mathcal{X}_T)$ under $\|\cdot\|_\infty$ (every continuous sequence-to-vector map on a compact domain can be uniformly approximated by a sequence of generalized D-retrieval tasks.)

Assumptions and intrinsic dimension

Assumptions

We impose some constraints and assumptions:

- **Model:** embeddings are uniformly bounded; the post-attention map is a standard MLP; all attention/FFN weights are uniformly bounded.
- **Non-degeneracy of Target:** each feature map f_i has a unique, well-separated minimizer (distinct across i, with positive-definite Hessian); the outer map F_0 depends on every coordinate (all partial derivatives are nonzero).

Corollary 1 (intrinsic dimension is indeed intrinsic)

If the same target H admits two generalized retrieval representations with dimensions D_1 and D_2, both satisfying the above Assumptions and with $D_1^2 + D_2^2 \leq \frac{1}{50} T$, then $D_1 = D_2$.

Theorem: Transition at the Intrinsic Dimension

Theorem 1: Transition at the Intrinsic Dimension (Informal)

Let $H \in \mathcal{F}_D^{d,T}$ and consider a single-layer transformer with h heads and per-head dimension n.

- If $h \geq D$ and $n \geq 2$, then there exists a constant $C_{d,D,T} > 0$ such that for all $M > \frac{C_{d,D,T}}{\epsilon^\gamma}$, there exists a transformer with at most M parameters that ϵ-approximates H.
- If $h < D$, define $k = \frac{\left(\frac{1}{4}T - h - D + 1\right)}{(n+1)h+1} - 1$. Then to ϵ-approximate H, this transformer must have at least $\Omega(1/\epsilon^k)$ parameters.

Here $\gamma > 0$ is the FFN approximation-rate exponent: achieving error δ for f_i and F_0 requires $O(\delta^{-\gamma})$ parameters for an FFN.

Explanation of Theorem 1

- **Phenomena:**
  - When the head count is sufficient (no less than the intrinsic dimension D), the transformer model approximates the target efficiently.
  - When the head count is below the intrinsic dimension D, the required FFN parameter budget must grow polynomially in $1/\epsilon$ with an exponent that increases linearly with sequence length T.
- **Mechanism:**
  - A generalized D-retrieval target has D distinct minima coordinates. If $h \geq D$, the heads can specialize; if $h < D$, some head must serve multiple coordinates.
  - Attention outputs weighted averages, compressing multiple relevant positions into an n-dimensional vector ($n \ll T$). As T grows, more positions must be distinguished under this fixed-dimensional compression, forcing $\hat{F}$ to do increasingly hard separation; this leads to exponential-in-T complexity.

Theorem: Single Head with Large Embedding Dimension

Theorem 2: Parameter Cost for Single-Head Transformer with Large Embedding Dimension (Informal)

Let $H \in \mathcal{F}_D^{d,T}$. For $h = 1$ and per-head embedding dimension $n \geq Td$, if the feed-forward block $\hat{F}$ is a 5-layer ReLU network, then there exists a constant $C_{d,D,T} > 0$ such that for all $M > \frac{C_{d,D,T}}{\epsilon^{1+\gamma}}$, there exists a single-head transformer with at most M parameters ϵ-approximating H.

Remark

With embedding dimension $n \gtrsim Td$, the model can encode the entire sequence into one token, so attention does not need to perform feature selection. The feed-forward block can then implement the target relation using this stored information.

Synthetic experiment: setting

Target function

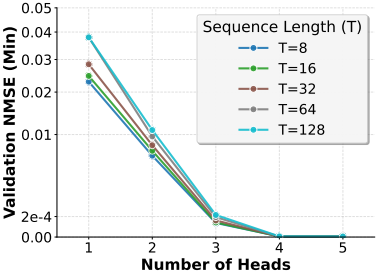
Given $X = \{x(1), \ldots, x(T)\}$ with $x(t) \in \mathbb{R}^4$, define

$$y = \sum_{i=1}^{4} \max_{1 \le t \le T} a_i^\top x(t),$$

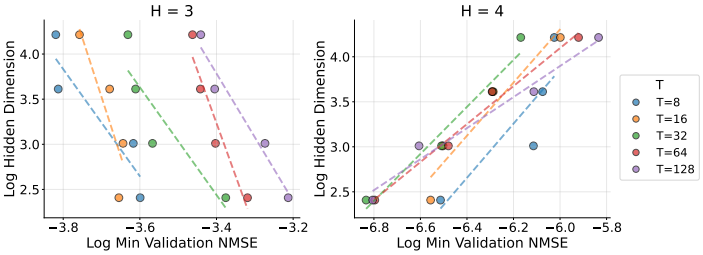
where $a_1, \ldots, a_4 \in \mathbb{R}^4$ are fixed, and $x(t) \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, I_4)$.

- Architecture: one-layer Transformers; token embedding = 2-layer ReLU MLP; add trainable cls token c_0; single MHA block (no residual / no norm); output = 2-layer GELU MLP. Both MLPs have hidden width N.

Synthetic experiment: results



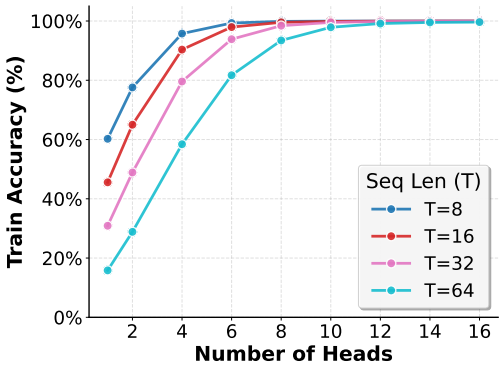
(a) **NMSE vs. Number of Heads h.**



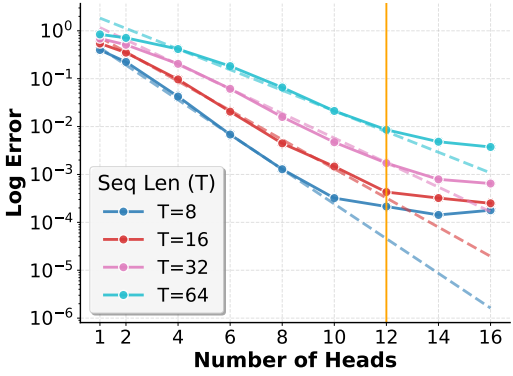
(b) **Log N vs. Log Accuracy (NMSE)**

Real experiment: MS MARCO subset

Two-layer Transformer encoder trained on an MS MARCO [Bajaj et al., 2016] subset.



(a) **Accuracy vs. Number of Heads for different T (Text Retrieval).**



(b) **Log(1-Accuracy) and its prediction (Text Retrieval)**

Thank you!

References I

Payal Bajaj, Daniel Campos, Nick Craswell, Li Deng, Jianfeng Gao, Xiaodong Liu, Rangan Majumder, Andrew McNamara, Bhaskar Mitra, Tri Nguyen, et al. Ms marco: A human generated machine reading comprehension dataset. *arXiv preprint arXiv:1611.09268*, 2016.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.