Structure-conforming Operator Learning via Transformers

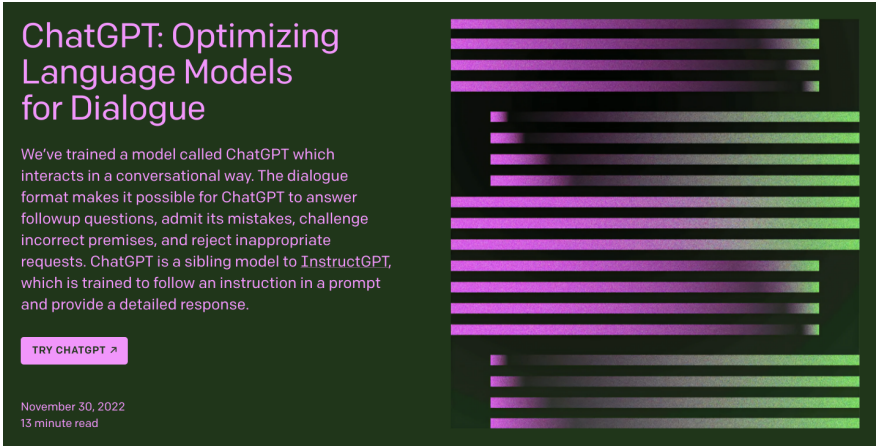
Shuhao Cao (UMKC)

joint work with Long Chen (UC Irvine) and Ruchi Guo (UC Irvine)

NSF-CBMS Conference: Deep Learning and Numerical PDEs
Morgan State University, June 19-23, 2023

What is a Transformer?

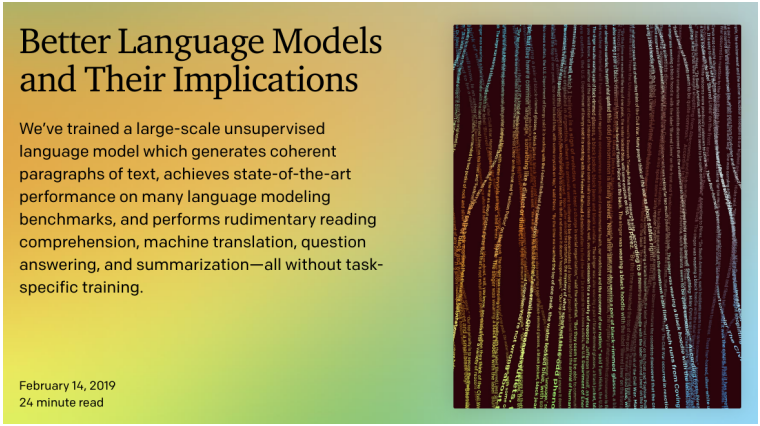
What is a Transformer?



ChatGPT. OpenAI[1].

[1] N. Stiennon et al. (2020). "Learning to summarize with human feedback". In: *Advances in Neural Information Processing Systems (NeurIPS)*.

What is a Transformer?



GPT (117m), GPT-2 (1.2b) [2], GPT-3 (175b)[3]. OpenAI.

[2] `https://github.com/karpathy/minGPT`

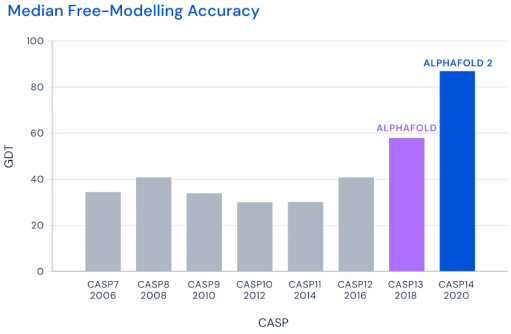
[3] J. Kaplan et al. (2020). "Scaling laws for neural language models". In: *arXiv preprint arXiv:2001.08361*.

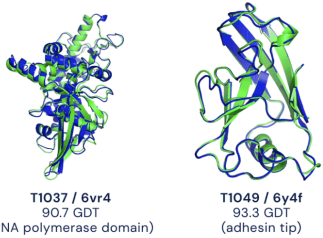
What is a Transformer?

(Left) Stable Diffusion by Stability AI[4]. (Right) AlphaTensor by Deepmind.

[4] R. Rombach et al. (2022). "High-resolution image synthesis with latent diffusion models". In: *Proceedings of the IEEE/CVF Conference on CVPR*.

What is a Transformer?





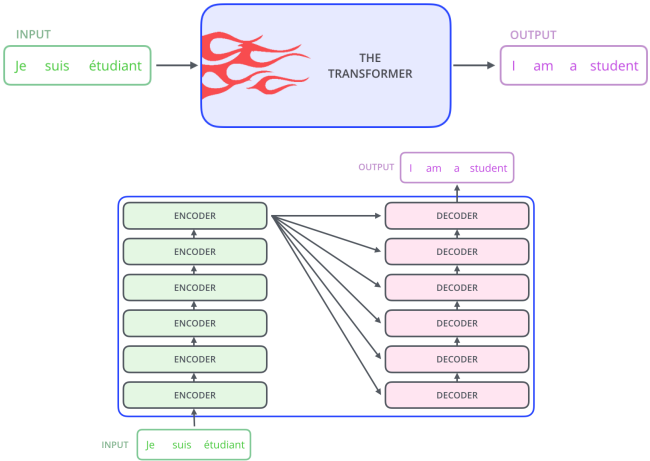
July 2021: AlphaFold 2 uses a Transformer to map the input of a multiple sequence alignment (MSA) consisting amino acids to the output of the 3D structure of a protein.

Source: Nature & Deepmind. [5][6]

[5] AlphaFold: a solution to a 50-year-old grand challenge in biology `https://deepmind.com/blog/article/alphafold-a-solution-to-a-50-year-old-grand-challenge-in-biology`

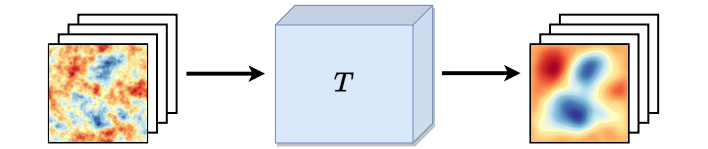
[6] J. Jumper et al. (2021). "Highly accurate protein structure prediction with AlphaFold". In: *Nature*.

What is a Transformer?



The Transformer is a deep neural network architecture to solve the machine translation problem in Natural Language Processing. Source: Jay Alammar. *The Illustrated Transformers*.

Tensor2tensor DNN



- `seq2seq` or `tensor2tensor` in Neural Machine Translation maps various sized matrices to matrices of the same size.
- Sentence in one language, embedded into high dimensional spaces, "translated" to another language's embedding

$$T : \mathbb{R}^{n \times d} \to \mathbb{R}^{n \times d}, \quad \text{or} \quad T : \mathbb{R}^{n \times n \times d} \to \mathbb{R}^{n \times n \times d}.$$

- Columns: numbers of latent/embedding dimension/channels (fixed in a given layer). Row: token embedding, patch embedding, or a DoF's embedding (in a discretization).
- The model can be trained on a lower "resolution" (short sentences) and evaluated at a higher "resolution" (longer sentences).

The state-of-the-art tensor2tensor model

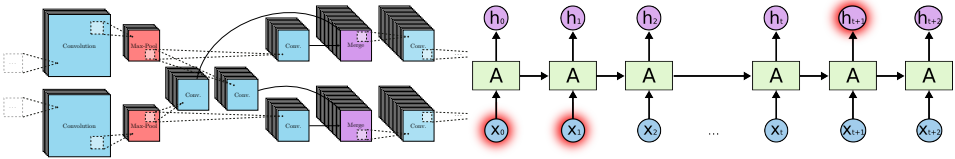
> The dominant sequence transduction models are based on complex recurrent or convolutional neural networks that include an encoder and a decoder. The best performing models also connect the encoder and decoder through an attention mechanism. We propose a new simple network architecture, the Transformer, based solely on attention mechanisms, dispensing with recurrence and convolutions entirely.

– Attention Is All You Need.[7]

- Key words: encoder, decoder, recurrent, convolutional, attention mechanism.
- Why these Google brainers want to dispense the recurrent or convolutional neural network?

[7] A. Vaswani et al. (2017). "Attention is all you need". In: *Advances in Neural Information Processing Systems (NIPS)*.

Traditional neural network-based models



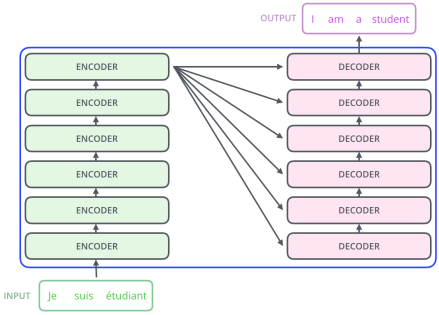
Source: (left) Cross convolution neural network[8].
(Right) Chris Olah's blog on Long Short-Term Memory.

Fully-connected (or sparsely-connected) neural network, and recurrent neural network lack in the following aspects

- Learning long range spacial (time) dependencies is an arduous process.
- Resolution-independent input output.
- Information preserving in an efficient way (full GPU/TPU saturation).

[8] P. Veličković et al. (2016). "X-CNN: Cross-modal convolutional neural networks for sparse datasets". In: *2016 IEEE symposium series on computational intelligence (SSCI)*. IEEE.

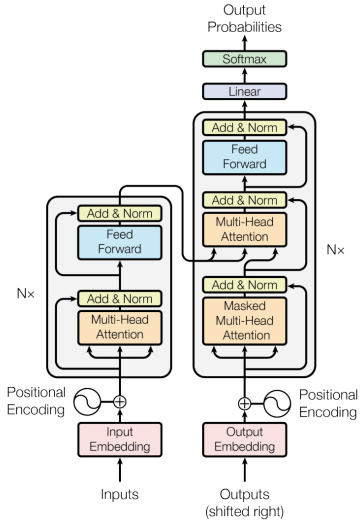
What is a Transformer?



A Transformer consists of a sequence of encoder blocks with *identical* architectures, and decoder blocks with *identical* architectures. Source: Jay Alammar. *The Illustrated Transformers*.

- Like RNN in Neural Machine Translation, the Transformer block in each layer has the same architecture (and the number of parameters).
- Unlike CNN, after the initial embedding layer (from words to vector), the latent representations propagated in the hidden layers are of the *same* discretization size.

What is a Transformer?



Source: Figure 1 in *Attention Is All You Need.*

- Softmax: $\frac{\exp(\hat{z}_k)}{\sum_{j=1}^{n} \exp\left(\hat{z}_j\right)}$.
- Linear & Feedforward: fully-connected neural network (with shared weights) at each position.
- Add: skip-connection $x \mapsto x + f(x)$.
- Positional encoding: a hard-coded mapping to differ different position in different dimensions.
- Norm: layer normalization (a learned diagonal row-scaling of the latent representation).

What is Multi-head Attention?

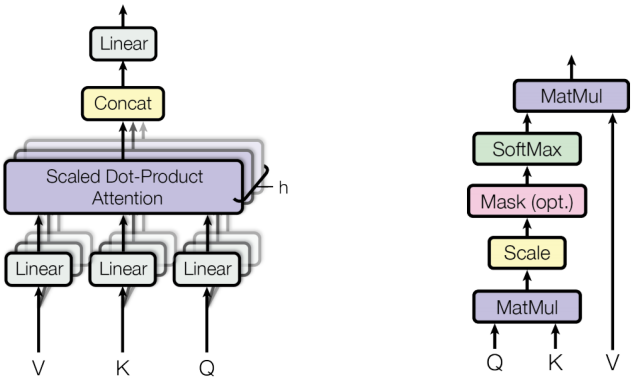
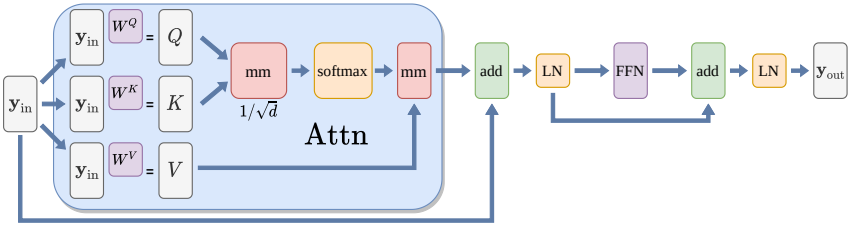


Image source: (Left) Multi-head Attention mapping. (Right) Scaled dot-product attention Softmax($QK^\top$)V. Figure 2 in *Attention Is All You Need*.

Single-head Self-Attention



Self-attention mechanism in the classical Transformer (figure reproduced for a single attention head).

- $y_{in}, y_{out} \in \mathbb{R}^{n \times d}$, input/output sequences; positional encodings added.
- Latent representations: query Q, key K, value V generated by 3 learnable matrices $W^Q, W^K, W^V \in \mathbb{R}^{d \times d}$: $Q = \boldsymbol{y} W^Q$, $K = \boldsymbol{y} W^K$, $V = \boldsymbol{y} W^V$.
- The scaled dot-product attention: $\mathrm{Attn}_s(\boldsymbol{y}) := \mathrm{Softmax}\left(d^{-1/2} Q K^\top\right) V$.
- The full attention operator (add&norm, feedforward) is then

$$\mathrm{Attn} : \mathbb{R}^{n \times d} \to \mathbb{R}^{n \times d}, \quad \boldsymbol{z} = \boldsymbol{y} + \mathrm{Attn}_s(\boldsymbol{y}), \quad \boldsymbol{y} \mapsto \mathrm{Ln}\left(\boldsymbol{z} + g\Big(\mathrm{Ln}(\boldsymbol{z})\Big)\right).$$

Single-head Self-Attention

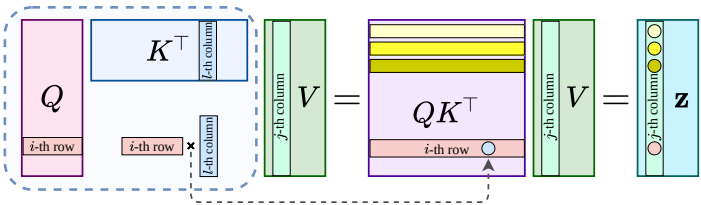
The full attention operator (single-layer, single-head):

$$\text{Attn} : \mathbb{R}^{n\times d} \to \mathbb{R}^{n\times d}, \quad \boldsymbol{z} = \boldsymbol{y} + \text{Attn}_s(\boldsymbol{y}), \quad \boldsymbol{y} \mapsto \text{Ln}\left(\boldsymbol{z} + g\Big(\text{Ln}(\boldsymbol{z})\Big)\right).$$

n: length of the input, d: latent dimensions, Ln: layer normalization, $g(\cdot)$: a trainable map (pointwise FFN in the classic Transformer).

- Positional embeddings.
- The softmax normalization makes the matrix multiplication like taking an expectation (convex combination).
- Long range spatial (time) dependencies.
- Information preserving efficiently.
- Temporal dependencies are not computed in a sequential order unlike RNNs.
- Much more parameters than RNNs yet easier/faster to train (fully GPU/TPU saturation).

Single-head Self-Attention

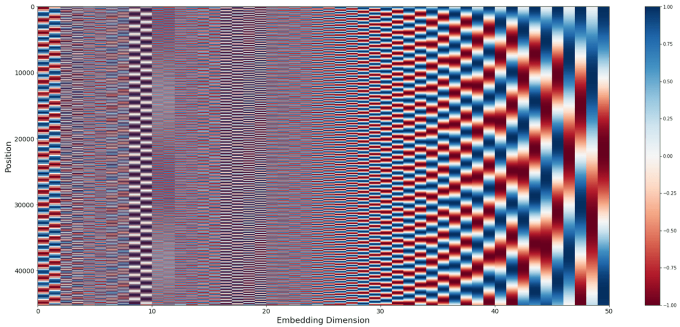


(Still) Open problem

What exactly is the mechanics of the attention mechanism?

- Kernel interpretation, RKHS: Tsai et al. (2019), Wright and Gonzalez (2021), Zhang et al. (2022).
- Fourier (change of basis): Li et al. (2021), Nguyen, Pham, et al. (2022).
- Low-rank or sparse: Y. Xiong et al. (2021), Nguyen, Suliafu, et al. (2021), Tay et al. (2020), Han et al. (2022).
- Random feature interpretation: Choromanski et al. (2021), Peng et al. (2021a).
- Iterative "solver": Yu et al. (2023)

Positional embedding



PE from *Attention is All You Need*.

- Positional embedding (PE): $\boldsymbol{x} \in \mathbb{R}^{n \times d}$ has the same dimension with the latent representation, and $\boldsymbol{y} \mapsto \boldsymbol{y} + \boldsymbol{x}$ for the $\boldsymbol{y}$ right after the input embedding.
- M: maximum discretization size; c: channel index.

$$\boldsymbol{x}_{(i,c)} = \sin\left(\frac{i}{M^{c/d}}\right) \text{ if } c \text{ is even; } \boldsymbol{x}_{(i,2c+1)} = \cos\left(\frac{i}{M^{(c-1)/d}}\right) \text{ if } c \text{ is odd}$$

Positional embeddings

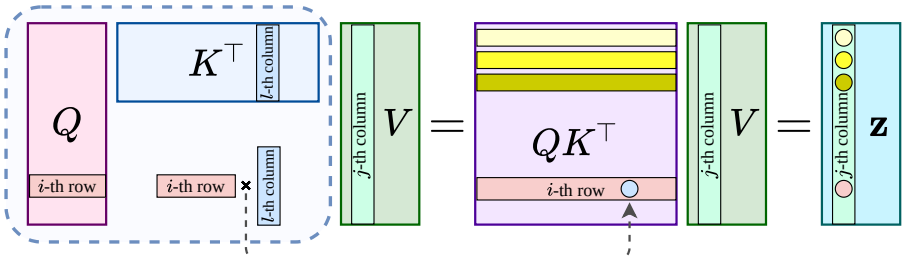
Different interpretations of the latent representation (Query/Key/Value) which is an $\mathbb{R}^{n \times d}$ matrix.

- Row: a high-dimensional embedding (vector representation) of a token.
- Column: certain discretization of "basis" function.

Open problems: Positional embeddings

- What role exactly does PE plays in attention?
- How PE shapes the topological structure of the latent representation space?
- How to design "nice" problem-oriented PE to achieve problem-specific attributes of traditional models?
- Is PE "$\approx$" coordinates? ViT: Dosovitskiy et al. (2021); DeiT: Touvron et al. (2021); Swin: Liu, Lin, et al. (2021).

Scaled dot-product Softmax($QK^\top$)V



$$(\boldsymbol{z}_i)_j = h\,\mathrm{Softmax}(QK^\top)_{i\bullet}\,\boldsymbol{v}^j = hm_i^{-1}\exp\left(\boldsymbol{q}_i\cdot\boldsymbol{k}_1,\ldots,\boldsymbol{q}_i\cdot\boldsymbol{k}_l,\ldots,\boldsymbol{q}_i\cdot\boldsymbol{k}_n\right)^\top\cdot\boldsymbol{v}^j$$

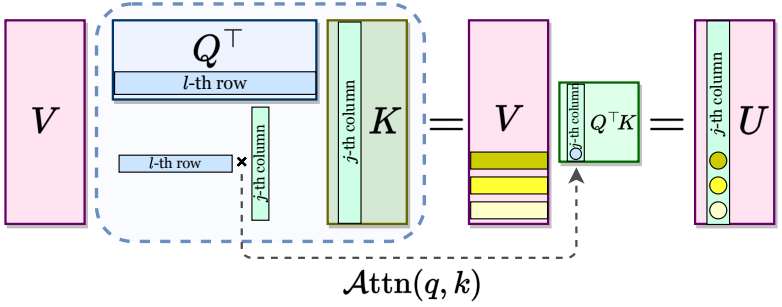
$$= hm_i^{-1}\sum_{l=1}^{n}\exp(\boldsymbol{q}_i\cdot\boldsymbol{k}_l)(\boldsymbol{v}^j)_l \approx m^{-1}(x_i)\int_\Omega \kappa(x_i,\xi)v_j(\xi)\mathrm{d}\xi,$$

The i-th row in the output computes approx. an integral transform with a non-symmetric normalized learnable low-rank "kernel" function $\kappa(x,\xi)$

$$z(x)\approx \lambda v(x)+m^{-1}(x)\int_\Omega \kappa(x,\xi;\theta)v(\xi;\theta)\mathrm{d}\xi, \qquad \text{where } \boldsymbol{q}_i=q(x_i),\boldsymbol{k}_i=k(x_i),\boldsymbol{v}_i=v(x_i).$$

Galerkin-type attention is inspired by FEM

- While it makes sense to ask the kernel to be positive (similarity between rows), it does not to ask the interaction between bases (columns) to be positive.



$$(\boldsymbol{z}^j)_i = z_j(x_i) = h\sum_{l=1}^{d}(\boldsymbol{k}^l \cdot \boldsymbol{v}^j)(\boldsymbol{q}^l)_i \approx \sum_{l=1}^{d}\left(\int_\Omega k_l(\xi)v_j(\xi)\mathrm{d}\xi\right) q_l(x_i).$$

- This is a (learnable) Petrov-Galerkin projection if if K and V are properly normalized and orthogonalized.

A preliminary result on the Galerkin-type attention

Theorem (Approximation capacity of a single layer of Galerkin attention [9])

$\mathbb{Q}_h \subset \mathcal{Q}$ and $\mathbb{V}_h \subset \mathcal{V}$ are the current approximation space, suppose there exists a continuous key-to-value map that is bounded below on the discrete approximation space, i.e., the functional norm of $v \mapsto b(q, v)$ is bounded below for any q, then for g_θ consists a Galerkin attention composed with a channel reduction map

$$\min_\theta \|f - g_\theta(\mathrm{y})\| \leq \underbrace{c^{-1}}_{\|b(q,\cdot)\|_{\mathbb{V}_h'} \geq c} \underbrace{\min_{q \in \mathbb{Q}_h} \max_{v \in \mathbb{V}_h} \frac{|b(\Pi f - q, v)|}{\|v\|}}_{\text{(Error of the Petrov-Galerkin projection)}} + \underbrace{\|f - \Pi f\|}_{\text{(Consistency)}}.$$

- Intepretation: for a "query" (a function in a Hilbert space), to deliver the best approximator in "value" (trial space), the "key" space (test space) has to be big enough so that for every value there is a key to unlock it.
- discrete Ladyzhenskaya-Babuška-Brezzi inf-sup condition: why Transformers have capacity to generalize so well with respect to the length of the sequence.

[9] S. Cao (2021). "Choose a Transformer: Fourier or Galerkin". In: *Advances in Neural Information Processing Systems (NeurIPS)*

Galerkin-type attention

- Fourier transform, revisited: consider a simple Galerkin projection in a set of orthogonal basis $\{q_j(\cdot)\}_{j=1}^d$ (wavelet, Fourier, etc.)

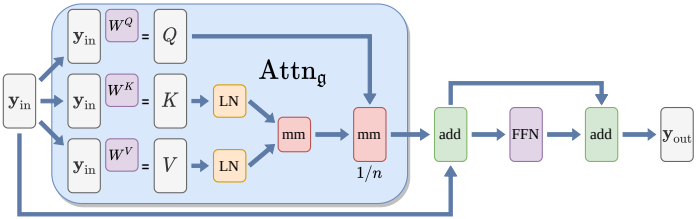
$$\min_{a_i} \left\| f - \sum_{i=1}^{d} a_i q_i(\cdot) \right\|^2_{L^2(\Omega)}, \quad \text{and} \quad z(x) := \sum_{l=1}^{d} \frac{(f, q_l)}{(q_l, q_l)} q_l(x),$$

- See also the Gram matrix inverse normalization in the saddle point problem.
- Note: this is similar to the Channel Attention with softmax[10].
- Can we use this heuristics to improve the evaluation accuracy for the attention operator without softmax? Inspired by the proof of the Ceá type lemma, and Xiong et al 2020[11], we present the following changes.

[10] S. Woo et al. (2018). "CBAM: Convolutional block attention module". In: *Proceedings of the European conference on computer vision (ECCV)*.

[11] R. Xiong et al. (2020). "On layer normalization in the transformer architecture". In: *International Conference on Machine Learning*. PMLR.

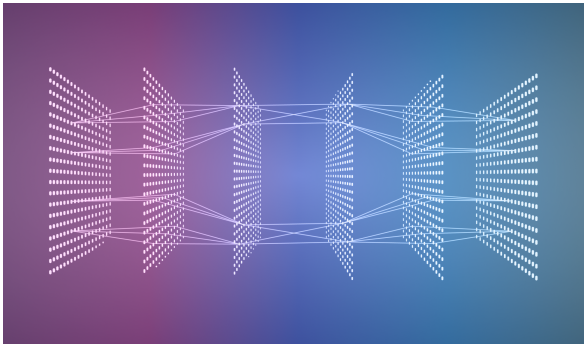
Galerkin-projection inspired attention layer



- A Galerkin projection-like layer normalization scheme together with mesh-weighted $(1/\sqrt{n})$ instead to tame the explosive matrix product.
- Positional encoding concatenated in every encoder layer and every head, unlike only once in the classic Transformer. The importance of this trick is discovered concurrently in AlphaFold 2.[12]
- Computational complexity is $\mathcal{O}(nd^2)$, cheaper than those with exponential feature maps (Random Feature Attention, FAVOR+ in Performer, etc).

[12] J. Jumper et al. (2021). "Highly accurate protein structure prediction with AlphaFold". In: *Nature*.

Representational capcity



- The embedding layer in Transformer constructs a ultra-high-dimensional vector representation of each token in a sentence or each patch in an image.
- In the encoder layer, this (latent) representation interacts with itself nonlinearly to get a "better" representation.
- This interaction can be position-wise (row-wise), or channel-wise (column-wise).

Image source: *Transformers: What They Are and Why They Matter*, Mehreen Saeed.

Representational capcity

Open problem about representational capacity

How to prove the universal representation (approximation) theorem for Transformer when the number of layers increase?

- What is representational power of (stacked) attention layer exactly[13]?
- Random feature model[14]: each channel (column) of the latent representation is similar to an RF-RR model

$$\boldsymbol{f} \mapsto \Phi(\boldsymbol{f}; \boldsymbol{\theta}) = \frac{1}{d} \sum_{j=1}^{d} \alpha_j(\boldsymbol{\theta}) g(\boldsymbol{f}; \boldsymbol{\theta})$$

where

$$\boldsymbol{\theta} = \operatorname{argmin} \frac{1}{N} \sum_{i=1}^{N} \|\boldsymbol{u}_i - \Phi(\boldsymbol{f}_i; \theta)\|_V^2 + \text{regularizations}$$

[13] C. Yun et al. (2020). "Are Transformers universal approximators of sequence-to-sequence functions?" In: *International Conference on Learning Representations*.

[14] A. Rahimi and B. Recht (2008). "Weighted sums of random kitchen sinks: Replacing minimization with randomization in learning". In: *Advances in neural information processing systems*.

Representational capcity and positional embedding

- PE plays an important role in shaping the representational capacity.
- Learnable PE[15], rotational-invariant PE[16], etc.

Theorem (Universal representater theorem (informal simplified version)[17])

Given fixed n and d, the function class of Transformers $\{u(\boldsymbol{y}) : u(\boldsymbol{y}) = g(\boldsymbol{y} + \boldsymbol{x}),$ *where* $g := g_L \circ \cdots \circ g_1\}$ *with the absolute fixed PE* $\boldsymbol{x}$ *is a universal approximator for continuous functions that map a compact domain in* $\mathbb{R}^{n \times d}$ *to* $\mathbb{R}^{n \times d}$.

Open problem: representational capacity

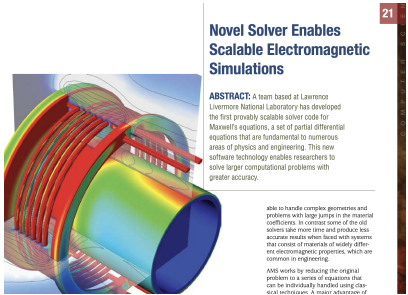
Can the theoretical results on the approximation capacity of Transformer with different PEs be reflected in specially designed experiments?

[15] J. Gehring et al. (2017). "Convolutional sequence to sequence learning". In: *International conference on machine learning*. PMLR.

[16] J. Su et al. (2021). "Roformer: Enhanced transformer with rotary position embedding". In: *arXiv preprint arXiv:2104.09864*.

[17] S. Luo et al. (2022). "Your Transformer May Not be as Powerful as You Expect". In: *Advances in Neural Information Processing Systems*. Ed. by A. H. Oh et al.

Inspiration: standing on the shoulder of giants

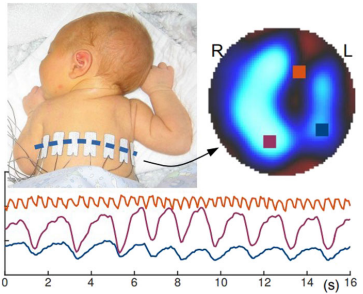


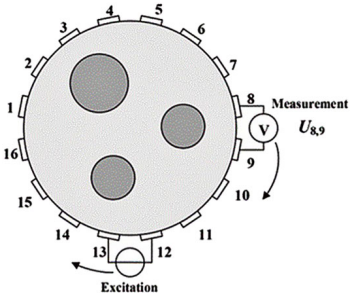
> “AMS is a perfect example of how fundamental mathematical research can lead to important software advances in high-performance computing.”

– On HX-preconditioner for Maxwell problems,
Report of The Panel on Recent Significant Advancements in Computational Science,
U.S. Department of Energy Office of Scientific and Technical Information, (2008).

Transformer Meets Boundary Value Inverse Problem

Electrical Impedance Tomography (EIT)

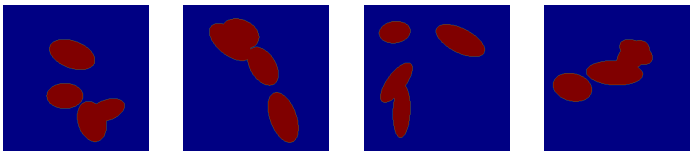




(Left)A 10-day-old infant with EIT electrodes[18]. By performing lung function imaging of newborns, timely diagnosis and treatment of lung diseases in early development of newborns without radiation damage can be done. (Right) Working principle of a 16 electrode system. Adjacent excitation is to select a pair of adjacent electrodes to input safe current, and then measure the output voltage between several pairs of adjacent electrodes except the excitation source.

[18] Y. Shi et al. (2021). "The research progress of electrical impedance tomography for lung monitoring". In: *Frontiers in Bioengineering and Biotechnology*.

EIT revisited



The forward model of EIT

$$\nabla \cdot (\sigma \nabla u) = 0 \quad \text{in } \Omega, \qquad \text{where } \sigma = \sigma_1 \text{ in } D, \text{ and } \sigma = \sigma_0 \text{ in } \Omega \backslash \overline{D}. \tag{$\star$}$$

- Current: $g = \sigma \nabla u \cdot \boldsymbol{n}|_{\partial\Omega}$ (Neumann boundary condition)
- Voltages: $f = u|_{\partial\Omega}$ (Dirichlet boundary condition)

Neumann-to-Dirichlet (NtD) mapping:

$$\Lambda_\sigma : H^{-1/2}(\partial\Omega) \to H^{1/2}(\partial\Omega), \quad g = \sigma \nabla u \cdot \boldsymbol{n}|_{\partial\Omega} \xrightarrow{\text{solve } (\star)} f = u|_{\partial\Omega}.$$

Inverse Problem of EIT

Forward and inverse operator

$$\mathcal{F} \ : \ \sigma \mapsto \Lambda_\sigma, \quad \text{and} \quad \mathcal{F}^{-1} \ : \ \Lambda_\sigma \mapsto \sigma.$$

- The measurement on $\partial\Omega$.
- The coefficient to be recovered.
- What we need (optimistically) is "knowing Λ_σ": for a set of basis $\{g_l\}_{l=1}^{\infty}$ of the corresponding Hilbert space, one can measure all the current-to-voltage pairs $\{g_l, f_l := \Lambda_\sigma g_l\}_{l=1}^{\infty}$ and construct the infinite dimensional matrix $\boldsymbol{A}_\sigma$.

$$\boldsymbol{f} = \boldsymbol{A}_\sigma \boldsymbol{g},$$

  where $\boldsymbol{g}$ and $\boldsymbol{f}$ are (infinite dimensional) vector representations of functions g and f.
- BCR-Net[19] is a DNN approximation of $\mathcal{F}^{-1}$ based on a large but finite sized matrix $\widetilde{\mathrm{A}}_\sigma$ as an accurate approximation to A_σ.

[19] Y. Fan and L. Ying (2020). "Solving electrical impedance tomography with deep learning". In: *Journal of Computational Physics*.

Inverse Problem of EIT

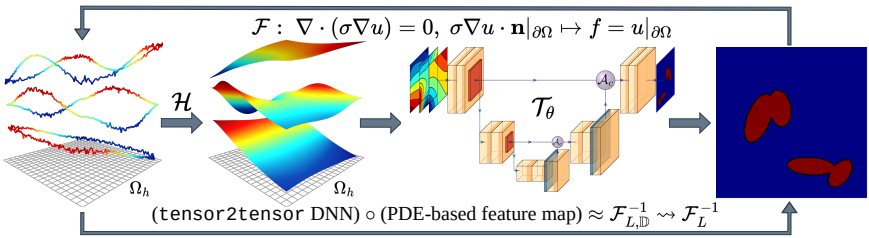
Forward and inverse operator with limited data pairs: use only a few data pairs $\{(g_l, f_l)\}_{l=1}^{L}$ for reconstruction.

$$\mathcal{F}_L : \sigma \mapsto \{(g_1, \Lambda_\sigma g_1), ..., (g_L, \Lambda_\sigma g_L)\} \quad \text{and} \quad \mathcal{F}_L^{-1} : \{(g_1, \Lambda_\sigma g_1), ..., (g_L, \Lambda_\sigma g_L)\} \mapsto \sigma.$$

- Extremely ill-posed or even not well-defined: the same boundary measurements may correspond to different σ.[20].
- For $\boldsymbol{g}_l = \boldsymbol{e}_l$, $l = 1, ..., L$, with $\boldsymbol{e}_l$ being unit vectors of a chosen basis, $(\boldsymbol{f}_1, ..., \boldsymbol{f}_L)$ only gives the first L columns of $\boldsymbol{A}_\sigma$.
- Restricting $\mathcal{F}_L^{-1}$ at a compact set of sampled data $\mathbb{D} := \{\sigma^{(k)}\}_{k=1}^{N}$.

[20] V. Isakov and J. Powell (1990). "On the inverse conductivity problem with one measurement". In: *Inverse Probl.*

From EIT to deep learning



- Learn an approximation to $\mathcal{F}_{L,\mathbb{D}}^{-1} : \{(g_1, \Lambda_{\sigma^{(k)}} g_1), ..., (g_L, \Lambda_{\sigma^{(k)}} g_L)\} \mapsto \sigma^{(k)}$.
- "Well-defined" enough as a high-dimensional interpolation (learning) problem on a compact data submanifold[21] with an end-to-end setting. Then generalization can be done for newly incoming σ's.
- The incomplete information of Λ_σ due to small L for one single σ is compensated by a large $N \gg 1$ sampling of different σ's.

[21] O. Ghattas and K. Willcox (2021). "Learning physics-based models from data: perspectives from inverse problems and model reduction". In: *Acta Numerica*.

From EIT to deep learning

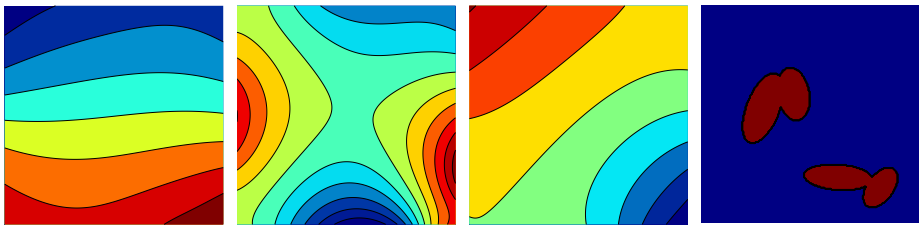
- What is an appropriate finite dimensional data format as inputs to the neural network?
- Is there a suitable neural network matching the mathematical structure?

Inspiration: direct sampling

Generate ϕ_l: the *harmonic extension* of $f_l - \Lambda_{\sigma_0} g_l$

$$\nabla \cdot (\sigma \nabla u) = 0 \quad \text{in } \Omega, \qquad \text{where } \sigma = \sigma_1 \text{ in } D, \text{ and } \sigma = \sigma_0 \text{ in } \Omega \backslash \overline{D}.$$

$$-\Delta \phi_l = 0 \quad \text{in} \quad \Omega, \quad \mathrm{n} \cdot \nabla \phi_l = (f_l - \Lambda_{\sigma_0} g_l) = (\Lambda_\sigma - \Lambda_{\sigma_0}) g_l \quad \text{on} \quad \partial\Omega, \quad \int_{\partial\Omega} \phi_l \mathrm{d}s = 0,$$



(a) (b) (c) (d)

EIT problem: (a)–(c) the input $\{\phi_l\}_{l=1}^{L}$ are harmonic extensions "features" for true σ (d).

Inspiration: direct sampling

Direct sampling method for EIT[22]: $f - \Lambda_{\sigma_0} g \to \phi \to \boldsymbol{d} \to \eta_x$.

$$\mathcal{I}_1^D(x) := \frac{\boldsymbol{d}(x) \cdot \nabla \phi(x)}{\|f - \Lambda_{\sigma_0} g\|_{L^2(\partial\Omega)} |\eta_x|_{H^s(\partial\Omega)}}.$$

where

$$-\Delta \eta_x = -\boldsymbol{d}(x) \cdot \nabla \delta_x \quad \text{in} \quad \Omega, \quad \boldsymbol{n} \cdot \nabla \eta_x = 0 \quad \text{on} \quad \partial\Omega, \quad \int_{\partial\Omega} \eta_x \mathrm{d}s = 0$$

- The empirical formula of $\mathcal{I}^D(x)$ can be written as an integral with Gaussian-like density, that attains maximum values for $x \in D$.
- The accuracy is much limited by some empirical choices of quantities such as the probing direction $\boldsymbol{d}(x)$ and $s = 3/2$.
- The this type of simple formula in direct sampling can be derived for only for a single data pair.

[22] Y. T. Chow, K. Ito, and J. Zou (2014). "A direct sampling method for electrical impedance tomography". In: *Inverse Probl.*

Operator learning: learn maps between function spaces

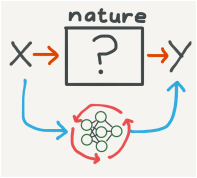
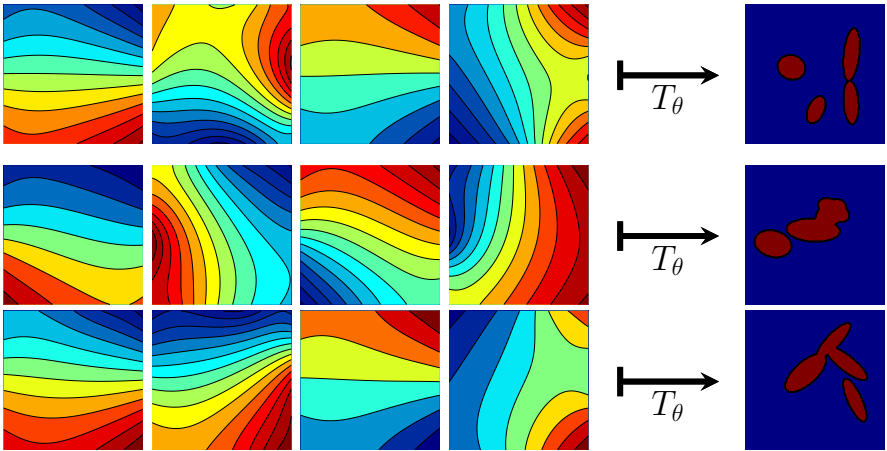


Figure courtesy of Jonty Sinai's blog: `https://jontysinai.github.io/`

Examples of the operator of interest T:

- Parametric PDE: given $a \in \mathcal{A}$ parameter and/or $f \in \mathcal{Y}$, find $u \in \mathcal{X}$ such that $\mathcal{L}_a(u) = f$. The operator T to be learned can be
  - The mapping between a varying parameter $a \in L^\infty$ to the solution $u \in H_0^1$ in $-\nabla \cdot (a \nabla u) = f$.
- Nonlinear initial value problem: given $u_0 \in \mathcal{H}$, find $u \in C([0, T]; \mathcal{H})$ such that $u_t + N(u) = 0$. The operator T to be learned can be
  - Direct inference from the initial condition to the solution at a much later time: $u_0(\cdot) \mapsto u(t_1, \cdot)$ with $t_1 \gg \Delta t$ (large time step).
- Boundary-value inverse problem:
  - From the PDE-based features ϕ in direct sampling to the index map $\mathcal{I}_D$.

Operator learning for EIT



More examples of direct sampling: (Ideal) NtD map Λ_σ's whole spectrum ($L = \infty$) can recover the inclusion σ with various interfaces. (Practice) "learn" a *single* parametrized operator T_θ that maps (a few, $L \leq 3$) harmonic extension features to reconstruct the inclusions.

From direct sampling to attention integral

- The global information of ϕ used as "keys" to locate a point x to probe.

$$\hat{\mathcal{I}}_1^D(x) := R(x) \int_\Omega \boldsymbol{d}(x) \cdot \mathcal{K}(x,y) \nabla\phi(y) \mathrm{d}y.$$

- The probing direction $\boldsymbol{d}(x)$ as "query" is assumed to depend globally on ϕ

$$\boldsymbol{d}(x) := \int_\Omega \mathcal{Q}(x,y) \nabla\phi(y) \mathrm{d}y.$$

  Choice of the probing direction in direct sampling[23]: If $\mathcal{Q}(x,y) = \delta_x(y)/\|\nabla\phi(x)\|$, then $\boldsymbol{d}(x) = \nabla\phi(x)/\|\nabla\phi(x)\|$.

- In $R(x)$, $|\cdot|_Y$ is assumed to be $|\eta_x|_Y^2 := (\mathcal{V}\eta_x, \eta_x)_{L^2(\partial\Omega)}$, where η_x is the potential using the probing as source. If $\mathcal{V}$ induces a Gaussian-like kernel which the attention kernel does induce[24], the index function can achieve maximum values for points inside D.

[23] M. Ikehata (2000). "Reconstruction of the support function for inclusion from boundary measurements". In: *Journal of Inverse and Ill-posed Problems*.

[24] H. Peng et al. (2021b). "Random Feature Attention". In: *International Conference on Learning Representations*.

"Attention is all we need"?

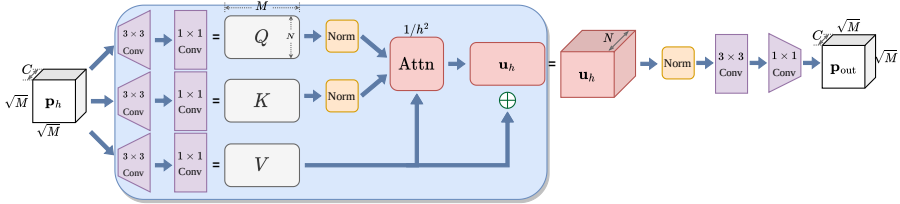


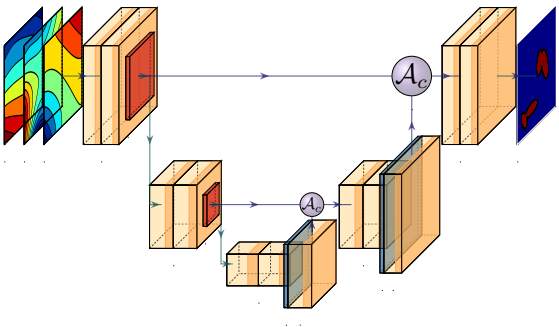
Figure: Schematics of a modified attention layer of the Transformer-based operator learner.

- Positional embedding: At each resolution, The 2D $\sqrt{M} \times \sqrt{M}$ Cartesian grid.
- ResNet `DoubleConv`: The double convolution block is modified[25] from that commonly seen in Computer Vision CNN[26].
- The "interaction" (attention matrix) between different latent representations can be computed using coarse latent representations.

[25] Z. Liu, H. Mao, et al. (2022). "A convnet for the 2020s". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

[26] K. He et al. (2016). "Deep residual learning for image recognition". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*.

Cross-attention $\times$ U-Net[27][28]



- Input: the concatenation of discretizations of ϕ and $\nabla\phi$.
- Output: the approximation to the index map $\mathcal{I}^D$.
- : 3×3 convolution + ReLU;
- : normalization;
- : interpolation;
- : cross attention from the coarse grid to the fine grid;
- : input and output discretized functions.

[27] O. Ronneberger, P. Fischer, and T. Brox (2015). "U-net: Convolutional networks for biomedical image segmentation". In: *International Conference on Medical image computing and computer-assisted intervention*. Springer.

[28] R. Guo and J. Jiang (2021). "Construct Deep Neural Networks based on Direct Sampling Methods for Solving Electrical Impedance Tomography". In: *SIAM Journal on Scientific Computing*.

An architectural advantage of $(QK^\top)V$

Theorem (Frequency bootstrapping (simplified informal 1D version)[29])

Suppose there exists a channel l in the current latent representation such that $(V_i)_l = \sin(az_i)$ for some $a \in \mathbb{Z}^+$, the current finite-channel sum attention kernel approximates a "nice" kernel to an error of $O(\epsilon)$ with only "lower frequency" modes. Then, there exists a set of weights such that certain channel k' in the output of the attention layer approximates $\sin(a'z)$, $\mathbb{Z}^+ \ni a' > a$ with comparable error.

- Heuristics: multiplicative neural architecture can use data-driven basis functions to characterize operators.

$$u_l(z) = h^2 \sum_{x \in \mathcal{M}} \big(q(z) \cdot k(x)\big) v_l(x)\, \delta_x \approx \int_\Omega \kappa_\theta(z, x) v_l(x)\, d\mu(x).$$

- Proof: use the tools of Pincherle-Goursat (degenerate) kernels for $\kappa_\theta(z, x; v) = \sum_{l=1}^N q_l(x; v) k_l(z; v)$.

[29] R. Guo, S. Cao, and L. Chen (2023). "Transformer Meets Boundary Value Inverse Problems". In: *The Eleventh International Conference on Learning Representations (ICLR)*

Electrical impedance tomography (EIT)

- Noise: $\xi = \xi(x)$ is assumed to be $\xi(x) = (f(x) - \Lambda_{\sigma_0} g(x))\tau G(x)$ where τ specifies the percentage of noise, and $G(x)$ is a Gaussian distribution.
- Train-test: train 10800 samples, test 2000 samples. 50 epochs of `1CYCLE`+ `ADAMW`.

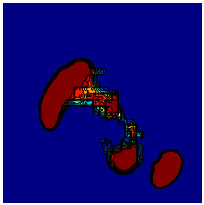
	Relative L^2 error			Position-wise cross entropy			Dice coefficient			# params
	$\tau = 0$	$\tau = 0.05$	$\tau = 0.2$	$\tau = 0$	$\tau = 0.05$	$\tau = 0.2$	$\tau = 0$	$\tau = 0.05$	$\tau = 0.2$	
U-Net	0.200	0.341	0.366	0.0836	0.132	0.143	0.845	0.810	0.799	7.7m
FNO2d[30]	0.318	0.492	0.502	0.396	0.467	0.508	0.650	0.592	0.582	10.4m
Hybrid UT[31]	0.185	0.320	0.333	0.0785	0.112	0.116	0.877	0.829	0.821	11.9m
Cross-Attention UT[32]	0.171	0.305	0.311	0.0619	0.105	0.109	0.887	0.840	0.829	11.4m
U-Net+Coarse Attn	0.184	0.343	0.360	0.0801	0.136	0.147	0.852	0.807	0.804	8.4m
UIT (ours)	**0.163**	**0.261**	**0.272**	**0.0564**	**0.0967**	**0.0981**	**0.897**	**0.858**	**0.845**	11.4m
UIT+(L=3) (ours)	**0.147**	**0.250**	**0.254**	**0.0471**	**0.0882**	**0.0900**	**0.914**	**0.891**	**0.880**	11.4m

[30] Z. Li et al. (2021). "Fourier Neural Operator for Parametric Partial Differential Equations". In: *The Ninth International Conference on Learning Representations (ICLR)*.

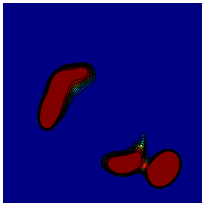
[31] Y. Gao, M. Zhou, and D. N. Metaxas (2021). "UTNet: a hybrid transformer architecture for medical image segmentation". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer.

[32] H. Wang et al. (2022). "Mixed transformer u-net for medical image segmentation". In: *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE.

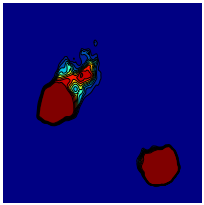
Reconstruction for unseen samples



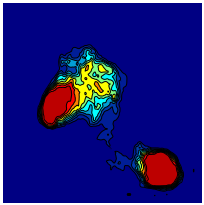
(a) U-Net (7.7m) $L = 1$



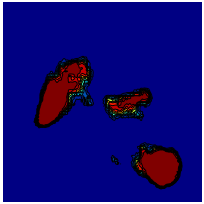
(b) U-Net (33m) $L = 3$



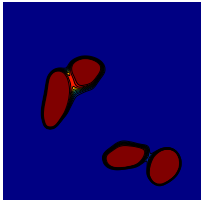
(c) FNO (10.4m) $L = 1$



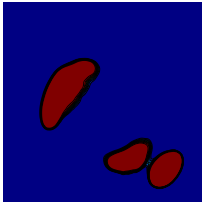
(d) Adaptive FNO (10.7m) $L = 1$



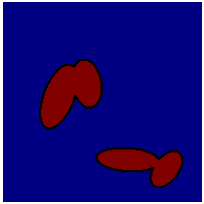
(e) Multiwavelet NO (9.8m) $L = 1$



(f) Hybrid UT (10.1m) $L = 1$



(g) UIT (11.4m) $L = 1$



(h) Ground truth inclusion

Acknowledgments

- The organizing committee and all the helpers.
- Fel., Hon., Prof., Dr. Jinchao Xu for the delivery of enlightening lectures.
- Dr. Jun Zou (CUHK) and Dr. Bangti Jin (CUHK) for the comments on direct sampling methods, Dr. Jingrong Wei for proof-reading our paper.
- Source codes to replicate the experiments are available at
  - `https://github.com/scaomath/galerkin-transformer`
  - `https://github.com/scaomath/eit-transformer`

This research is supported in part by NSF awards DMS-2136075 and DMS-2309778.