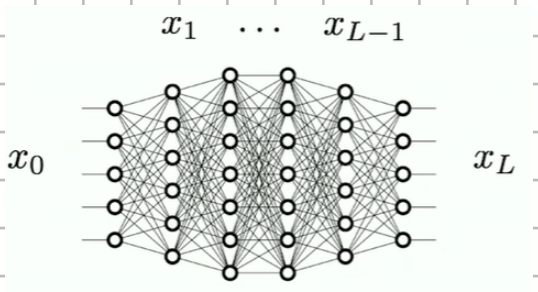
Resnet



$x_{k+1} = x_k + \sigma(W_k x_k + b_k)$ $\quad k = 1, 2, \cdots, L-1$

Nonlinear function, eg. $\max(\cdot, 0)$ — Affine function, $f(x) = ax + b$

NN = Map $x_0 \to x_L$

continuous time.

$\dot{x}(t) = \sigma(W_t x(t) + b_t)$

direction change

NN = flow map $x(0) \to x(T)$

Transformer

transformer input $\{x_1, x_2, \cdots x_n\} \in \mathbb{R}^d$: tokens have position encoding.

$\{x_1, x_2, \cdots, x_n\} \Leftrightarrow \frac{1}{n}\sum_{i=1}^{n} \delta_{x_i}$

prompt = input = probability measure on tokens ($P(\Omega) = 1$) $\Rightarrow$ likelihood of token being next

transformer output: predict the next token

output = probability measure on tokens $\Rightarrow$ empirical dist. of tokens in prompt

flow map

transformer = flow map f_θ: $\mu(0) \to \mu(T)$

$P(\mathbb{R}^d) \to P(\mathbb{R}^d)$

$\mu(0) = \frac{1}{n}\sum_{i=1}^{n} \delta_{x_i}$ $\Rightarrow$ δ: dot indicator, indicates that there is a token at this coordinate

Transformer is a "Mean-field interacting partical (token) system"

$\dot{x}_i(t) = X_t[\mu(t)](x_i(t))$ $\quad i = 1, 2, \cdots, n$ — depends on itself

rate of change — mean-field: depends on all of the other tokens' aggregate distribution

Continuity Equation

Divergence: { div < 0: gather; div > 0: diffusion }

$$\partial_t \mu(t) + \operatorname{div}(\mu(t) X_t[\mu(t)]) = 0$$

density — movement speed, depends on $\mu(t)$

self-attention dynamics

$\dot{x}_i(t) = X_t[\mu(t)](x_i(t))$

Self-attention dynamics = special choice of $X_t[\mu(t)](\cdot)$

$$X_t[\mu(t)](x) = V_t \frac{\int y\, e^{\langle Q_t x,\, K_t y\rangle} \mu(t)(dy)}{\int e^{\langle Q_t x,\, K_t y\rangle} \mu(t)(dy)} = V_t \mathbb{E}_{\check{\mu}_x}[Y_t]$$

V_t, Q_t, K_t: $d \times d$ matrices learned during traing.

Q_t (Query): What I want — eg. "I" wants a verb

K_t (key): my characteristics — eg. "I" is a pronoun

V_t (Value): information package — eg: "river": flowing, moist, blue, ...

$\langle \cdot, \cdot \rangle$: inner product, indicating similarit

$\langle Q_t x, K_t y \rangle$ means that the more the features of y ($K_t y$) match what x wants ($Q_t x$), the larger $\langle Q_t x, K_t y \rangle$ is

$$X_t[\mu(t)](x) = V_t \frac{\int y\, e^{\langle Q_t x,\, K_t y\rangle} \mu(t)(dy)}{\int e^{\langle Q_t x,\, K_t y\rangle} \mu(t)(dy)} = V_t \mathbb{E}_{\check{\mu}_x}[Y_t]$$

$\check{\mu}_x$ → attention-weighted density

Since $\mu(t) = \frac{1}{n}\sum_{i=1}^{n} \delta_{x_i(t)}$

$$X_t[\mu(t)](x_i(t)) = V_t\left(\sum_{j=1}^{n} P_{ij}(t)\, x_j(t)\right)$$

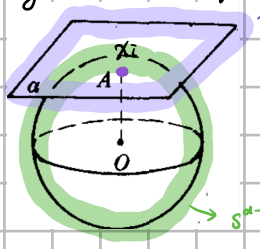
$$P_{ij}(t) = \frac{e^{\langle Q_t x_i(t),\, K_t x_j(t)\rangle}}{\sum_{k=1}^{n} e^{\langle Q_t x_i(t),\, K_t x_k(t)\rangle}}$$

Layer Normalization

LayerNorm (prevents explosion)

Each self-attention involves a weighted sum of a large number of vectors, causing the magnitude of the embedding vectors to increase (gradient exploring) or decrease (gradient vanishing) sharply between layers.

Dynamics on the sphere:



tangent plane of x_i

original self-attention speed (F)

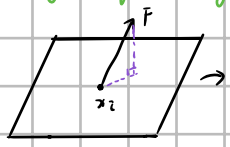
$$\dot{x}_i(t) = \mathrm{Proj}_{T_{x_i(t)}S^{d-1}} V_t\left(\sum_{j=1}^{n} P_{ij}(t)\, x_j(t)\right)$$

correction operation

tangent plane: $T_{x_i} S^{d-1}$

S^{d-1}: region representing vectors with a fixed magnitude.

Proj:



$F(t)$ → F' $\mathrm{Proj}\, F = F'$ → tangent plane of x_i

Bells And Whistles

MLP layer in-between two self-attention layers

$$\dot{x}_i(t) = \mathrm{Proj}_{T_{x_i(t)}S^{d-1}} \mathrm{MLP}_t \circ V_t\left(\sum_{j=1}^{n} P_{ij}(t)\, x_j(t)\right)$$

Multiple heads

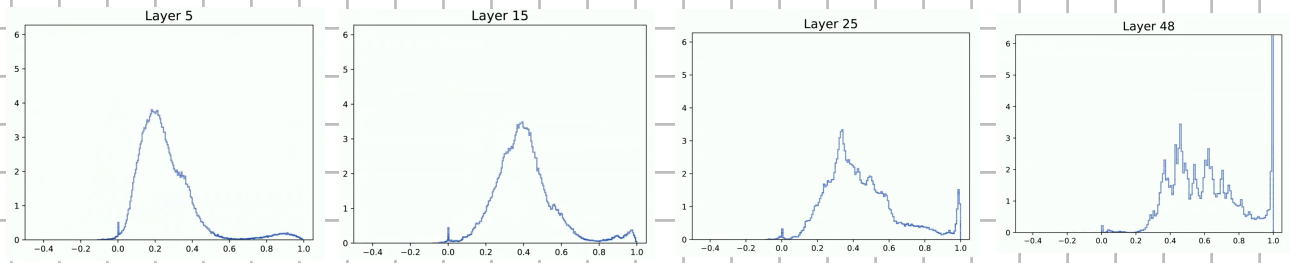
$$\dot{x}_i(t) = \frac{1}{K}\sum_{k=1}^{K} \mathrm{Proj}_{T_{x_i(t)}S^{d-1}} \mathrm{MLP}_t^{[k]} \circ V_t^{[k]}\left(\sum_{j=1}^{n} P_{ij}^{[k]}(t)\, x_j(t)\right)$$

Long-Time Asymptotics

$$\dot{x}_i(t) = \mathrm{Proj}_{T_{x_i(t)}S^{d-1}} V_t\left(\sum_{j=1}^{n} P_{ij}(t)\, x_j(t)\right) \quad \text{when } t \to \infty \quad \mu(t) \to \text{clustering}$$

e.g. $\mu(t)$ can be confusing: "bank" may be in the financial and geographical zone.

After multiple layers ($t\uparrow$), all tokens representing 'finance' converge into a financial cluster, while all tokens representing 'geography' converge into a geographical cluster



A Simple Attention Model

$$\dot{x}_i(t) = \mathrm{Proj}_{T_{x_i(t)}S^{d-1}} V_t\left(\sum_{j=1}^{n} P_{ij}(t)\, x_j(t)\right) \qquad P_{ij}(t) = \frac{e^{x_i(t)^T Q_t^T K_t x_j(t)}}{\sum_{\ell=1}^{n} e^{x_i(t)^T Q_t^T K_t x_\ell(t)}}$$

$\Rightarrow$ Simplify: $V_t = Id \quad Q_t^T K_t = \beta Id$

$$\dot{x}_i(t) = \mathrm{Proj}_{T_{x_i(t)}S^{d-1}} \frac{\sum_{j=1}^{n} e^{\beta\langle x_i(t),\, x_j(t)\rangle} x_j(t)}{\sum_{k=1}^{n} e^{\beta\langle x_i(t),\, x_k(t)\rangle}}$$

Gradient Flow

$$\dot{x}_i(t) = \mathrm{Proj}_{T_{x_i(t)}S^{d-1}} \frac{\sum_{j=1}^{n} e^{\beta\langle x_i(t),\, x_j(t)\rangle} x_j(t)}{\sum_{k=1}^{n} e^{\beta\langle x_i(t),\, x_k(t)\rangle}}$$

→ it is a gradient flow of the repulsive interaction energy

$$F(\mu) = \frac{1}{2}\sum_{i,j} e^{-\frac{\beta}{2}\|x_i(t) - x_j(t)\|^2}$$

self-attention: semantic appeal

energy F: geometric repulsion

"apple" must be close to "fruit"

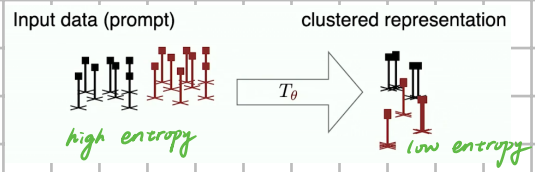
"apple" and "fruit" cannot be too close, otherwise the energy of the whole system will explore.

If the system is only affected by attraction, all tokens will gather to a point, causing all semantic information to overlap and the model to lose distinguishability.

Due to the presence of the repulsion energy F, transformers finally reach a state

of equilibrium (clustering) not a chaotic state, but rather a layout that maximizes geometric spacing on the sphere

What Happened To The Data



⇒ Map decreases entropy of input linearly

$\frac{d}{dt}\int \mu_t \log \mu_t = \Theta(t)$ → rate of linear change: representing a quantity that is linearly related to t

rate of change — free energy: representing the extent from chaos to order

Each layer of the transformer is transforming chaotic input data into an organized semantic structure at a linear, controllable pace

Schematic Diagram - from « Attention is all you need »