On the geometry of large transformers representations

Alberto Cazzaniga
alberto.cazzaniga@areasciencepark.it

Alessio Ansuini
alessio.ansuini@areasciencepark.it

AREA Science Park,
Institute for Research and Innovation Technology

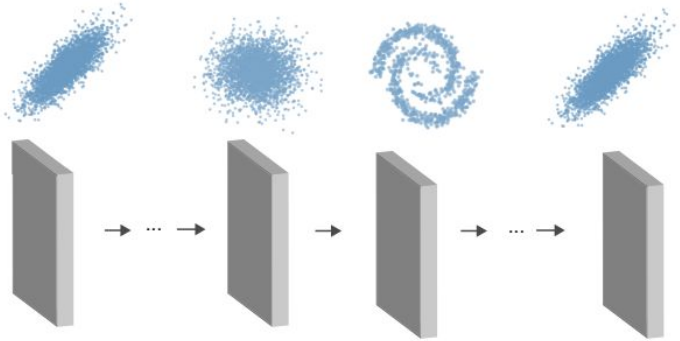
DTU SUMMER SCHOOL

Advanced Topics in Machine Learning

22/08/2023

Motivation

Data representations in neural networks undergo profound changes across the layers.



Q.: can changes in the geometry of the data cloud explain the rise of meaningful features in large transformer models?

L. Valeriani, F. Cuturello, A. Ansuini, A. Cazzaniga, NIPS Workshop MLSB, 2022
https://doi.org/10.1101/2022.10.24.513504

L.Valeriani, D. Doimo, F. Cuturello, A. Laio, A. Ansuini, A. Cazzaniga
https://arxiv.org/abs/2302.00294

“Beware reductionism”

Christopher Manning @chrmanning · Aug 2

Reflecting again on how knowing all the architecture & equations of the Transformer model is really of no use at all in convincingly explaining to someone how an LLM like ChatGPT can write paragraphs of lucid text in response to a prompt.

I guess I’m saying “Beware reductionism”.

#geometric_invariants

Outline

geometry
- *global geometry*: Intrinsic Dimension
- *local geometry*: Neighborood Overlap

large transformers
- self-supervision and self-attention
  - transformers for protein sequences
  - transformers for image generation

Q.: can changes in the geometry of the data cloud explain the rise of meaningful in large transformers?

rise of meaningful features
- local and global geometries are related
- synthetic representations are the most semantically meaningful

what about natural language???
- Llama-2 representations

Self-supervised tasks (from language games)

Masked Language Modelling
Game: fill in the blanks

I bring my _____ to the park.

I bring my _____ to the park to run.

I bring my _____ to the park to play football.

I bring my _____ to the park to play with other puppies.

J. Devlin et al.,BERT, 2018

Next-Token Prediction
Game: predict next word

I

I bring

...

I bring my book to the park to

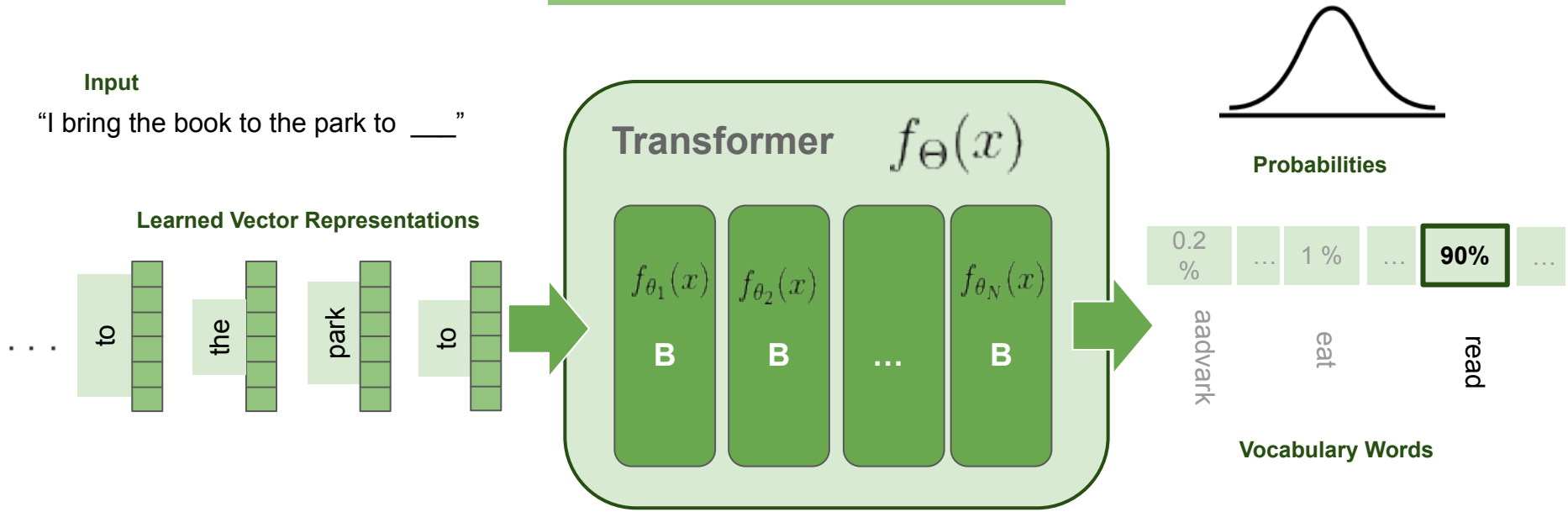
I bring my book to the park to read

Y. Bengio et al., A Neural Probabilistic LM, 2003

'A word is characterized by the company it keeps'

John R. Firth, “A synopsis of linguistic theory”, 1957

Transformer models



Self-supervision allows training on VERY large datasets.

Datapoints transformed by structurally identical blocks in $x^{(i)} \in \mathbb{R}^{L \times D}$

$$x_{in} = x^{(0)} \rightarrow x^{(1)} = f_{\theta_1}(x^{(0)}) \rightarrow \ldots \rightarrow x^{(N)} = f_{\theta_N}(x^{(N-1)})$$

Crucial innovation: (causal) self-attention

Classical NN transformation

Input is a **single vector**

$$x$$

The basic transformation is **absolute**

$$\phi(w \cdot x)$$

Self-attention transformation

Input is a **set (frame) of vectors**

$$x = (x_1, x_2, \ldots)$$

The basic transformation is **relative:**

- Compute how relevant is i-th word for others

$$\alpha_i = \frac{\exp(x \cdot x_i)}{Z}$$

- New value is **relative** to other words

$$x \longmapsto \sum_i \alpha_i x_i$$

A. Vaswani, “Attention is all you need”, 2016

Crucial innovation: (causal) self-attention

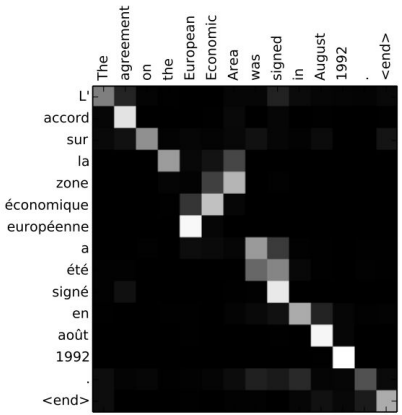
Classical NN transformation

Input is a **single vector**

$$x$$

The basic transformation is **absolute**

$$\phi(w \cdot x)$$



Self-attention transformation

Input is a **set (frame) of vectors**

$$x = (x_1, x_2, \ldots)$$

The basic transformation is **relative:**

- Compute how relevant is i-th word for others

$$\alpha_i = \frac{\exp((xW_q)(x_iW_k)^{\mathrm{T}})}{Z}$$

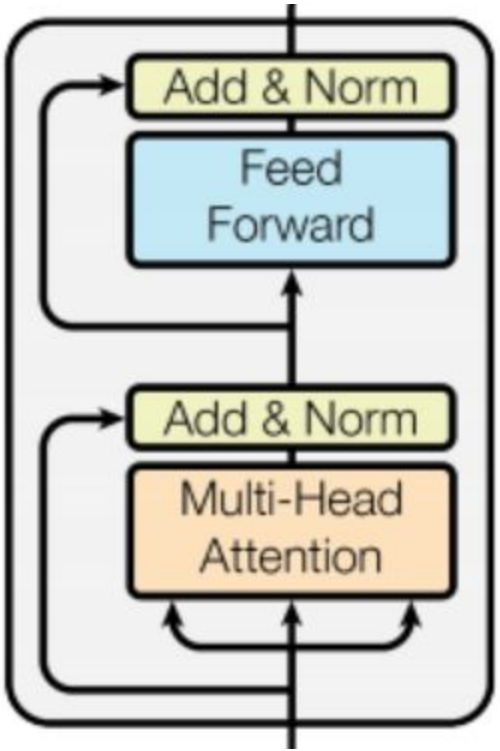
- New value is **relative** to other words

$$x \to \sum_i \alpha_i (x_i W_v)$$

A. Vaswani, “Attention is all you need”, 2016

Self-attention block

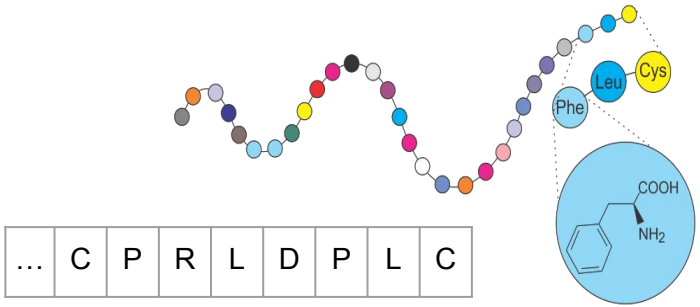
- apply several self-attention (SA) maps in parallel (multi-head SA)
- two layer MLP transforms each token (with shared parameters)
- residual connections are applied after each sub-component
- normalisation (along feature dimension) applied after any transformation



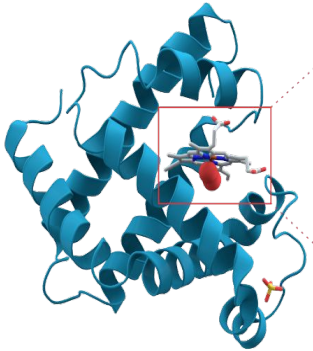
E. Turner, “An Introduction to Transformers”, arXiv:2304.10557, 2023

Why transformers for proteins?

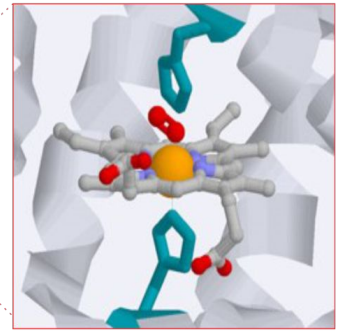
Sequence



Structure



Biological Function



Many known protein sequences (more than 200M)

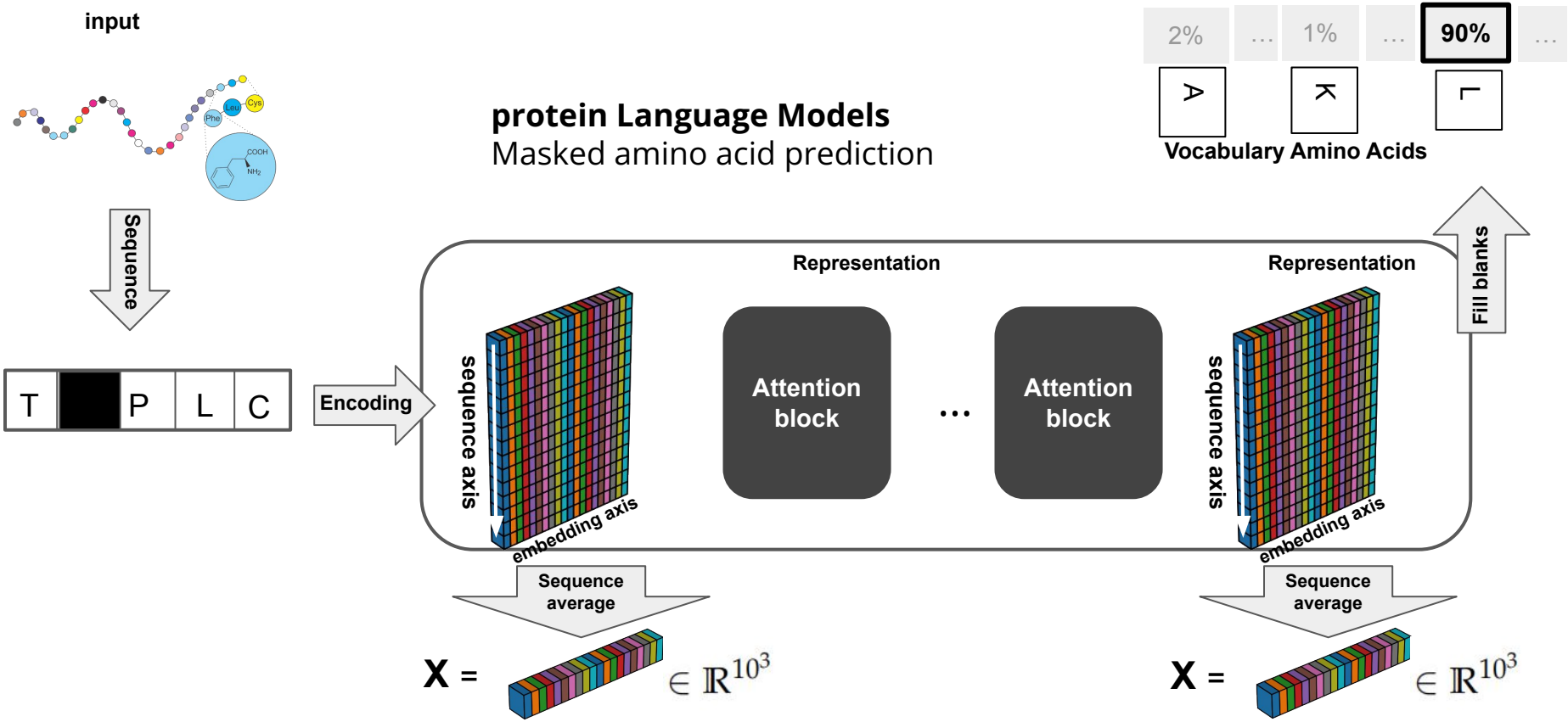
Few experimentally measured protein structures (approx. 200K)

UniProt

NCBI MGnify

RCSB PDB PROTEIN DATA BANK

Transformers models for protein sequences

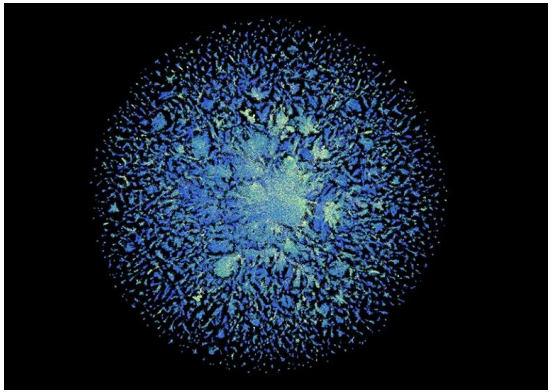


A. Rives et al., Biological structure and function emerges from scaling unsupervised learning, 2021

ESM2 model family

Evolutionary Scale Modelling

Representations give *meaningful* cartography of the protein universe



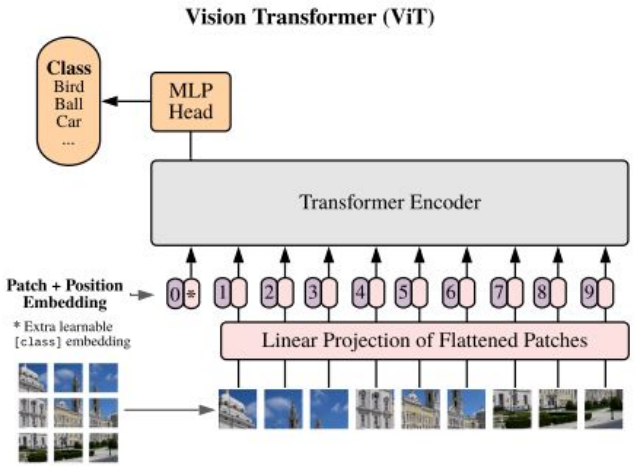
Great playground for scale analysis

Model	#Blocks	Emb. dim.	#Heads	#Params
ESM-2(8M)	6	320	20	8M
ESM-2(35M)	12	480	20	35M
ESM-2(150M)	30	640	20	150M
ESM-2(650M)	33	1280	20	650M
ESM-2(3B)	36	2560	40	3B
ESM-2(15B)	48	5120	40	15B

Z. Lin et al., Language models of protein sequences at the scale of evolution enable accurate structure prediction, Science, 2022

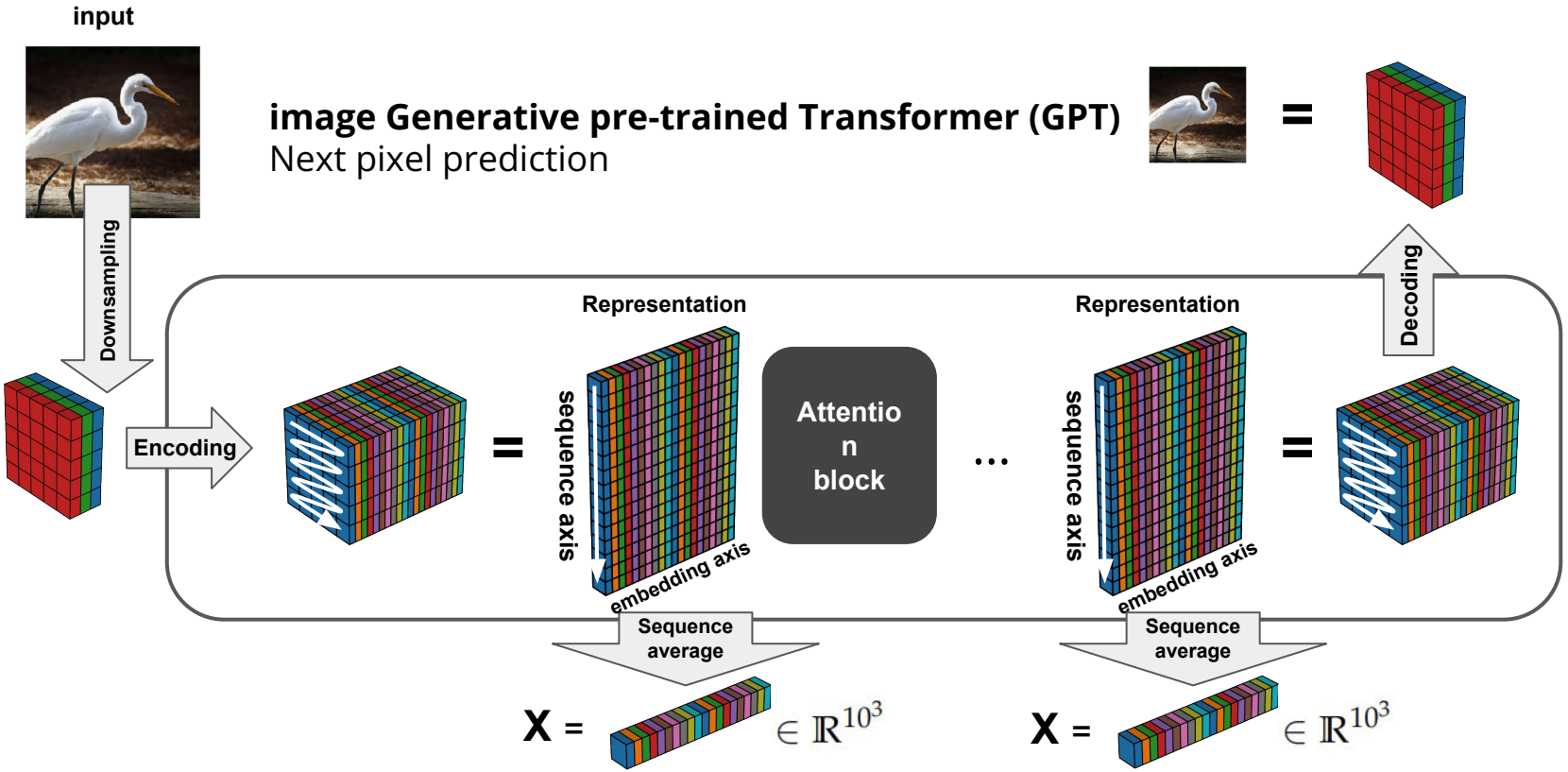
Why transformers for images?

- classify/construct meaningful representations of images without convolutional structure
- different approach to “locality” through patches
- attention mechanism detects mutual relations between different local properties
- leverage large amount of unsupervised or weakly supervised data efficiently



Alexander Kolesnikov et al., Transformers for Image Recognition at Scale, ICLR, 2021

Transformer models for image generation



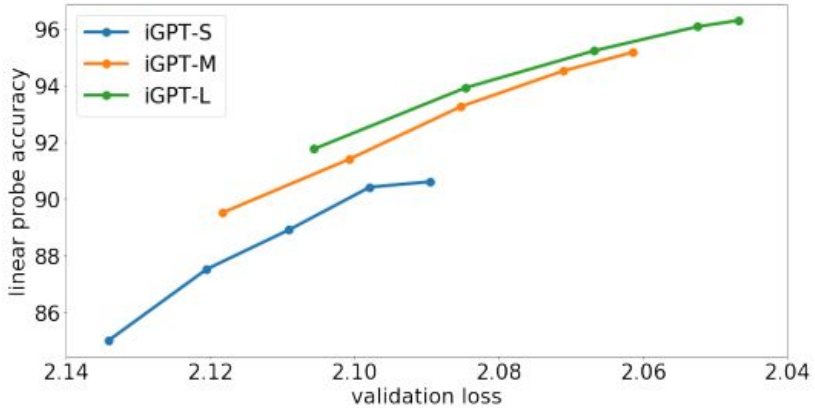
M. Chen et al., “Generative Pretraining from Pixels”, 2020

iGPT family

Generative pretraining gives rise to *meaningful* features

- SOTA on unsupervised transfer for CIFAR-10 and CIFAR-100

Model	#Blocks	Emb. dim.	#Params
iGPT-XL	60	3072	6.8B
iGPT-L	48	1536	1.4B
iGPT-M	36	1024	455M
iGPT-S	24	512	76M



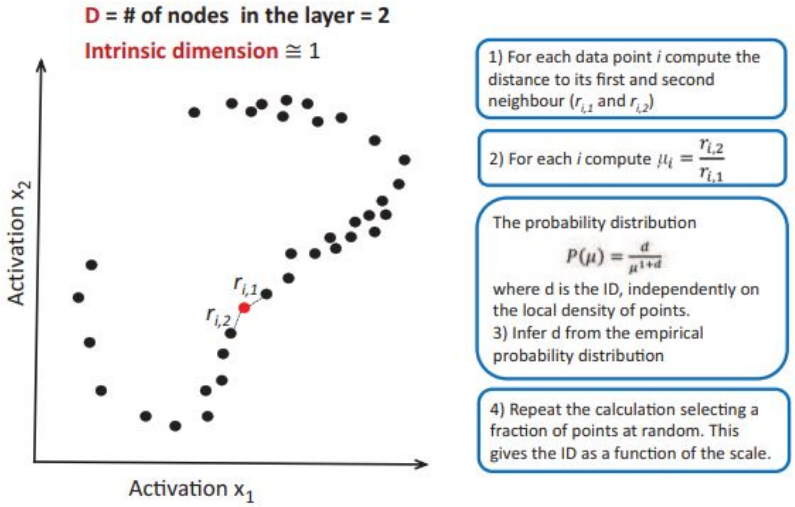
- Weights for different stages of training are available
- Models trained also with BERT-type loss

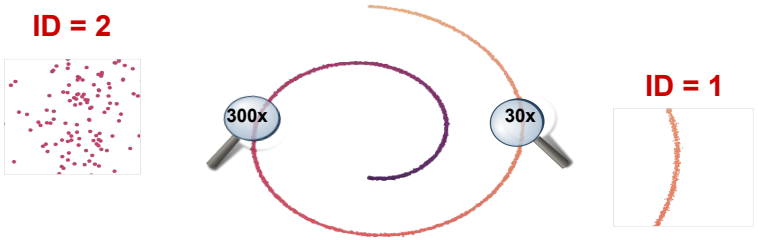
M. Chen et al., “Generative Pretraining from Pixels”, 2020

Global geometry: Intrinsic Dimension (ID)

Manifold hypothesis: many datasets in high dimension lie close to low dimensional manifolds.

Estimate the **intrinsic dimension** of the embedded manifold approximating the data.





E. Facco et al., 2017

Some care needed:

- scale matters, need to look for **persistent** phenomena
- tends to underestimate for large ID
- needs approx. locally constant density

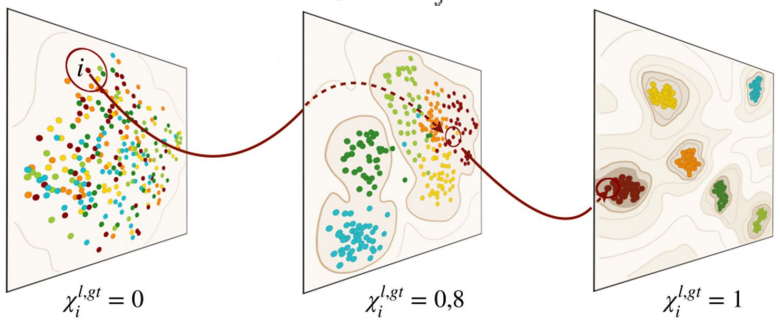
Local geometry: Neighborood Overlap (NO)

How to compare data clouds in different layers?
(Think of representations of transformers at different blocks l and m)

$$\chi_k^{l,m} = \frac{1}{N} \sum_i \frac{1}{k} \sum_j A^l{}_{ij} A^m_{ij}$$

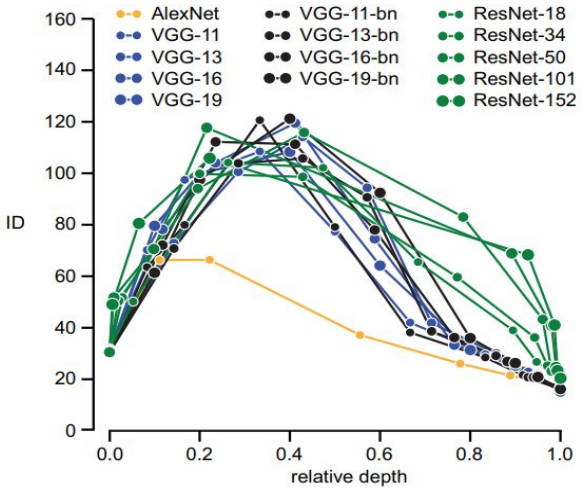
When given discrete labels: how many neighbors are in the same class?

$$\chi_k^{l,gt} = \frac{1}{N} \sum_i \frac{1}{k} \sum_j A^l_{ij} A^{gt}_{ij}$$



D. Doimo et al., “Hierarchical nucleation in deep neural networks”, NIPS, 2020

Geometry of NN for image classification

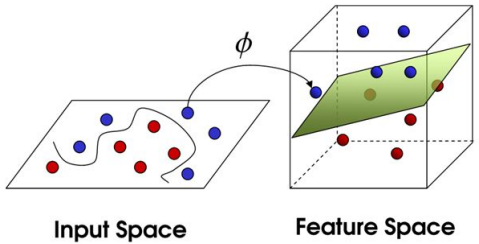


Global geometry:

- hunchback shape
- more compression implies better classification

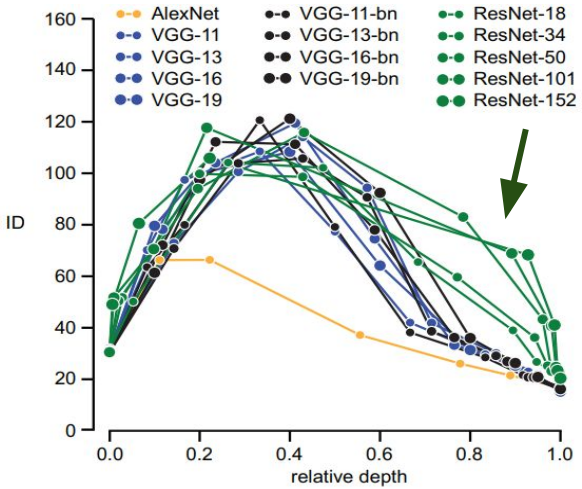
Kernel trick

Cover's Theorem (1965): In sufficiently high dimension any dataset is *linearly separable*



A. Ansuini et al., "Intrinsic dimension of data representations in deep neural networks", NIPS, 2019

Geometry of NN for image classification

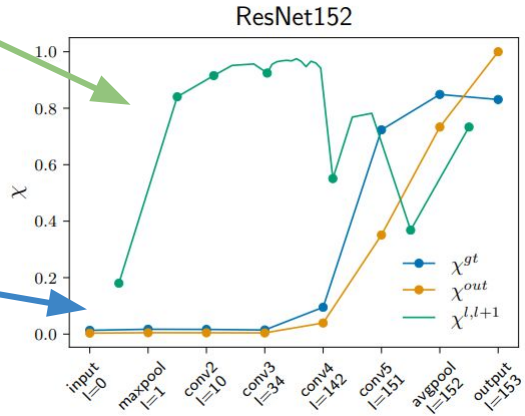


Global geometry:
- hunchback shape
- more compression implies better classification

Local geometry:
- neighborhood overlap with class high only in last layers
- Local and global geometry not as related as expected

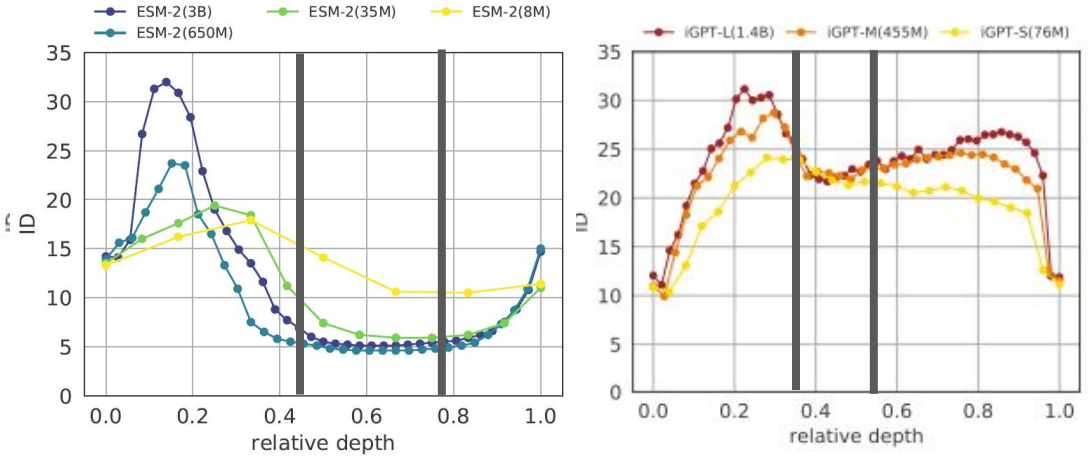
Consecutive layers

Overlap with ImageNet labels



Global geometry of transformers

ID of transformers representation has a three phased behaviour.
Analogy with a sophisticated autoencoder (NOT REQUIRED).

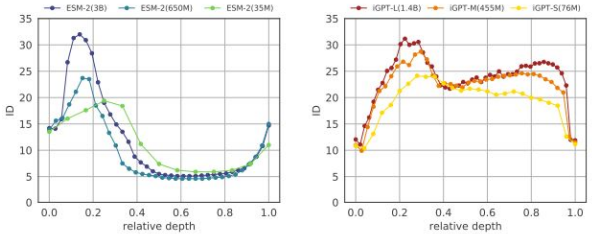


protein Language Models

Image GPT

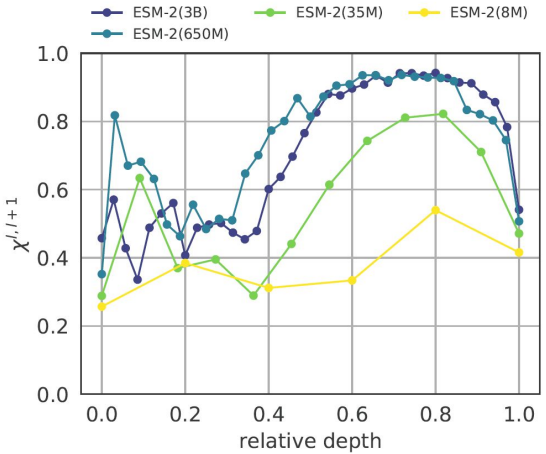
Local and global geometry of transformers

Local and global geometry are tightly related.



Measure of neighborhood overlap for consecutive blocks.

protein LM



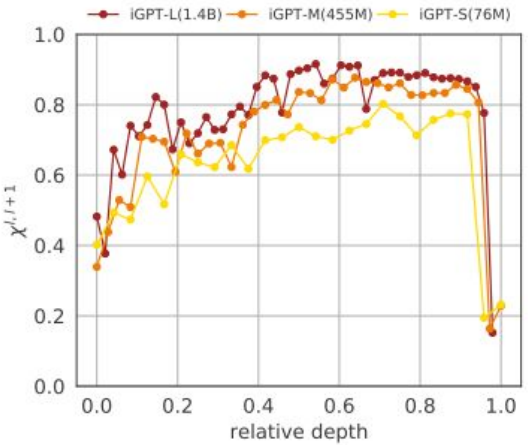
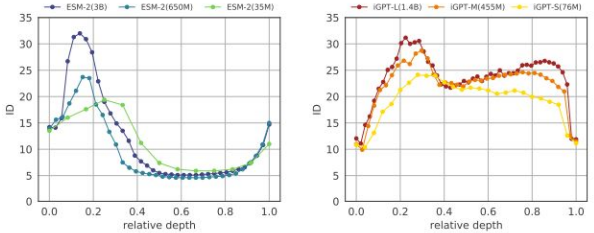


Image GPT

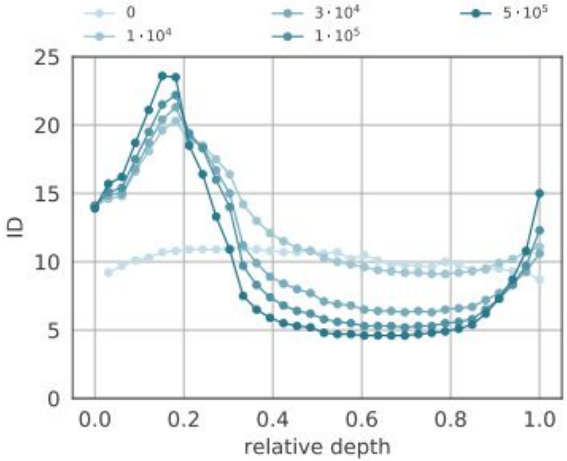
Global geometry during training

Three phased behaviour emerges hierarchically during training.



ID curve at various stages of training.

protein LM



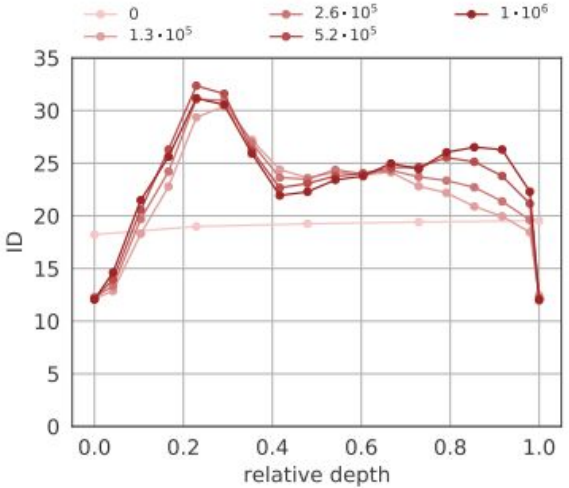
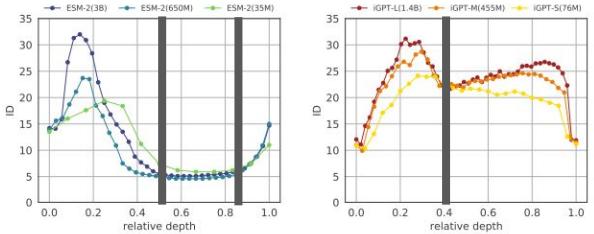


Image GPT

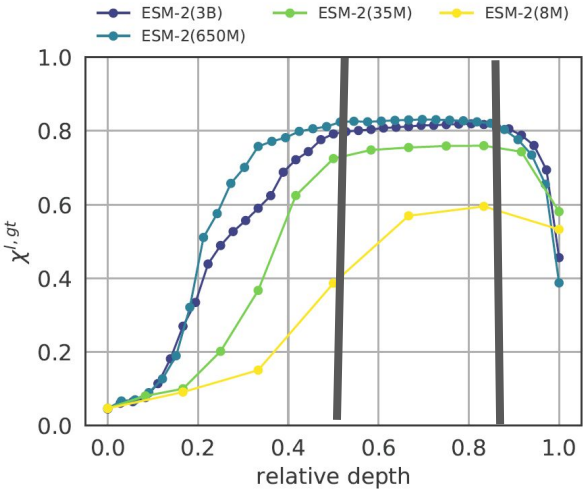
Lower ID implies more meaningful features

Intermediate low dimensional representations are the most semantically rich.



Neighborhood overlap with “fold class” (pLMs) and “ImageNet class” (iGPT).

protein LM



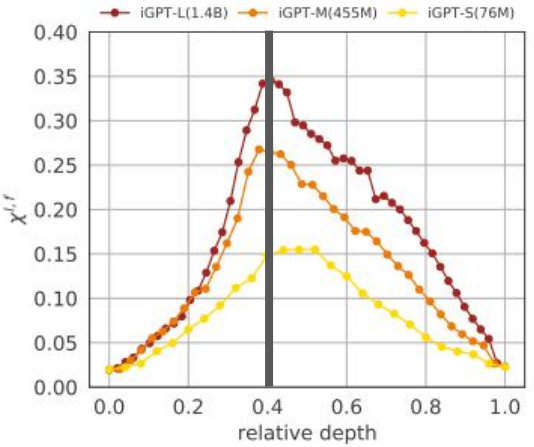


Image GPT

What about natural language???

LLama-2(70B) ~ GPT-3

Still no sign of saturation

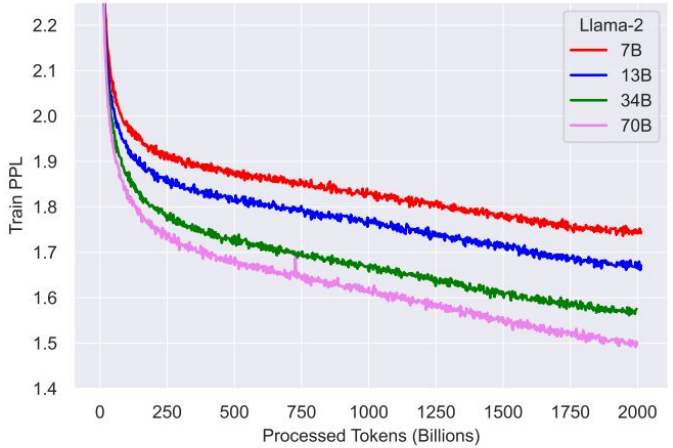
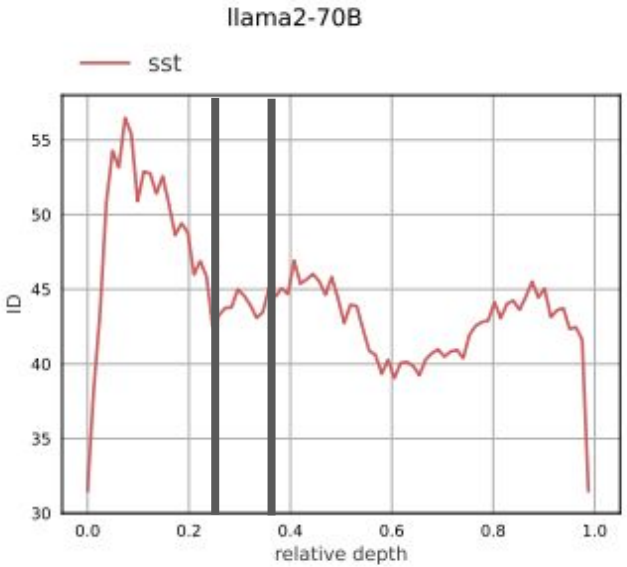


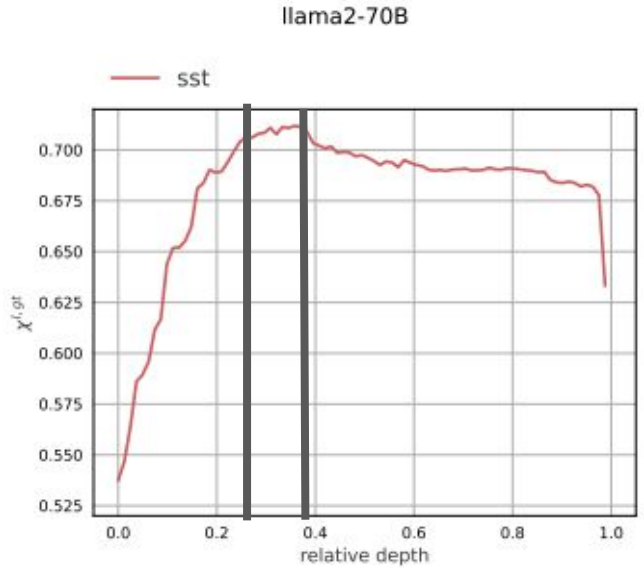
Figure 5: Training Loss for Llama 2 models. We compare the training loss of the Llama 2 family of models. We observe that after pretraining on 2T Tokens, the models still did not show any sign of saturation.

H. Touvron, Llama 2: Open Foundation and Fine-Tuned Chat Models, 2023

Llama reps: global geometry and semantic

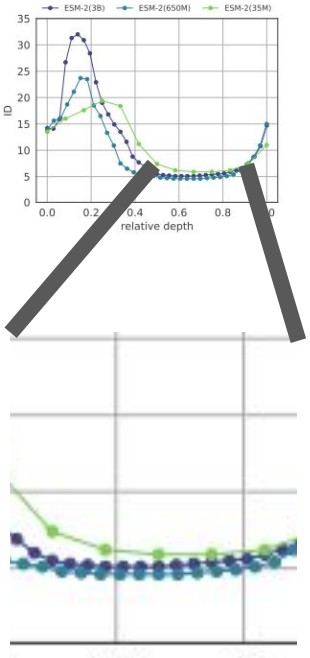
Emergence of double plateau ID-profile
First plateau is rich in semantic information



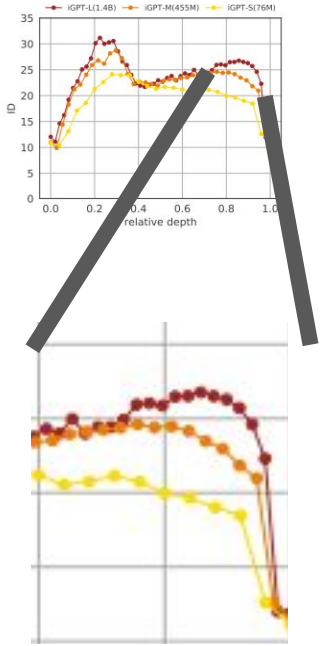


Future directions

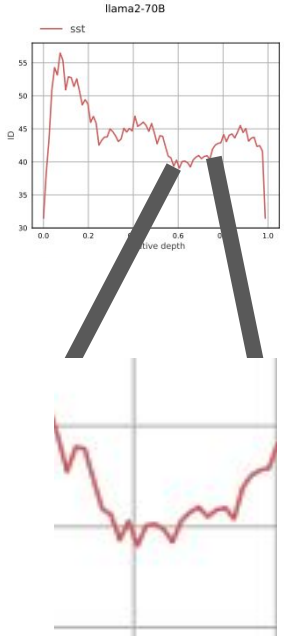
Zoom on subtle phenomena



Which degree of universality?



Meaning of second peak?



Meaning of second relative minumum?

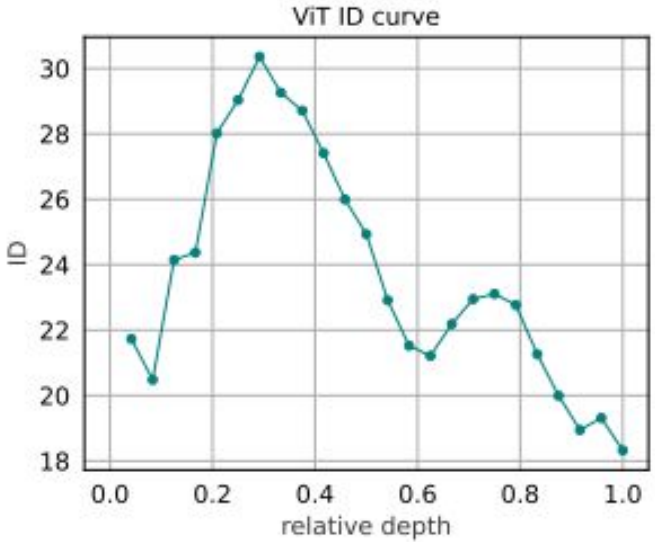
Thanks for the attention




Further Results

Self-supervision is crucial for three-phased behaviour

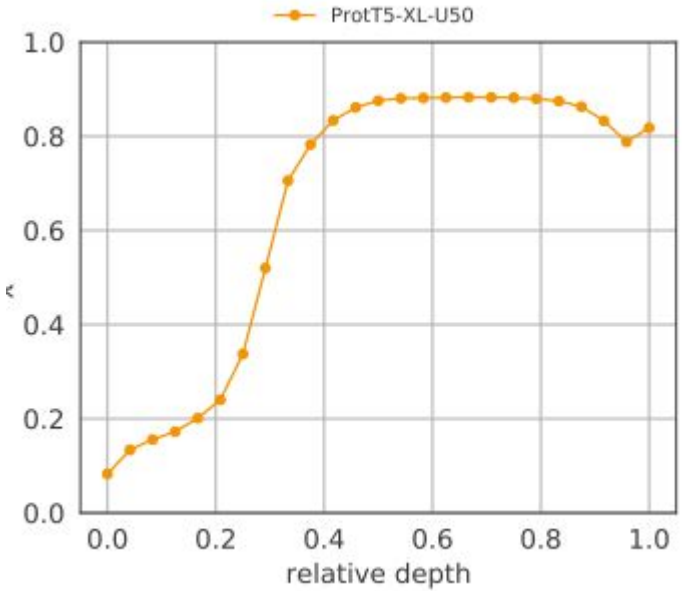
Vision Transformer (ViT) trained to classify Imagenet-1k does not exhibit a clear three-phased beaviour.



Much less pronounced second peak
Final ID does not match input ID

Application

Nearest neighbor search in plateau layers improves identification of protein relations.



Looking for closest match (homolog) in plateau layers improves performance of 6% for free.

Results are robust wrt the choice of k