Transformers are universal in-context learners

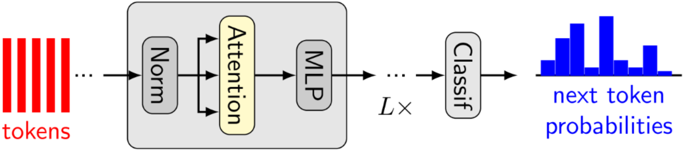
Takashi Furuya

Doshisha University

Joint work with

Maarten V. de Hoop (Rice University)
Gabriel Peyré (CNRS, ENS, PSL)

Transformers



Transformer's architecture

The transformer [Vaswani et al 2017] was initially developed for NLP, and is a model to learn contexts (mapping sequences to sequences).

Q. How do transformers have an approximation ability?

Transformers

- $X = (x_i)_{i=1}^n \in \mathbb{R}^{d_{\mathrm{tok}} \times n}$; a set of n tokens, $x_i \in \mathbb{R}^{d_{\mathrm{tok}}}$
- $\mathrm{MAtt}_\theta : \mathbb{R}^{d_{\mathrm{tok}} \times n} \to \mathbb{R}^{d_{\mathrm{tok}} \times n}$, a multiple heads attention map with a skip-connection defined by

$$\mathrm{MAtt}_\theta(X) := X + \sum_{h=1}^{H} W^h V^h X \,\mathsf{SoftMax}(X^\top (Q^h)^\top K^h X / \sqrt{k})$$

$\theta := (W^h, K^h, Q^h, V^h)_{h=1}^H \subset \mathbb{R}^{d_{\mathrm{tok}} \times d_{\mathrm{head}}} \times \mathbb{R}^{k \times d_{\mathrm{tok}}} \times \mathbb{R}^{k \times d_{\mathrm{tok}}} \times \mathbb{R}^{d_{\mathrm{head}} \times d_{\mathrm{tok}}}$

- SoftMax function defined by

$$\forall Z \in \mathbb{R}^{n \times n}, \quad \mathsf{SoftMax}(Z) := \left(\frac{e^{Z_{i,j}}}{\sum_{\ell=1}^n e^{Z_{i,\ell}}} \right)_{i,j=1}^n \in \mathbb{R}_+^{n \times n},$$

- A transformer $T : \mathbb{R}^{d \times n} \to \mathbb{R}^{d' \times n}$ as a composition of L attention maps and MLPs:

$$T(X) = \mathrm{MLP}_{\xi_L} \circ \mathrm{MAtt}_{\theta_L} \circ \ldots \circ \mathrm{MLP}_{\xi_1} \circ \mathrm{MAtt}_{\theta_1}(X),$$

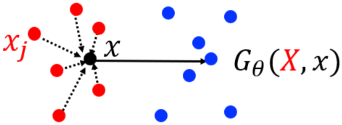
In-context mappings form

- The mapping $X \mapsto \mathrm{MAtt}_\theta(X)$ can be re-written as $G_\theta(X, \cdot)$ to each token,

$$\mathrm{MAtt}_\theta(X) = (G_\theta(X, x_i))_{i=1}^n,$$

where $x \mapsto G_\theta(X, x)$ is "in-context" mapping (depending on context X)

$$G_\theta(X, x) := x + \sum_{h=1}^{H} W^h \sum_{j=1}^{n} \frac{\exp\left(\frac{1}{\sqrt{k}}\langle Q^h x,\ K^h x_j\rangle\right)}{\sum_{\ell=1}^{n} \exp\left(\frac{1}{\sqrt{k}}\langle Q^h x,\ K^h x_\ell\rangle\right)} V^h x_j.$$



In-context mappings

In-context mappings form

- The composition of two in-context mappings denoted by $\diamond$:

$$(G_2 \diamond G_1)(X, x) \coloneqq G_2(Y, G_1(X, x)) \text{ where } Y \coloneqq (G_1(X, x_i))_{i=1}^n,$$

- The transformer's definition translated into

$$T(X) = (F_{\xi_L} \diamond G_{\theta_L} \diamond \ldots \diamond F_{\xi_1} \diamond G_{\theta_1}(X, x_i))_{i=1}^n.$$

Note that MLPs F_{ξ_ℓ} are "context-free" maps (i.e., $F_\xi(X, x) = F_\xi(x)$), while attentions G_{θ_ℓ} are "in-context" maps (i.e., $G_{\theta_\ell}(X, \cdot)$ depend on the context X).

- Here, we focus on the approximation ability of the in-context map

$$\mathbb{R}^{d\times n} \times \mathbb{R}^d \ni (X, x) \mapsto F_{\xi_L} \diamond G_{\theta_L} \diamond \ldots \diamond F_{\xi_1} \diamond G_{\theta_1}(X, x) \in \mathbb{R}^{d'}$$

Main result

Theorem 1

Let $\Omega \subset \mathbb{R}^d$ *be a compact set and* $\Lambda^\star : \Omega^n \times \Omega \to \mathbb{R}^{d'}$ *be continuous. Then for all* $\varepsilon > 0$*, there exist* L *and parameters* $(\theta_\ell, \xi_\ell)_{\ell=1}^L$*, such that*

$$\forall (X, x) \in \Omega^n \times \Omega, \quad |F_{\xi_L} \diamond G_{\theta_L} \diamond \ldots \diamond F_{\xi_1} \diamond G_{\theta_1}(X, x) - \Lambda^\star(X, x)| \leq \varepsilon,$$

with $d_{\mathrm{tok}}(\theta_\ell) \leq d + 3d'$, $d_{\mathrm{head}}(\theta_\ell) = k(\theta_\ell) = 1$, $H(\theta_\ell) \leq d'$.

Previous work: [Chulhee et al 2019], the transformers operate over an embedding dimension which grows with the number n of tokens.

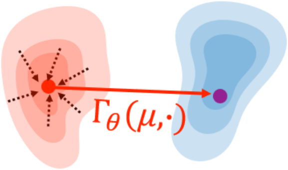
Novelty : The embedding dimensions $d_{\mathrm{tok}}, d_{\mathrm{head}}$ do not depend on ε and n.

Proof of main result

We extend to the measure-theoretic in-context maps defined as, $\forall(\mu, x) \in \mathcal{P}(\mathbb{R}^{d_{\text{tok}}}) \times \mathbb{R}^{d_{\text{tok}}}$,

$$\Gamma_\theta(\mu, x) \coloneqq x + \sum_{h=1}^{H} W^h \int \frac{\exp\left(\frac{1}{\sqrt{k}}\langle Q^h x,\, K^h y\rangle\right)}{\int \exp\left(\frac{1}{\sqrt{k}}\langle Q^h x,\, K^h z\rangle\right) \mathrm{d}\mu(z)} V^h y \,\mathrm{d}\mu(y).$$

Arbitrary number of token can be inputted.



Measure-theoretic in-context maps

Proof of main result

- When $\mu = \frac{1}{n}\sum_{i=1}^{n}\delta_{x_i}$, coincides with the discrete case, i.e.,

$$\forall X = (x_i)_{i=1}^{n}, \quad G_\theta(X, x) = \Gamma_\theta\Big(\frac{1}{n}\sum_{i=1}^{n}\delta_{x_i}, x\Big).$$

- The definition of composition generalized as

$$(\Gamma_2 \diamond \Gamma_1)(\mu, x) \coloneqq \Gamma_2(\nu, \Gamma_1(\mu, x)), \quad \text{where} \quad \nu \coloneqq \Gamma_1(\mu, \cdot)_\sharp \mu,$$

- The measure-theoretic transformer $\mathcal{T} : \mathcal{P}(\mathbb{R}^d) \to \mathcal{P}(\mathbb{R}^{d'})$ is defined by

$$\mathcal{T}(\mu) := \left(F_{\xi_L} \diamond \Gamma_{\theta_L} \diamond \ldots \diamond F_{\xi_1} \diamond \Gamma_{\theta_1}(\mu, \cdot)\right)_\sharp \mu.$$

Proof of main result

Show that

$$(\mu, x) \mapsto F_{\xi_L} \diamond \Gamma_{\theta_L} \diamond \ldots \diamond F_{\xi_1} \diamond \Gamma_{\theta_1}(\mu, x)$$

is universal in $\mathcal{C}(\mathcal{P}(\Omega) \times \Omega)$.

- Apply the Stone-Weierstrass theorem.
- We define a generalized Laplace-like transform

$$L(\mu)(a, c) := \int \frac{e^{c\langle a,\, y\rangle} \langle a,\, y\rangle}{\int e^{c\langle a,\, z\rangle}\, \mathrm{d}\mu(z)}\, \mathrm{d}\mu(y).$$

The following lemma is useful for showing the separation of points.

Lemma 2

The map $\mu \mapsto L(\mu)$ is injective.