Transformers' Spectral Bias and The Symmetric Group

Itay Lavie, Guy Gur-Ari, Zohar Ringel

Background and Contributions

Transformers are widely used and show state-of-the-art performance, yet our understanding of them is still fragmented and lacking. We study inductive bias in transformers in the infinitely over-parameterized kernel limit and argue transformers tend to be biased towards more permutation symmetric functions in sequence space.

Contributions.

- We give explicit analytical predictions for the generalization performance of a NN with linear attention at the kernel limit. We show how irreducible representations of the symmetric group can be built and used to predict learnability in this case.
- We extend our results to a transformer block with standard softmax attention. We show experimentally the learnability bounds found based on the dimension of the relevant irreducible representations are tight.
- We analyze WikiText-2 and show evidence for permutation symmetry in its principal components, suggesting that the toolbox presented can be of use on natural language datasets.

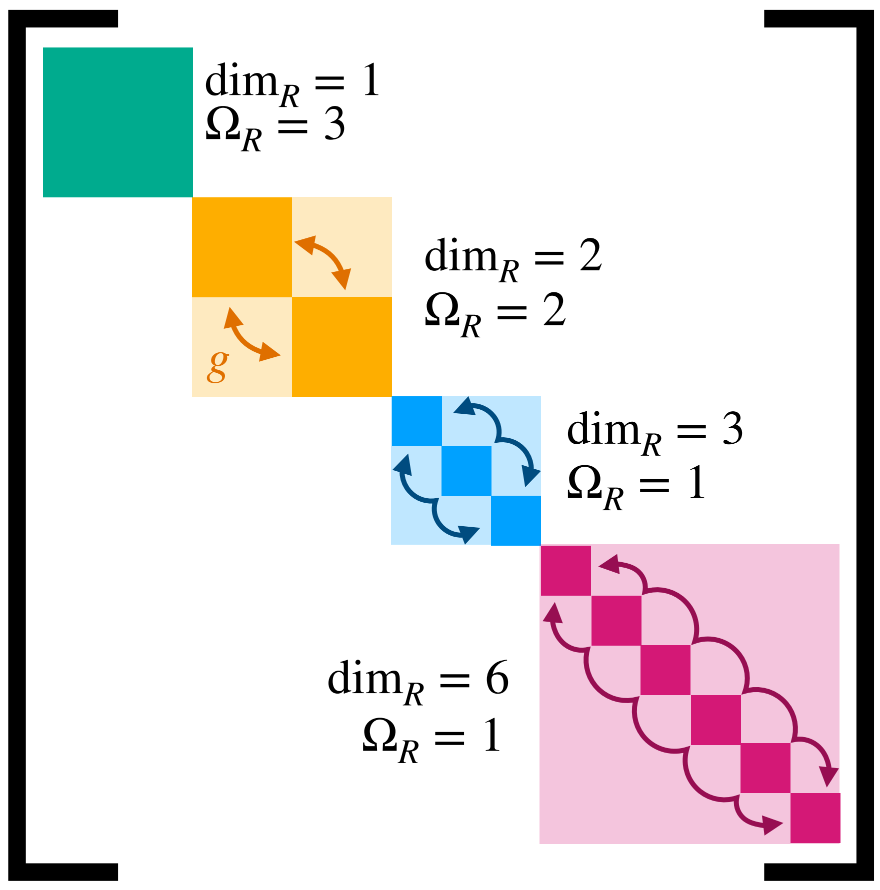


Illustration of diagonalization using symmetries

Theory

Infinitely wide transfomers ($d_k, N_h \to \infty$) admit kernel limits, where Bayesian inference is described by the regression with the NNGP kernel and learning with gradient flow is described by regression with the NTK [1].

$$\hat{f}(X_*) = \sum_{i=1}^{\infty} \frac{\lambda_i}{\lambda_i + \delta/N} g_i \varphi_i(X_*) \qquad \hat{K}\varphi_i(X) = \mathbb{E}_{Y\sim p_{\text{train}}}\left[k(X,Y)\varphi_i(Y)\right] = \lambda \varphi_i(X)$$

$$g_i = \langle g(x), \varphi_i(x)\rangle_x = \mathbb{E}_{x\sim p_{\text{train}}}\left[g(x)\varphi_i(x)\right]$$

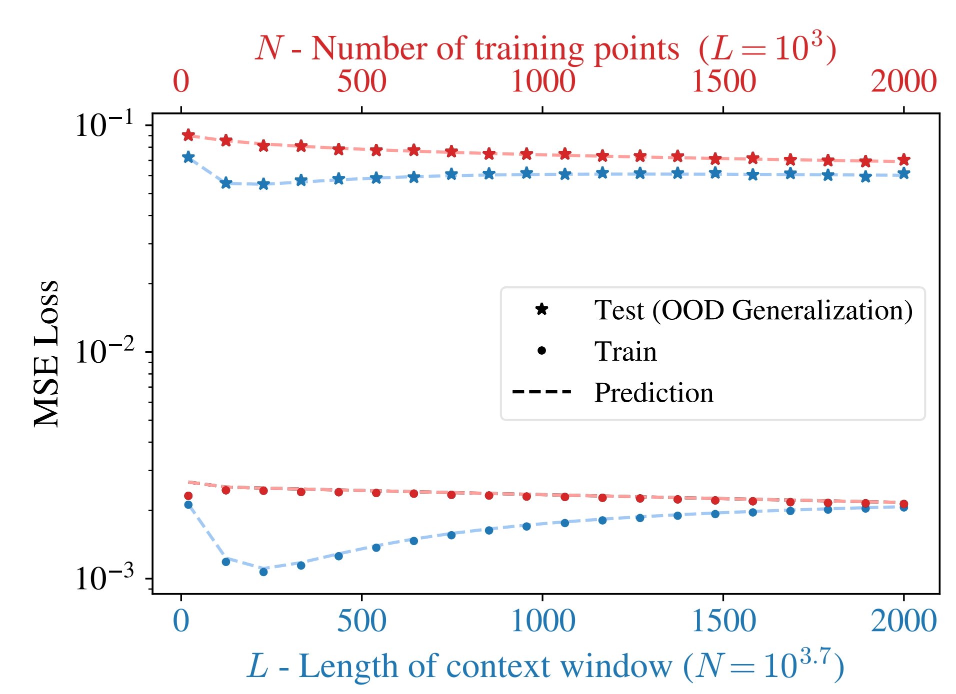
From the predictor expression, we see the sample complexity for $\varphi_i(x)$ is $N^* \simeq \sigma^2\lambda^{-1}$. We simplify the eigenvalue problem by capitalizing on the permutation symmetry present in transformer models with learned positional encoding.

Proposition. *An operator (such as $\hat{K}$) that is symmetric under the action of a group G via a faithful representation T, such that $\forall g \in G, k(T_g\vec{x}, T_g\vec{y}) = k(\vec{x},\vec{y})$ & $p(T_g\vec{x}) = p(\vec{x})$, can be decomposed into degenerate blocks. Each one of the blocks corresponds to an irrep R of G and its eigenvalue is bounded by the dimension of the irrep $\lambda_R = O(\dim_R^{-1})$.*

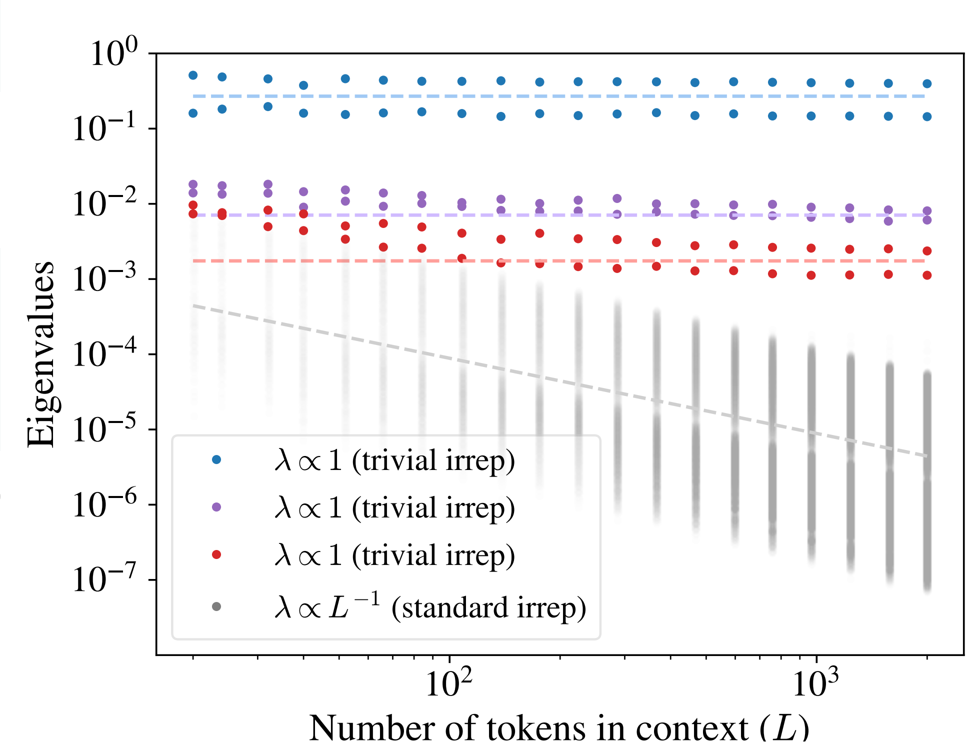
Characterizing the irreps of the symmetric group over the polynomials of one-hot encoded vectors allows us to bound the sample complexity scaling with L.

Theorem. *The space of homogeneous multilinear polynomials in n variables of degree d can be fully decomposed into $\min\{d+1, L-d+1\}$ unique irreps of the symmetric group S_L labeled by the partitions $(L-m, m)$ for $0 \le m \le d, n-d$. The $(L-m,m)$ irrep has dimension $\dim_R \sim L^m$.*

The result is an asymptotic bound on the sample complexity: $N \simeq \lambda_{(L-m,m)}^{-1}\sigma^2 = \Omega(L^m)$



Theory Vs. Experiment

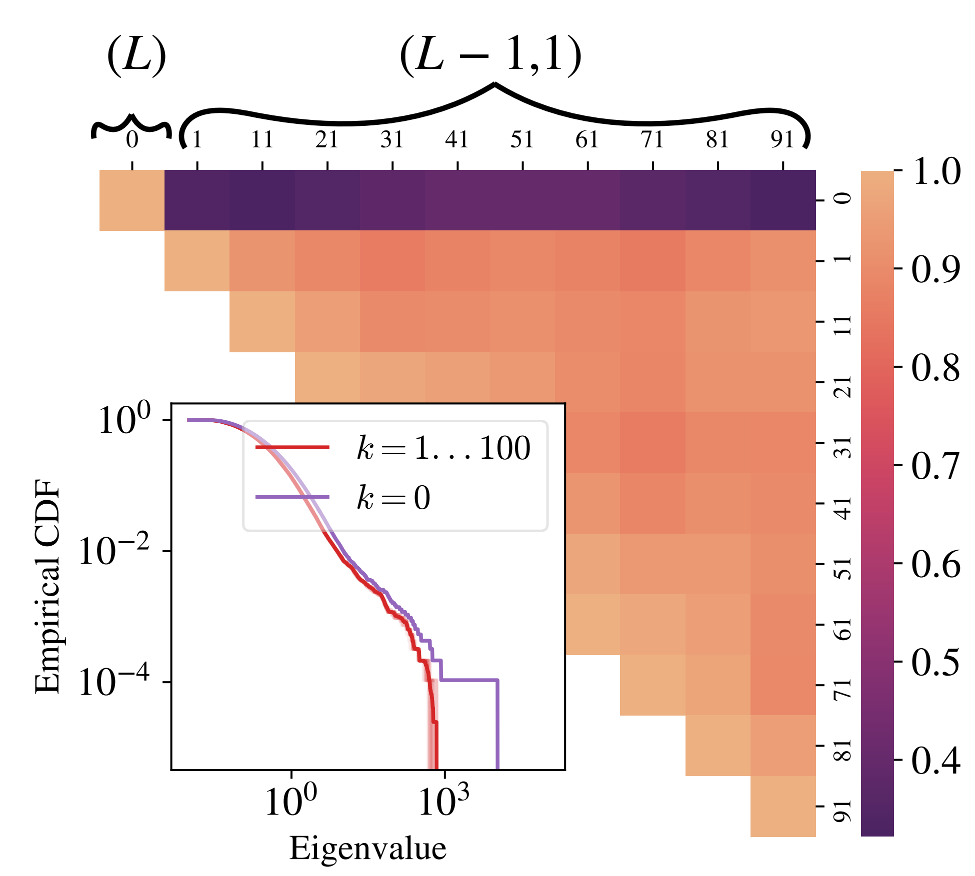


Kernel eigenvalues scaling law

Model

Network. Embedding and learned positional encoding → multi-head self-attention with a non-linearity Φ → one hidden layer MLP with non-linearity ϕ → linear readout.

Dataset. Each of the N samples is a sequence of L one-hot encoded tokens drawn from Hidden Markov Model. The HMM is drawn from a mixture for each sample.



Evidence for permutation symmetry in WikiText

Experimental Results

The top figure shows the predictions for the loss as a function of N and L together with exact Bayesian inference. We find good agreement both on train and test (OOD).

In the middle figure the spectrum of the kernel, for a NN with softmax attention and linear MLP is shown. The eigenvalues take the maximum scaling possible based on the degeneracy of the irrep $\lambda = O(\dim_R^{-1})$.

In the bottom figure, we probe the permutation symmetry in the first-order correlations of WikiText-2. We find a large similarity in the $(L-1,1)$ irrep ($k \neq 0$), that does not exist with the (L) irrep ($k = 0$). The spectrum of the different correlation matrices inside the $(L-1,1)$ irrep is almost identical as well, as indicated by the eigenvalue CDF in the same figure. This similarity, again, does not exist between the two irreps (i.e. $k = 0, k \neq 0$).

References

[2] Jiri Hron et al. (2020). "Infinite attention: NNGP and NTK for deep attention networks". In: Proceedings of the 37th International Conference on Machine Learning.