Vision Language Models

Generative AI and Foundation Models

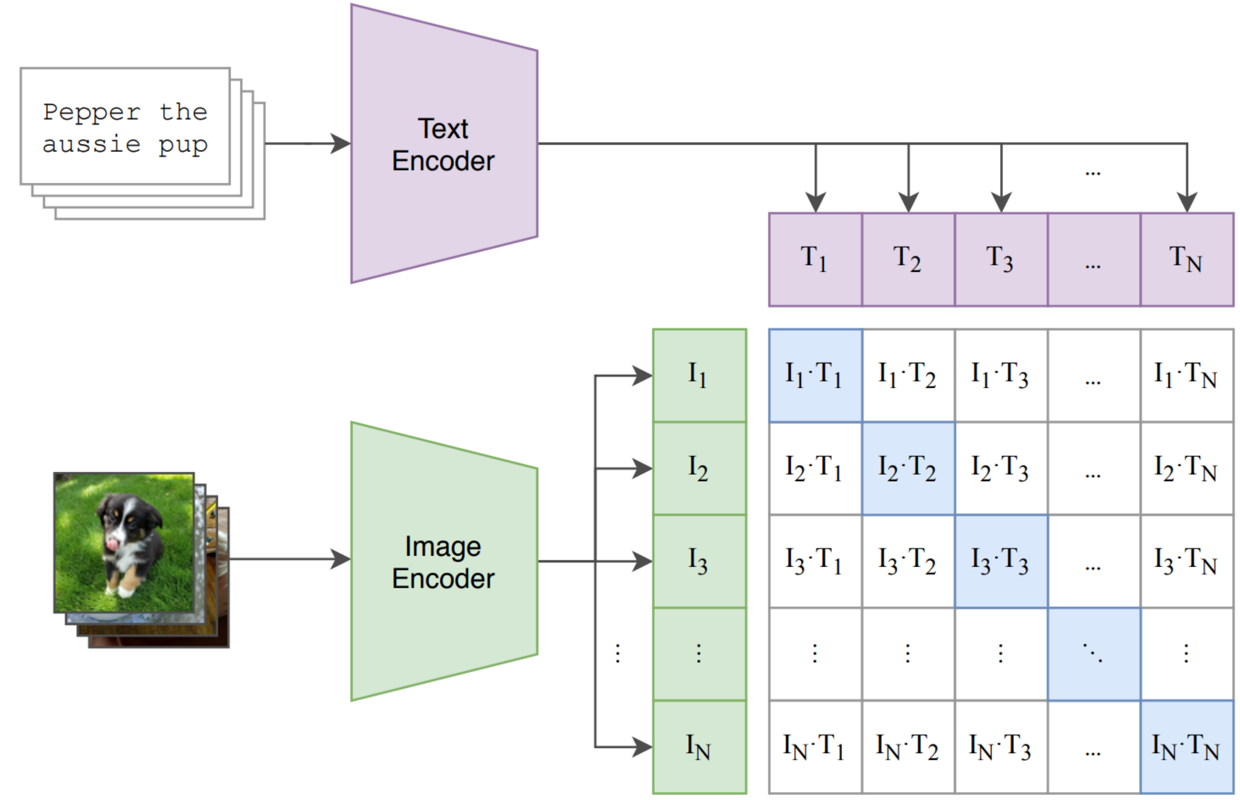
Spring 2024

Department of Mathematical Sciences

Ernest K. Ryu

Seoul National University

CLIP



Consider a dataset of image-caption pairs $\{(X_i, C_i)\}_{i=1}^{N}$.

Contrastive Language Image Pre-training (CLIP) find an image encoder $f_\theta : \mathcal{X} \to \mathbb{R}^d$ and text encoder $g_\phi : \mathcal{C} \to \mathbb{R}^d$ be the text encoder. Such that $f_\theta(X) \cdot g_\phi(C) > 0$ if X and C are related and $f_\theta(X) \cdot g_\phi(C) < 0$ or $f_\theta(X) \cdot g_\phi(C) \approx 0$ if X and C are not related.

A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, Learning transferable visual models from natural language supervision, *ICLR*, 2021.

InfoNCE loss

Let $\{(X_i, Y_i)\}_{i=1}^N$ be IID data pairs sampled from $p(\cdot,\cdot)$. We call

$$\mathcal{L}_{\mathrm{NCE}} = \frac{1}{N}\sum_{i=1}^{N} \log \frac{e^{f(X_i,Y_i)}}{\frac{1}{N}\sum_{j=1}^{N} e^{f(X_i,Y_j)}}$$

the InfoNCE (Noise Contrastive Estimation) loss.

Note that

$$\mathcal{L}_{\mathrm{NCE}} = \sum_{i=1}^{N} \log \frac{e^{f(X_i,Y_i)}}{\sum_{j=1}^{N} e^{f(X_i,Y_j)}}$$

is equivalent as a loss function as it differs only by a constant factor ($1/N$) and a constant term ($\log N$).

MI ≥ InfoNCE

Let $I(X;Y) = I(Y;X)$ denote the mutual information between X and Y.

Theorem. Let $\{(X_i, Y_i)\}_{i=1}^{N}$ be IID data pairs sampled from $p(\cdot,\cdot)$. Then, for any $f : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$, we have

$$I(X_1; Y_1) \geq \mathop{\mathbb{E}}_{\substack{(X_i,Y_i)\sim p \\ i=1,\ldots,N}} \left[\frac{1}{N} \sum_{i=1}^{N} \log \frac{e^{f(X_i,Y_i)}}{\frac{1}{N}\sum_{j=1}^{N} e^{f(X_i,Y_j)}} \right]$$

By symmetry, we also have

$$I(X_1; Y_1) \geq \mathop{\mathbb{E}}_{\substack{(X_i,Y_i)\sim p \\ i=1,\ldots,N}} \left[\frac{1}{N} \sum_{j=1}^{N} \log \frac{e^{f(X_j,Y_j)}}{\frac{1}{N}\sum_{i=1}^{N} e^{f(X_i,Y_j)}} \right]$$

(When $N < \infty$, the two InfoNCE losses are not exactly equal.)

B. Poole, S. Ozair, A. van den Oord, A. A. Alemi, and G. Tucker, On variational bounds of mutual information, *ICML*, 2019.

Proof. Let $p(x, y)$ be a joint probability density function on random variables X and Y. Let p_X and p_Y be the marginals for X and Y. Write $p(X|Y)$ for the conditional distribution of X conditioned on Y. Let $q(x|y)$ be any conditional distribution. Then,

$$
\begin{aligned}
I(X;Y) &= \mathop{\mathbb{E}}_{(X,Y)\sim p}\left[\log \frac{p(X,Y)}{p_X(X)p_Y(Y)}\right] \\
&= \mathop{\mathbb{E}}_{(X,Y)\sim p}\left[\log \frac{p(X\,|\,Y)}{p_X(X)}\right] \\
&= \mathop{\mathbb{E}}_{(X,Y)\sim p}\left[\log \frac{q(X\,|\,Y)}{p_X(X)}\right] + \mathop{\mathbb{E}}_{(X,Y)\sim p}\left[\log \frac{p(X\,|\,Y)}{q(X\,|\,Y)}\right] \\
&= \mathop{\mathbb{E}}_{(X,Y)\sim p}\left[\log \frac{q(X\,|\,Y)}{p_X(X)}\right] + \mathop{\mathbb{E}}_{Y\sim p_Y}\left[\mathop{\mathbb{E}}_{X\sim p(\cdot\,|\,Y)}\left[\log \frac{p(X\,|\,Y)}{q(X\,|\,Y)}\right] \,\middle|\, Y\right] \\
&= \mathop{\mathbb{E}}_{(X,Y)\sim p}\left[\log \frac{q(X\,|\,Y)}{p_X(X)}\right] + \mathop{\mathbb{E}}_{Y\sim p_Y}\left[D_{\mathrm{KL}}(p(\cdot\,|\,Y)\|q(\cdot\,|\,Y))\right] \\
&\geq \mathop{\mathbb{E}}_{(X,Y)\sim p}\left[\log \frac{q(X\,|\,Y)}{p_X(X)}\right]
\end{aligned}
$$

Now let $h(x, y)$ be an arbitrary function such that $Z(Y) = \mathbb{E}_{X \sim p_X}\left[e^{h(X,Y)}\right] < \infty$ for all Y. Let

$$q(x \mid y) = p_X(x) \frac{e^{h(x,y)}}{Z(y)}$$

and plug it into our bound to get

$$\begin{aligned}
I(X;Y) &\geq \mathbb{E}_{(X,Y)\sim p}[h(X,Y)] - \mathbb{E}_{(X,Y)\sim p}[\log Z(Y)] \\
&= \mathbb{E}_{(X,Y)\sim p}[h(X,Y)] - \mathbb{E}_{Y\sim p_Y}[\log Z(Y)] \\
&\overset{\text{(i)}}{\geq} \mathbb{E}_{(X,Y)\sim p}[h(X,Y)] - \log \mathbb{E}_{Y\sim p_Y}[Z(Y)] \\
&\overset{\text{(ii)}}{\geq} \mathbb{E}_{(X,Y)\sim p}[h(X,Y)] - \frac{1}{e} \mathbb{E}_{Y\sim p_Y}[Z(Y)] \\
&= \mathbb{E}_{(X,Y)\sim p}[h(X,Y)] - \frac{1}{e} \mathbb{E}_{\substack{X\sim p_X \\ Y \sim p_Y}}\left[e^{h(X,Y)}\right]
\end{aligned}$$

where (i) follows from Jensen's inequality and (ii) follows from the inequality $\log(x) \leq x/e$. Note that $X \sim p_X$ and $Y \sim p_Y$ means $(X, Y) \sim p_X(X)p_Y(Y)$, i.e., X and Y are sampled independently. This is different from sampling $(X, Y) \sim p$ (except in the special case of $p(x, y) = p_X(x)p_Y(x)$).

So far, we have not made any assumptions on the dimensions of X and Y. Let $X_1 \in \mathcal{X}$ and $Y = (Y_1, \dots, Y_N) \in \mathcal{Y}^N$. Let

$$p(X_1, Y) = p(X_1, Y_1) \prod_{i=2}^{N} p_Y(Y_i),$$

i.e., sample a dependent pair $(X_1, Y_1) \sim p$ and otherwise sample $Y_2, \dots, Y_N$ independently. Then,

$$I(X_1; Y_1) = I(X_1; Y) = I(X_1; Y_1, Y_2, \dots, Y_N)$$

since (X_1, Y_1) and $(Y_2, \dots, Y_N)$ are independent. (Follows from the chain rule of mutual information.)

Using the previous bound, we have

$$I(X_1;Y) \geq \mathop{\mathbb{E}}_{\substack{(X_1,Y_1)\sim p \\ Y_i\sim p_Y,\ i=2,\dots,N}} [h(X_1,Y)] - \frac{1}{e} \mathop{\mathbb{E}}_{\substack{X_1\sim p_X \\ Y_i\sim p_Y,\ i=1,\dots,N}} \left[e^{h(X_1,Y)}\right]$$

If we set

$$h(X_1,Y) = 1 + \log \frac{e^{f(X_1,Y_1)}}{\frac{1}{N}\sum_{j=1}^N e^{f(X_1,Y_j)}}$$

then we have

$$\begin{aligned} I(X_1;Y_1) &= I(X_1;Y) \\ &\geq 1 + \mathop{\mathbb{E}}_{\substack{(X_1,Y_1)\sim p \\ Y_i\sim p_Y,\ i=2,\dots,N}} \left[\log \frac{e^{f(X_1,Y_1)}}{\frac{1}{N}\sum_{j=1}^N e^{f(X_1,Y_j)}}\right] - \mathop{\mathbb{E}}_{\substack{X_1\sim p_X \\ Y_i\sim p_Y,\ i=1,\dots,N}} \left[\frac{e^{f(X_1,Y_1)}}{\frac{1}{N}\sum_{j=1}^N e^{f(X_1,Y_j)}}\right] \\ &= 1 + \mathop{\mathbb{E}}_{\substack{(X_1,Y_1)\sim p \\ Y_i\sim p_Y,\ i=2,\dots,N}} \left[\log \frac{e^{f(X_1,Y_1)}}{\frac{1}{N}\sum_{j=1}^N e^{f(X_1,Y_j)}}\right] - \mathop{\mathbb{E}}_{\substack{X_1\sim p_X \\ Y_i\sim p_Y,\ i=1,\dots,N}} \left[\frac{\frac{1}{N}\sum_{j=1}^N e^{f(X_1,Y_j)}}{\frac{1}{N}\sum_{j=1}^N e^{f(X_1,Y_j)}}\right] \\ &= \mathop{\mathbb{E}}_{\substack{(X_1,Y_1)\sim p \\ Y_i\sim p_Y,\ i=2,\dots,N}} \left[\log \frac{e^{f(X_1,Y_1)}}{\frac{1}{N}\sum_{j=1}^N e^{f(X_1,Y_j)}}\right] \\ &= \mathop{\mathbb{E}}_{(X_i,Y_i)\sim p,\, i=1,\dots,N} \left[\frac{1}{N}\sum_{i=1}^N \log \frac{e^{f(X_i,Y_i)}}{\frac{1}{N}\sum_{j=1}^N e^{f(X_i,Y_j)}}\right] \end{aligned}$$

■

MI = InfoNCE at optimum as $N \to \infty$

Theorem. Let $\{(X_i, Y_i)\}_{i=1}^{N}$ be IID data pairs sampled from $p(\cdot,\cdot)$. Let

$$f_\star(x, y) = \log \frac{p(x, y)}{p_X(x) p_Y(y)} + \text{constant}$$

Then, $\mathcal{L}_{\text{NCE}} \to I(X_1; Y_1)$ as $N \to \infty$.

(The $f_\star$ is not the optimum/maximizer for finite sample (batch) size N, but it is optimal in the limit as $N \to \infty$ since it attains the MI upper bound.)

Proof. Recall $$\mathcal{L}_{\text{NCE}} = \frac{1}{N}\sum_{i=1}^{N} \log \frac{e^{f_\star(X_i,Y_i)}}{\frac{1}{N}\sum_{j=1}^{N} e^{f_\star(X_i,Y_j)}}, \qquad f_\star(X,Y) = \log \frac{p(X,Y)}{p_X(X)p_Y(Y)} + \text{constant}.$$

First consider the denominator:

$$\begin{aligned}
\frac{1}{N}\sum_{j=1}^{N} e^{f_\star(X_i,Y_j)} &= e^{\text{constant}} \frac{1}{N}\sum_{j=1}^{N} \frac{p(X_i,Y_j)}{p_X(X_i)p_Y(Y_j)} \\
&= e^{\text{constant}} \frac{1}{N} \frac{p(X_i,Y_i)}{p_X(X_i)p_Y(Y_i)} + e^{\text{constant}} \frac{1}{N} \sum_{\substack{j=1 \\ j\neq i}}^{N} \frac{p(X_i,Y_j)}{p_X(X_i)p_Y(Y_j)} \\
&= \mathcal{O}(1/N) + e^{\text{constant}} \frac{N-1}{N} \frac{1}{N-1} \sum_{\substack{j=1 \\ j\neq i}}^{N} \frac{p(X_i,Y_j)}{p_X(X_i)p_Y(Y_j)} \\
&\to e^{\text{constant}} \mathop{\mathbb{E}}_{\substack{X\sim p_X \\ Y\sim p_Y}} \left[\frac{p(X,Y)}{p_X(X)p_Y(Y)} \right] \\
&= e^{\text{constant}} \int_{\mathcal{X}} \int_{\mathcal{Y}} \frac{p(x,y)}{p_X(x)p_Y(y)} p_X(x)p_Y(y)\, dxdy \\
&= e^{\text{constant}} \int_{\mathcal{X}} \int_{\mathcal{Y}} p(x,y)\, dxdy \\
&= e^{\text{constant}}
\end{aligned}$$

Indeed, with $(X_i, Y_i) \sim p$ for $i = 1, \ldots, N$,

$$\begin{aligned}
\mathcal{L}_{\mathrm{NCE}} &= \frac{1}{N} \sum_{i=1}^{N} \log \frac{e^{f_\star(X_i, Y_i)}}{\frac{1}{N} \sum_{j=1}^{N} e^{f_\star(X_i, Y_j)}} \\
&\approx \frac{1}{N} \sum_{i=1}^{N} \log \frac{p(X_i, Y_i)}{p_X(X_i) p_Y(Y_i)} \\
&\approx \mathop{\mathbb{E}}_{(X_1, Y_1) \sim p} \left[\log \frac{p(X_1, Y_1)}{p_X(X_1) p_Y(Y_1)} \right] \\
&= I(X_1; Y_1)
\end{aligned}$$

■

InfoNCE loss and CE loss

Consider the InfoNCE loss

$$\mathcal{L}_{\mathrm{NCE}} = \sum_{i=1}^{N} \underbrace{\log \frac{e^{f(X_i, Y_i)}}{\sum_{j=1}^{N} e^{f(X_i, Y_j)}}}_{\stackrel{\mathrm{def}}{=} \ell_{\mathrm{NCE}}(Y_:, X_i)}$$

Each term $\ell_{\mathrm{NCE}}(Y_:, X_i)$ can be viewed as the cross entropy loss applied to classifying X_i into N classes with ground truth label/class i with prediction probabilities

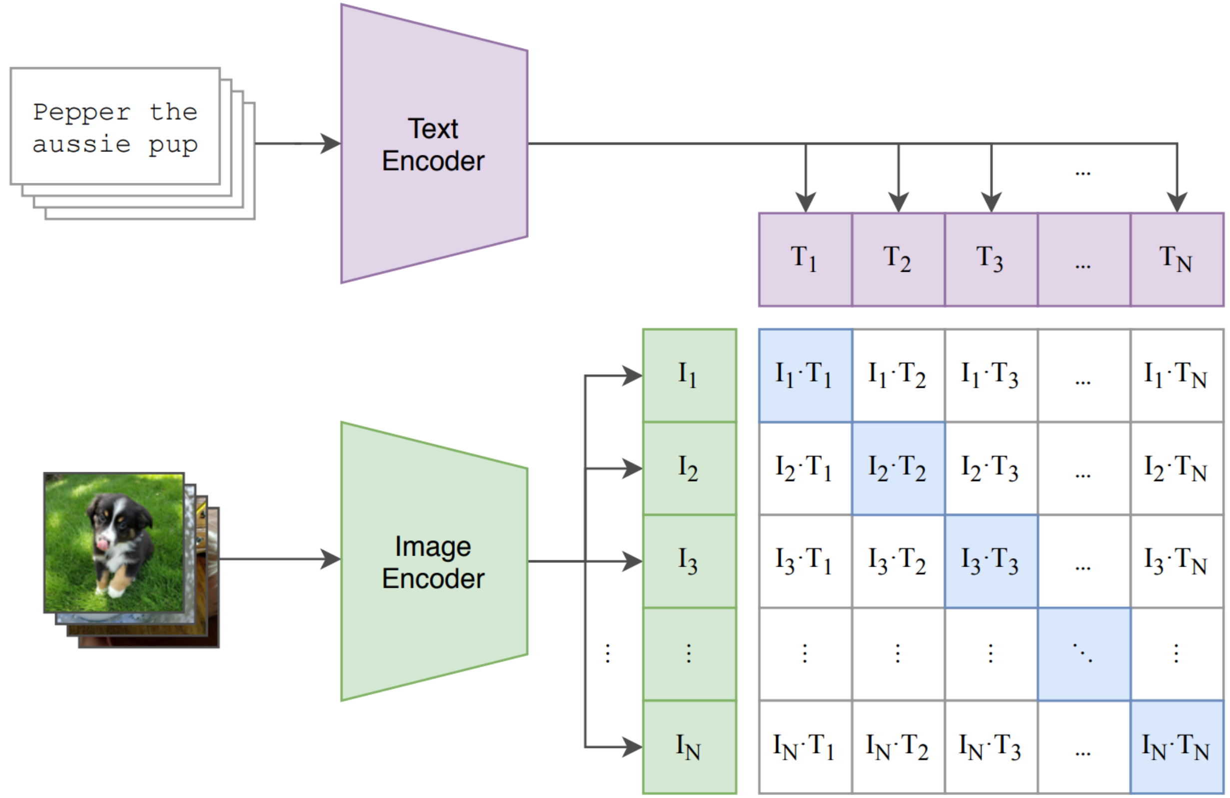
$$\mathbb{P}(\text{class of } X_i = j) \propto \exp\left(f\left(X_i, Y_j\right)\right)$$

To put it differently, $F(\cdot; Y) = \left(f(\cdot; Y_1), f(\cdot; Y_2), \ldots, f(\cdot; Y_N)\right)$ is the pre-softmax neural network for classifying an input x. Remember,

$$\ell^{\mathrm{CE}}(F(x), i) = -\log\left(\frac{\exp\left(F_i(x)\right)}{\sum_{j=1}^{k} \exp\left(F_j(x)\right)}\right)$$

CLIP

Consider a dataset of image-caption pairs $\{(X_i, C_i)\}_{i=1}^{N}$. Let $f_\theta : \mathcal{X} \to \mathbb{R}^d$ be the image encoder and $g_\phi : \mathcal{C} \to \mathbb{R}^d$ be the text encoder.



Contrastive Language Image Pre-training (CLIP) maximizes

$$\mathcal{L}_{\text{NCE}}(\theta,\phi) = \frac{1}{N}\sum_{i=1}^{N}\log\frac{\exp(f_\theta(X_i)\cdot g_\phi(C_i)/\tau)}{\frac{1}{N}\sum_{j=1}^{N}\exp(f_\theta(X_i)\cdot g_\phi(C_j)/\tau)} + \frac{1}{N}\sum_{i=1}^{N}\log\frac{\exp(f_\theta(X_i)\cdot g_\phi(C_i)/\tau)}{\frac{1}{N}\sum_{j=1}^{N}\exp(f_\theta(X_j)\cdot g_\phi(C_i)/\tau)}$$

$$\cong \sum_{i=1}^{N}\log\frac{\exp(f_\theta(X_i)\cdot g_\phi(C_i)/\tau)}{\sum_{j=1}^{N}\exp(f_\theta(X_i)\cdot g_\phi(C_j)/\tau)} + \sum_{i=1}^{N}\log\frac{\exp(f_\theta(X_i)\cdot g_\phi(C_i)/\tau)}{\sum_{j=1}^{N}\exp(f_\theta(X_j)\cdot g_\phi(C_i)/\tau)}$$

A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, Learning transferable visual models from natural language supervision, *ICLR*, 2021.

CLIP approximates MI

Roughly, CLIP trains embeddings in $\mathbb{R}^d$ such that $f_\theta(X) \cdot g_\phi(C)$ is large if X and C are related (C describes the contents of image X) and small if X and C are not related.

By the data processing inequality

$$I(X;C) \geq I(f_\theta(X);C) \geq I(f_\theta(X);g_\phi(C))$$

By our previous analysis, we have

$$I(X;C) \geq I(f_\theta(X);g_\phi(C)) \geq \frac{1}{2}\mathbb{E}[\mathcal{L}_{\mathrm{NCE}}]$$

By our previous analysis the bound is attained ($I(X;C) = (1/2)\mathcal{L}_{\mathrm{NCE}}$) if $N \to \infty$ and

$$f_{\theta^\star}(X) \cdot g_{\phi^\star}(C) + \text{constant} = \tau \log \frac{p(X,C)}{p(X)p(C)} = \tau \log p(C \mid X) - \tau \log p(C)$$

Are joint embeddings universal?

Is the approximation

$$f_{\theta^\star}(X) \cdot g_{\phi^\star}(C) + \text{constant} \approx \tau \log p(C \mid X) - \tau \log p(C)$$

possible? The RHS is, in general, a very complicated function jointly depending on X and C while the inner product structure of LHS feels like a separable-ish structure.

To rephrase the question, given that f_θ and g_ϕ are, in some sense, universal approximators, is

$$f_\theta(X) \cdot g_\phi(C) = \sum_{k=1}^{d} (f_\theta(X))_k (g_\phi(C))_k$$

a universal approximator of any function $h(X, C)$? The answer is yes, if d is large.

Universality of joint embeddings I

Let $\mathcal{X}$ and $\mathcal{Y}$ be locally compact Hausdorff (LCH) spaces. LCH spaces include the space of images, usually represented as $\mathbb{R}^n$, and the space of sentences, discrete spaces usually represented as $\mathcal{V}^*$.

Let $\mathcal{F} \subset \mathcal{C}(\mathcal{X}; \mathbb{R})$ and $\mathcal{G} \subset \mathcal{C}(\mathcal{Y}; \mathbb{R})$ be dense sub-vector spaces in the topology of uniform convergence on compacta. Then the Stone–Weierstrass theorem tells us that

$$\left\{ \sum_{k=1}^{d} f_k(x) g_k(y) \,\middle|\, f_1, \ldots, f_k \in \mathcal{F},\, g_1, \ldots, g_k \in \mathcal{G},\, d \in \mathbb{N} \right\} \subset \mathcal{C}(\mathcal{X} \times \mathcal{Y}; \mathbb{R})$$

which forms an algebra, is dense in the topology of uniform convergence on compacta. In other words, if we have a joint embedding $f_\theta : \mathcal{X} \to \mathbb{R}^d$ and $g_\phi : \mathcal{Y} \to \mathbb{R}^d$, then $h_{\theta,\phi}(x, y) = f_\theta(x) \cdot g_\phi(y)$ is a universal approximator if $d \to \infty$ and $f_\theta(x)$ and $g_\phi(y)$ has depth ≥ 2 and width$\to \infty$.

Universality of joint embeddings II

Assume $L^2(\mathcal{X};\mathbb{R})$ and $L^2(\mathcal{Y};\mathbb{R})$ be separable Hilbert spaces. (Essentially all Hilbert spaces arising in "real life" are separable.)

Then, $L^2(\mathcal{X};\mathbb{R}) \otimes L^2(\mathcal{Y};\mathbb{R}) = L^2(\mathcal{X}\times\mathcal{Y};\mathbb{R})$, i.e.,

$$\left\{ \sum_{k=1}^{d} f_k(x) g_k(y) \,\middle|\, f_1,\ldots,f_k \in L^2(\mathcal{X};\mathbb{R}),\ g_1,\ldots,g_k \in L^2(\mathcal{Y};\mathbb{R}),\ d\in\mathbb{N} \right\} \subset L^2(\mathcal{X}\times\mathcal{Y};\mathbb{R})$$

is dense. In other words, $h_{\theta,\phi}(x,y) = f_\theta(x)\cdot g_\phi(y)$ is a universal approximator if $d\to\infty$ and $f_\theta(x)$ and $g_\phi(y)$ has depth ≥ 2 and width$\to\infty$.