LIMITATIONS FOR LEARNING FROM POINT CLOUDS

Anonymous authors
Paper under double-blind review

ABSTRACT

In this paper we prove new universal approximation theorems for deep learning on point clouds that do not assume fixed cardinality. We do this by first generalizing the classical universal approximation theorem to general compact Hausdorff spaces and then applying this to the permutation-invariant architectures presented in *PointNet* (Qi et al) and *Deep Sets* (Zaheer et al). Moreover, though both architectures operate on the same domain, we show that the constant functions are the only functions they can mutually uniformly approximate. In particular, DeepSets architectures cannot uniformly approximate the diameter function but can uniformly approximate the center of mass function but it is the other way around for PointNet.

1 INTRODUCTION

Recently, architectures proposed in *PointNet* (Qi et al., 2017) and *Deep Sets* (Zaheer et al., 2017) have allowed for the direct processing of point clouds within a deep learning framework. These methods produce outputs that are permutation-invariant with respect to the member points and work for point clouds of arbitrarily large cardinality. Zaheer et al. (2017) also presents a permutation-equivariant architecture which we do not discuss here.

Each of of these works provide their own universal approximation theorem (UAT) to support the empirical success of their architectures. However, both results assume the cardinality of the point cloud is fixed to some size n. In this work we refine these results, remove the cardinality limitation and arrive at two main results which can be summarized roughly as follows (assuming unrestricted finite cardinality for the input point clouds):

1) PointNet (DeepSets) architectures can uniformly approximate real-valued functions that are uniformly continuous with respect to the Hausdorff (Wasserstein) metric and nothing else.
2) Only the constant functions can be uniformly approximated by both architectures. In particular, PointNet architectures can uniformly approximate the diameter function but DeepSets architectures cannot. Conversely, DeepSets architectures can uniformly approximate the center of mass function but PointNet architectures cannot.

To do this we extend the many universal approximation results for feed-forward networks (Cybenko, 1989; Hornik et al., 1989; Leshno et al., 1993; Stinchcombe, 1999) to the abstract setting of general compact Hausdorff spaces. We then find appropriate compact metric spaces over which PointNet and DeepSets architectures can be easily analyzed and then finally we observe the resulting consequences in the original setting of interest, i.e. point clouds.

2 PRELIMINARIES

2.1 POINTNET AND DEEPSETS ARCHITECTURES

In practice, the implementations of the architectures presented in *PointNet* and *Deep Sets* can involve many additional tricks, but the essential ideas are quite simple. We do however make a small modification to the *Deep Sets* model. For $A \subseteq \mathbb{R}^n$ of cardinality $|A| < \infty$, we have Qi et al. (2017)

and Zaheer et al. (2017) suggesting scalar-output neural networks of the form

$$F_{PN}(A) = \rho\left(\max_{\boldsymbol{a}\in A} \boldsymbol{\varphi}(\boldsymbol{a})\right), \quad \text{and} \quad F_{DS}(A) = \rho\left(\boldsymbol{b} + \frac{1}{|A|}\sum_{\boldsymbol{a}\in A} \boldsymbol{\varphi}(\boldsymbol{a})\right),$$

respectively. Here $\boldsymbol{\varphi} : \mathbb{R}^n \to \mathbb{R}^m$ creates features for each point in A, then a symmetric operation is applied, and then $\rho : \mathbb{R}^m \to \mathbb{R}$ combines these features into a scalar output (here $\max$ is the component-wise maximum). In practice, we need both ρ and $\boldsymbol{\varphi}$ to be neural networks. Note that because we use a symmetric operation before ρ, the output will not depend on the ordering of points in the point cloud, and because the max and sum operations scale to arbitrary finite cardinalities the size of the point cloud is not an issue. The original model in *Deep Sets* did not have a bias term $\boldsymbol{b}$ and used a sum instead of the averaging we use here. This change will help us later in our theoretical analysis.

It will help to introduce some simplifying notation. Let $\mathcal{F}(\Omega)$ denote the set of all nonempty finite subsets of a set Ω (i.e. point clouds in Ω) and let $\mathcal{F}^k(\Omega)$ be the set of k-point subsets. Now consider $\Omega \subseteq \mathbb{R}^N$ and define $\max_f, \mathrm{ave}_{f,b} : \mathcal{F}(\Omega) \to \mathbb{R}$ which are given by $\max_f(A) = \max_{\boldsymbol{a}\in A} f(\boldsymbol{a})$ and $\mathrm{ave}_{f,b}(A) = b + \frac{1}{|A|}\sum_{\boldsymbol{a}\in A} f(\boldsymbol{a})$ respectively. We make sense of this in the natural way if we use vector-valued $\boldsymbol{f}$ and $\boldsymbol{b}$ by operating component-wise. We call these operations max neurons and biased-averaging neurons respectively.

Once again letting ρ and $\boldsymbol{\varphi}$ be neural networks, $F_{PN} = \rho \circ \max_{\boldsymbol{\varphi}}$ and $F_{DS} = \rho \circ \mathrm{ave}_{\boldsymbol{\varphi},\boldsymbol{b}}$ will be the general form of what we call the PointNet and DeepSets architectures (resp.) in this paper.

Some natural questions are 1) is there a topology for $\mathcal{F}(\Omega)$ that makes these architectures continuous, 2) how expressive are these approaches, and 3) how deep is deep enough for function approximation?

2.2 Function Spaces and Uniform Approximation

From now on, we only consider $\mathbb{R}$-valued functions unless otherwise stated. Let $B(A)$ be the set of bounded functions on a set A, let $C(X)$ and $C_b(X)$ be the set of continuous and bounded continuous functions on a topological space X (respectively), and let $U(M)$ and $U_b(M)$ be the uniformly continuous and bounded uniformly continuous functions on a metric space (M, d) (respectively). We equip all of these with the uniform norm i.e. $\|f\|_A = \sup_{a\in A} |f(a)|$. This makes them all normed spaces, with $B(A)$, $C_b(X)$ and $U_b(M)$ additionally being Banach spaces. Moreover, if X is compact and (M, d) has compact metric completion, then $C(X) = C_b(X)$ and $U(M) = U_b(M)$ and hence are also Banach spaces. For background see Rudin (2006).

If given an injective map $i : A \to X$, then we say that $\varphi : A \to \mathbb{R}$ (uniquely) continuously extends to X if there is a (unique) $\widetilde{\varphi} \in C(X)$ such that $\widetilde{\varphi} \circ i = \varphi$. We say a family of funcntions $\mathcal{N}$ on A (uniquely) continuously extends to X if every $\varphi \in \mathcal{N}$ (uniquely) continuously extends to X.

We will make use of the following lemma which is proved in the appendix.

Lemma 2.1. *Let $\mathcal{N} \subseteq B(D)$ where D is a dense subset of a compact metric space (X, d). Suppose $\mathcal{N}$ has a continuous extension to X denoted by $\mathcal{N}' \subseteq C(X)$ which is dense. Then the uniform closure of $\mathcal{N}$ in $B(D)$ is $r(C(X)) = U(D)$ where $r : C(X) \to C_b(D)$ is the domain restriction map.*

Letting $D = \mathcal{F}(\Omega)$, this lemma suggest the following plan of attack: find a compact metric space (X, d) in which we can realize $\mathcal{F}(\Omega)$ as a dense subset and hope that our class of neural networks $\mathcal{N}$ continuously extends to a dense subset of $C(X)$. If we can do that, then we know the uniform closure of our class of neural networks are precisely the uniformly continuous functions on $\mathcal{F}(\Omega)$ with respect to the metric inherited from X. This motivates the next subsection.

2.3 Metrics on the Space of Point Clouds

From now on we will assume (Ω, d) is a compact metric space and when $\Omega \subseteq \mathbb{R}^n$ it will be compact and equipped with the Euclidean metric. Let $\mathcal{K}(\Omega)$ denote the set of all compact subsets of Ω and $\mathcal{P}(\Omega)$ denote the set of all Borel probability measures on Ω. The Hausdorff metric d_H (Munkres, 2000) is a natural metric for $\mathcal{K}(\Omega)$ and 1-Wasserstein metric d_W (Villani, 2009) (also called the Earth-mover distance) is a natural metric for $\mathcal{P}(\Omega)$. With these metrics, $\mathcal{K}(\Omega)$ and $\mathcal{P}(\Omega)$ become

compact metric spaces of their own. From now on we will assume these two spaces are always equipped with the aforementioned metrics.

We also briefly mention $M(\Omega)$ the Banach space of finite signed regular Borel measures on Ω. By the Riesz-Markov theorem it is the topological dual space of $C(\Omega)$. Of interest to us is that $\mathcal{P}(\Omega) \subseteq M(\Omega)$ and that the weak-* topology on $\mathcal{P}(\Omega)$ coincides with the topology induced by d_W. This means that $d_W(\mu_n, \mu) \to 0$ iff $\int f \, \mathrm{d}\mu_n \to \int f \, \mathrm{d}\mu$ for all $f \in C(\Omega)$.

Next, note that $\mathcal{F}(\Omega) \subseteq \mathcal{K}(\Omega)$ and let $i_{\mathcal{K}}$ denote the natural inclusion map. We can also define an injective map $i_{\mathcal{P}} : \mathcal{F}(\Omega) \to \mathcal{P}(\Omega)$ by mapping $A \in \mathcal{F}(\Omega)$ to its associated empirical measure $i_{\mathcal{P}}(A) = \mu_A = \frac{1}{|A|} \sum_{a \in A} \delta_a \in \mathcal{P}(\Omega)$ where δ_a is the Dirac delta measure supported at a. The injective maps $i_{\mathcal{K}}$ and $i_{\mathcal{P}}$ allow us to induce the d_H and d_W metrics on $\mathcal{F}(\Omega)$. We will denote the metrized versions by $\mathcal{F}_H(\Omega)$ and $\mathcal{F}_W(\Omega)$ respectively and use the same convention for $\mathcal{F}_H^k(\Omega)$ and $\mathcal{F}_W^k(\Omega)$. Another important fact to know is that $i_{\mathcal{K}}$ and $i_{\mathcal{P}}$ embed $\mathcal{F}(\Omega)$ as dense subset of $\mathcal{K}(\Omega)$ and $\mathcal{P}(\Omega)$. The former follows from compactness of the members of $\mathcal{K}(\Omega)$ and to see why the latter is true see Fournier & Guillin (2015); Villani (2009)

For $f \in C(\Omega)$ and $b \in \mathbb{R}$, define $\mathrm{Max}_f : \mathcal{K}(\Omega) \to \mathbb{R}$ and $\mathrm{Ave}_{f,b} : \mathcal{P}(\Omega) \to \mathbb{R}$ as the functions $\mathrm{Max}_f(K) = \max_{x \in K} f(x)$ and $\mathrm{Ave}_{f,b}(\mu) = b + \int_\Omega f \, \mathrm{d}\mu$.

Lemma 2.2. *Let (Ω, d) be compact, $f \in C(\Omega)$, and $b \in \mathbb{R}$. Then $\mathrm{Max}_f \in C(\mathcal{K}(\Omega))$ and $\mathrm{Ave}_{f,b} \in C(\mathcal{P}(\Omega))$ and $\mathrm{Max}_f \circ i_{\mathcal{K}} = \max_f$ and $\mathrm{Ave}_{f,b} \circ i_{\mathcal{P}} = \mathrm{ave}_{f,b}$.*

This lemma (proved in the appendix) tells us that the max neurons and biased-averaging neurons continuously extend to $\mathcal{K}(\Omega)$ and $\mathcal{P}(\Omega)$ and hence so do PointNet and DeepSets architectures (since we merely compose with the continuous ρ after). Thus, we wil be able analyze such architectures as continuous functions on compact metric spaces, which is mathematically a much nicer problem than studying them as set-theoretic functions on an un-metrized $\mathcal{F}(\Omega)$.

2.4 Generalized Neural Network Notation

For $\mathcal{A}$ a collection of functions from X to Y and $\mathcal{B}$ a collection of functions from Y to Z, we denote the set of all compositions by $\mathcal{B} \circ \mathcal{A} = \{f \circ g \mid f \in \mathcal{B}, g \in \mathcal{A}\}$. In the case of a single function $\sigma : Y \to Z$ we let $\sigma \circ \mathcal{A} = \{\sigma \circ f \mid f \in \mathcal{A}\}$ and similarly for right-composition.

Next, let Aff denote the set of all affine functionals on $\mathbb{R}^N$, i.e. any function of the form $f(\boldsymbol{x}) = \boldsymbol{w} \cdot \boldsymbol{x} + b$. Let $\mathcal{N}^\sigma := \mathrm{span}\{\sigma \circ \mathrm{Aff}\}$ denote the set of single hidden-layer neural networks with linear output-layer and activation function σ, and then denote an H-layered network by $\mathcal{N}^{\boldsymbol{\sigma}}$ where $\boldsymbol{\sigma} = (\sigma_1, \ldots, \sigma_H)$ is a list of H-many activation functions where $\mathcal{N}^{(\boldsymbol{\sigma}, \tau)} := \mathcal{N}^{\boldsymbol{\sigma}, \tau} := \mathrm{span}(\tau \circ \mathcal{N}^{\boldsymbol{\sigma}})$.

Next we define various classes of PointNet networks whose weight functions are themselves neural networks. Let $\mathcal{N}_{PN}^\sigma := \mathcal{N}_{PN}^{\sigma;\varnothing} := \mathrm{span}\{\max_f \mid f \in \mathcal{N}^\sigma\}$ then define $\mathcal{N}_{PN}^{\sigma;\tau} := \mathrm{span}\{\tau \circ \mathcal{N}_{PN}^\sigma\}$. Like before, we can inductively define deeper networks, but we can also use deeper weight networks as well to create $\mathcal{N}_{PN}^{\boldsymbol{\sigma};\boldsymbol{\tau}}$ – thus we have two distinct notions of depth.

Next, we do the same for DeepSets. Let $\mathcal{N}_{DS}^\sigma := \mathcal{N}_{DS}^{\sigma;\varnothing} := \mathrm{span}\{\mathrm{ave}_{f,b} \mid f \in \mathcal{N}^\sigma, b \in \mathbb{R}\}$. Note that $\mathrm{ave}_{f,b} + \mathrm{ave}_{g,c} = \mathrm{ave}_{f+g,b+c}$ and $\alpha \, \mathrm{ave}_{f,b} = \mathrm{ave}_{\alpha f, \alpha b}$. Thus since $\mathcal{N}^\sigma$ is a linear space, taking the span has no effect and $\mathcal{N}_{DS}^\sigma = \{\mathrm{ave}_{f,b} \mid f \in \mathcal{N}^\sigma, b \in \mathbb{R}\}$. Going one layer deeper yields $\mathcal{N}_{DS}^{\sigma;\tau} := \mathrm{span}\{\tau \circ \mathcal{N}_{DS}^\sigma\}$ which gets us new functions. Like with PointNet, we can inductively develop deeper families in two ways.

By Lemma 2.2 we can extend all the operations of our neural networks to $\mathcal{K}(\Omega)$ and $\mathcal{P}(\Omega)$ in a natural way. This let's us talk about about PointNet networks on $\mathcal{K}(\Omega)$ and DeepSets networks on $\mathcal{P}(\Omega)$ which we'll define analogously by replacing $\max_f$ with Max_f and $\mathrm{ave}_{f,b}$ with $\mathrm{Ave}_{f,b}$. Thus

$$\mathcal{M}_{PN}^\sigma = \mathrm{span}\{\mathrm{Max}_f \mid f \in \mathcal{N}^\sigma\}, \qquad \mathcal{M}_{PN}^{\sigma;\tau} = \mathrm{span}\{\tau \circ \mathcal{N}_{PN}^\sigma\},$$
$$\mathcal{M}_{DS}^\sigma = \mathrm{span}\{\mathrm{Ave}_{f,b} \mid f \in \mathcal{N}^\sigma, b \in \mathbb{R}\}, \qquad \mathcal{M}_{DS}^{\sigma;\tau} = \mathrm{span}\{\tau \circ \mathcal{M}_{DS}^\sigma\}.$$

As before, the linear structure of $\mathcal{N}^\sigma$ makes $\mathcal{M}_{DS}^\sigma = \{\mathrm{Ave}_{f,b} \mid f \in \mathcal{N}^\sigma\}$

3 MAIN RESULTS

3.1 TOPOLOGICAL UAT

Leshno et al. (1993) prove that $\mathcal{N}^\sigma$ with $\sigma \in C(\mathbb{R})$ has universal approximation property iff σ is not a polynomial. For this reason, we will say a $\sigma \in C(\mathbb{R})$ is 'universal' if it is non-polynomial and denote the set of all such such functions by $\mathfrak{U}(\mathbb{R})$. Using this theorem and Stone-Weierstrass we prove a UAT for certain kinds of two-hidden-layer 'neural networks' on an abstract compact Hausdorff space.

Recall that a family of functions S on Ω separates points if for any $x \neq y$ there is an $f \in S$ so that $f(x) \neq f(y)$.

Theorem 3.1 (Topological-UAT)**.** *Let X be a compact Hausdorff space and $\sigma \in \mathfrak{U}(\mathbb{R})$. If $S \subseteq C(X)$ separates points and contains a nonzero constant, then $\operatorname{span}(\sigma \circ \operatorname{span} S)$ is dense in $C(X)$. Additionally, if S also happens to be a linear subspace, then $\operatorname{span}(\sigma \circ S)$ is dense in $C(X)$.*

Proof. Let S and σ satisfy the above and let $V = \operatorname{span} S$. Let $\operatorname{Alg}(V)$ denote the algebra generated by V, i.e. all possible finite products, sums and scalar multiples of the elements of V. Then $\operatorname{Alg}(V)$ is unital subalgebra of $C(X)$ that seperates points. By the Stone-Weierstrass theorem $\operatorname{Alg}(V)$ is dense in $C(X)$. Now let $F \in C(X)$ and $\epsilon > 0$ be arbitrary. By density there is a $G \in \operatorname{Alg}(V)$ such that $|F(a) - G(a)| < \epsilon/2$ for all $a \in X$. Since $G \in \operatorname{Alg}(V)$ there is an N-variable polynomial p and $\boldsymbol{s} = (s_1, \ldots, s_N)$ where $s_i \in S$, so that $G = p \circ \boldsymbol{s}$. Since all $s_i \in C(X)$ and X is compact, the image $\boldsymbol{s}(X) \subseteq \mathbb{R}^N$ is compact. By the classical UAT (Leshno et al., 1993), there exists an $\eta \in \mathcal{N}^\sigma$ such that $|p(\boldsymbol{x}) - \eta(\boldsymbol{x})| < \epsilon/2$ for all $\boldsymbol{x} \in \mathbb{R}^N$. Thus,

$$|F(a) - (\eta \circ \boldsymbol{s})(a)| \leq |F(a) - p(\boldsymbol{s}(a))| + |p(\boldsymbol{s}(a)) - \eta(\boldsymbol{s}(a))| < \epsilon/2 + \epsilon/2 = \epsilon$$

for every $a \in X$. Finally note that $\eta(\boldsymbol{s}(a)) = \sum_{i=1}^m a_i \sigma(\boldsymbol{w}_i \cdot \boldsymbol{s}(a) + b_i)$ for some $a_i, b_i \in \mathbb{R}$ and $\boldsymbol{w}_i \in \mathbb{R}^N$. Since S constains a nonzero constant, $\operatorname{span} S$ contains every constant and so $\boldsymbol{w}_i \cdot \boldsymbol{s} + b_i \in \operatorname{span} S$. Thus $\eta \circ \boldsymbol{s} \in \operatorname{span}(\sigma \circ \operatorname{span} S)$ as desired.

Lastly, if S is also linear subspace, then $S = \operatorname{span} S$ and so $\operatorname{span}(\sigma \circ S)$ is dense in $C(X)$. □

3.2 POINT CLOUD UAT

We have met almost all the conditions required to use the topological-UAT on $\mathcal{K}(\Omega)$ and $\mathcal{P}(\Omega)$. We just need to show that Max_f and $\operatorname{Ave}_{f,b}$ yield nonzero constants and can separate points even when we limit ourselves to $f \in \mathcal{N}^\sigma$.

Lemma 3.2 (Separation Lemma)**.** *Let $\Omega \subseteq \mathbb{R}^N$ be compact and $\sigma \in \mathfrak{U}(\mathbb{R})$. Then $S_{PN} = \{\operatorname{Max}_f \mid f \in \mathcal{N}^\sigma\}$ and $S_{DS} = \{\operatorname{Ave}_{f,b} \mid f \in \mathcal{N}^\sigma, b \in \mathbb{R}\}$ separate points and contain constants.*

Proof. Let d denote the Euclidean distance. First note that the constant function $h = \sigma(c) \in \mathcal{N}^\sigma$ for some $c \in \mathbb{R}$. Since σ is not a polynomial, there is a choice of c for which $\sigma(c) \neq 0$. This means $\operatorname{Max}_h \in S_{PN}$ and $\operatorname{Ave}_{h,0} \in S_{DS}$ are both constant. Now we just need to show that S_{PN} and S_{DS} separate points.

(S_{PN} separates points): Let $A, B \in \mathcal{K}(\Omega)$ with $A \neq B$. Without loss of generality, $A \setminus B \neq \varnothing$ so choose $a \in A \setminus B$. Let $f(x) = \min\{1, d(x, B)/d(a, B)\}$ and note that $f(a) = 1$, $f(B) = \{0\}$ and $f(\Omega) = [0, 1]$. By the classical UAT (Leshno et al., 1993) $\mathcal{N}^\sigma$ is dense in $C(\Omega)$, so there is a $g \in \mathcal{N}^\sigma$ so that $|f(x) - g(x)| < 1/2$ for all $x \in \Omega$. Note $\operatorname{Max}_g \in S_{PN}$ and that $\operatorname{Max}_g(A) > 1/2$ and $\operatorname{Max}_g(B) < 1/2$. Since A and B were arbitrary, this shows S_{PN} separates point in $\mathcal{K}(X)$.

(S_{DS} separates points): Given $\mu_1, \mu_2 \in \mathcal{P}(\Omega)$ with $\mu_1 \neq \mu_2$, by the Hahn-Banach separation theorem there exists a weak-* continuous linear functional $L : M(\Omega) \to \mathbb{R}$ that separates them. Let $\delta = |L(\mu_1) - L(\mu_2)|$. The topological dual of $M(\Omega)$ with the weak-* topology is equivalent to $C(\Omega)$ and so there is an $f \in C(\Omega)$ so that $L(\eta) = \int f \, \mathrm{d}\eta$ for all $\eta \in M(\Omega)$. Since $\mathcal{N}^\sigma$ is dense in $C(\Omega)$ there is a $g \in \mathcal{N}^\sigma$ so the that $|f(x) - g(x)| < \delta/2$ for all $x \in \Omega$. Define $J(\eta) = \int g \, \mathrm{d}\eta$. Then for all $\eta \in \mathcal{P}(\Omega)$ we have $|L(\eta) - J(\eta)| \leq \int |f - g| \, \mathrm{d}\eta < \frac{\delta}{2} \int \mathrm{d}\eta = \delta/2$. Applying the triangle inequality we obtain,

$$\delta = |L(\mu_1) - L(\mu_2)| \leq \underbrace{|L(\mu_1) - J(\mu_1)|}_{<\delta/2} + |J(\mu_1) - J(\mu_2)| + \underbrace{|J(\mu_2) - L(\mu_2)|}_{<\delta/2}$$

Thus $0 < |J(\mu_1) - J(\mu_2)|$ and so $J = \mathrm{Ave}_{g,0} \in S_{DS}$ separates μ_1 and μ_2. Since μ_1 and μ_2 were arbitrary, it follows that S_{DS} seperates points in $\mathcal{P}(\Omega)$. □

The following theorems show that one hidden layer in the weight networks and one hidden layer of the of the other kind suffice to prove the universal approximation theorems for PointNet and DeepSets.

Theorem 3.3. *Let $\Omega \subseteq \mathbb{R}^N$ be compact and $\sigma, \tau \in \mathfrak{U}(\mathbb{R})$. Then $\mathcal{M}_{PN}^{\sigma;\tau}$ and $\mathcal{M}_{DS}^{\sigma;\tau}$ are dense in $C(\mathcal{A})$ and $C(\mathcal{B})$ respectively, where $\mathcal{A} \subseteq \mathcal{K}(\Omega)$ and $\mathcal{B} \subseteq \mathcal{P}(\Omega)$ are closed subsets.*

Proof. Recall $\mathcal{M}_{PN}^{\sigma;\tau} = \operatorname{span}\{\tau \circ \operatorname{span} S_{PN}\}$ and $\mathcal{M}_{DS}^{\sigma;\tau} = \operatorname{span}\{\tau \circ S_{DS}\}$. Since $\mathcal{K}(\Omega)$ and $\mathcal{P}(\Omega)$ are compact metric spaces, $\mathcal{A}$ and $\mathcal{B}$ are compact Hausdorff. By Lemma 3.2 we know S_{PN} and S_{DS} separate points and contain nonzero constants and so the topological-UAT (Theorem 3.1) yields the desired result. □

Theorem 3.4 (Point-Cloud-UAT)**.** *Let $\Omega \subseteq \mathbb{R}^N$ be compact. If $\sigma, \tau \in \mathfrak{U}(\mathbb{R})$, then the uniform closure of $\mathcal{N}_{PN}^{\sigma;\tau}$ and $\mathcal{N}_{DS}^{\sigma;\tau}$ within $B(\mathcal{F}(\Omega))$ is $U(\mathcal{F}_H(\Omega))$ and $U(\mathcal{F}_W(\Omega))$ respectively.*

Proof. $\mathcal{F}_H(\Omega)$ and $\mathcal{F}_W(\Omega)$ are isometrically isomorphic to $i_{\mathcal{K}}(\mathcal{F}(\Omega))$ and $i_{\mathcal{P}}(\mathcal{F}(\Omega))$ which are in turn dense in $(\mathcal{K}(\Omega), d_H)$ and $(\mathcal{P}(\Omega), d_W)$. By Lemma 2.2 we have that $\mathcal{N}_{PN}^{\sigma;\tau}$ and $\mathcal{N}_{DS}^{\sigma;\tau}$ continuously extend to $\mathcal{K}(\Omega)$ and $\mathcal{P}(\Omega)$ as $\mathcal{M}_{PN}^{\sigma;\tau}$ and $\mathcal{M}_{DS}^{\sigma;\tau}$. By Theorem 3.3 we know $\mathcal{M}_{PN}^{\sigma;\tau}$ and $\mathcal{M}_{DS}^{\sigma;\tau}$ are dense in $C(\mathcal{K}(\Omega))$ and $C(\mathcal{P}(\Omega))$. Finally, by Lemma 2.1 we have the desired result. □

It's worth noting that we could have used Stinchcombe's generalization of the UAT to the case of neural networks on compact subsets of locally convex spaces (Stinchcombe, 1999) to prove that $\mathcal{M}_{DS}^{\sigma;\tau}$ is dense in $C(\mathcal{P}(\Omega))$ but we chose the above route for consistency of technique and to be self-contained.

We now prove as a corollary a refinement of the universal approximation theorems of Qi et al. (2017) and Zaheer et al. (2017), both of which applied to the the case of k-point point clouds. In this version of the theorem we are able to restrict the depth of the neural network to just two hidden layers. The proof is essentially the same as Theorem 3.4.

Corollary 3.5. *Let $\Omega \subseteq \mathbb{R}^N$ be compact. If $\sigma, \tau \in \mathfrak{U}(\mathbb{R})$, then the uniform closure of $\mathcal{N}_{PN}^{\sigma;\tau}$ and $\mathcal{N}_{DS}^{\sigma;\tau}$ within $B(\mathcal{F}_k(\Omega))$ are $U(\mathcal{F}_k(\Omega)_H)$ and $U(\mathcal{F}_k(\Omega)_W)$ respectively.*

Proof. $\mathcal{F}_H^k(\Omega)$ and $\mathcal{F}_W^k(\Omega)$ are isometrically isomorphic to $i_{\mathcal{K}}(\mathcal{F}^k(\Omega))$ and $i_{\mathcal{P}}(\mathcal{F}^k(\Omega))$ which are in turn dense in their respective closures which we denote $\mathcal{G}_H(\Omega) \subseteq \mathcal{K}(\Omega)$ and $\mathcal{G}_W(\Omega) \subseteq \mathcal{P}(\Omega)$. Thus by Lemma 2.2 and Theorem 3.3 we have that $\mathcal{M}_{PN}^{\sigma;\tau}$ and $\mathcal{M}_{DS}^{\sigma;\tau}$ are dense in $C(\mathcal{G}_H(\Omega))$ and $C(\mathcal{G}_W(\Omega))$. Finally, by Lemma 2.1 we have the desired result. □

3.3 Limitations of PointNets and DeepSets

Note that unlike the classical universal approximation theorem we should not expect to be able uniformly approximate $C(\mathcal{F}_H(\Omega))$ or $C(\mathcal{F}_W(\Omega))$ since their elements might not even be bounded functions. For example, $\alpha_K(A) = d_H(A, K)^{-1}$ is unbounded but continuous on $\mathcal{F}_H(\Omega)$ whenever $K \in \mathcal{K}(\Omega)$ but $K \notin \mathcal{F}(\Omega)$. Subtler still, we do not even obtain all elements of $C_b(\mathcal{F}_H(\Omega))$ and $C_b(\mathcal{F}_W(\Omega))$. As an example of this, observe that $\beta_K = \sin \circ \alpha_K$ is bounded and is also continuous on $\mathcal{F}_H(\Omega)$ because α_K is.

We'll now compare the representation power of these two architectures. Let $\Omega \subseteq \mathbb{R}^N$ be compact. We define the point cloud diameter function $\mathrm{Diam} : \mathcal{F}(\Omega) \to \mathbb{R}$ and point cloud center-of-mass function $\mathrm{Center} : \mathcal{F}(\Omega) \to \mathbb{R}^N$ by $\mathrm{Diam}(A) = \max_{\boldsymbol{x},\boldsymbol{y} \in A} d(\boldsymbol{x}, \boldsymbol{y})$ and $\mathrm{Center}(A) = \frac{1}{|A|} \sum_{\boldsymbol{x} \in A} \boldsymbol{x}$. Then we can obtain the following result.

Theorem 3.6. *Let (Ω, d) be an infinite compact metric space with no isolated points. Then a function $f : \mathcal{F}(\Omega) \to \mathbb{R}$ is continuous with respect to both d_H and d_W iff it is constant. As a corollary,* Diam *is uniformly approximable by PointNet networks but not DeepSets networks and* Center *is uniformly approximable by DeepSets networks but not PointNet networks.*

Proof. Assume $f : \mathcal{F}(\Omega) \to \mathbb{R}$ is both d_H-continuous and d_W-continuous. Let $A \in \mathcal{F}(\Omega)$ and let $p \in A$. For each $n = 1, 2, \ldots$ choose $A'_n \in \mathcal{F}(\Omega)$ to be an n-point set contained within the $1/n$-ball around p. We can do this because Ω is infinite without isolated points. Now let $A_n = A'_n \cup (A \setminus \{p\})$.

Observe that $A_n \overset{d_H}{\to} A$ and $A_n \overset{d_W}{\to} \{p\}$. Thus,

$$f(A) = f\left(\lim_{n\to\infty}^{d_H} A_n\right) = \lim_{n\to\infty} f(A_n) = f\left(\lim_{n\to\infty}^{d_W} A_n\right) = f(\{p\})$$

Note that A was arbitrary so f must always give a set and any of its singleton subsets the same value. Now let $B, C \in \mathcal{F}(\Omega)$ such that $B \neq C$. Without loss of generality assume $q \in B \setminus C$ and $q \neq r \in C$. Thus by the above,

$$f(B) = f(\{q\}) = f(\{q, r\}) = f(\{r\}) = f(C)$$

thus f must be constant. Conversely, constant maps are always continuous.

Finally, $|\mathrm{Diam}(A) - \mathrm{Diam}(B)| \leq 2d_H(A, B)$ and hence is d_H-continuous on $\mathcal{K}(\Omega)$ and Center is d_W-continuous on $\mathcal{P}(\Omega)$ because $\mathrm{Ave}_{\pi_i,0}$ is d_W-continuous (here π_i is the projection onto the i-th component map). This means they are uniformly continuous on $\mathcal{F}_H(\Omega)$ and $\mathcal{F}_W(\Omega)$ respectively and so the result follows from the above and Theorem 3.4. □

4 Conclusions and Future Work

The failure of the perceptron model to learn the XOR function was a blow that motivated the search for new models. When classical feed-forward neural networks began to be successful, the spectre of limited representation power loomed over the field until the universal approximation theorem Cybenko (1989)Hornik et al. (1989)Leshno et al. (1993) resolved the question. In this work we resolved the same question for the case of two current deep learning models for point clouds, and laid out a program that could work for other models as well. However, unlike the case of classical neural networks on compact domains of $\mathbb{R}^N$, the same question for point clouds has a less definitive answer even after having determined all of the uniformly approximable functions. Each method explored here has their strengths and limitations but the presence of useful functions out-of-reach of PointNet and DeepSets opens the door for further research. However, these difference may largely dissapear when restricted to point clouds of fixed size, though we were not able to determine if that was the case.

There are many questions left over for future work. Is it possible to merge PointNet and DeepSets in a way to obtain greater approximation power? Are there other useful topologies on $\mathcal{F}(\Omega)$ for which new kinds of continuous neural networks can be constructed? Also, how much do these limitations matter in practice? Can we get concrete bounds for how close in uniform-norm we can get to Diam and Center? Also, neural networks are usually trained on L^2-loss not L^∞-loss, so perhaps these differences disappear when that is taken into account. But what kind of measure would we put on $\mathcal{F}(\Omega)$ so as to give integration, and L^2-loss, a concrete meaning? And finally, are there ways of developing architectures from the ground up with desirable yet different approximation capabilities on point clouds and what are they like?

References

George Cybenko. Approximation by superpositions of a sigmoidal function. *Mathematics of control, signals and systems*, 2(4):303–314, 1989.

Nicolas Fournier and Arnaud Guillin. On the rate of convergence in wasserstein distance of the empirical measure. *Probability Theory and Related Fields*, 162(3-4):707–738, 2015.

Kurt Hornik, Maxwell Stinchcombe, and Halbert White. Multilayer feedforward networks are universal approximators. *Neural networks*, 2(5):359–366, 1989.

Moshe Leshno, Vladimir Ya Lin, Allan Pinkus, and Shimon Schocken. Multilayer feedforward networks with a nonpolynomial activation function can approximate any function. *Neural networks*, 6(6):861–867, 1993.

James Raymond. Munkres. *Topology*. Prentice Hall, 2000.

Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 652–660, 2017.

Walter Rudin. *Real and complex analysis*. Tata McGraw-hill education, 2006.

M.b. Stinchcombe. Neural network approximation of continuous functionals and continuous functions on compactifications. *Neural Networks*, 12(3):467477, 1999. doi: 10.1016/s0893-6080(98)00108-7.

Cedric Villani. *Optimal transport: old and new*. Springer, 2009.

Manzil Zaheer, Satwik Kottur, Siamak Ravanbakhsh, Barnabas Poczos, Ruslan R Salakhutdinov, and Alexander J Smola. Deep sets. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems 30*, pp. 3391–3401. Curran Associates, Inc., 2017. URL `http://papers.nips.cc/paper/6931-deep-sets.pdf`.

A APPENDIX

Proof of Lemma 2.1. First we show that the r is a linear isometry. Since X is compact for $f \in C(X)$ there is a $p \in X$ so that $|f(p)| = \|f\|_X$. By density of D there is a sequence $p_n \in D$ that limits to p. So

$$\|f\|_X = |f(p)| = \lim_{n\to\infty} |f(p_n)| \leq \sup_{x\in D} |f(x)| = \|r(f)\|_D \leq \sup_{x\in X} |f(x)| = \|f\|_X$$

Next, since $C(X)$ is complete, so is its isometric image $r(C(X))$ and because $C_b(X)$ is complete that means $r(C(X))$ is closed. Thus,

$$r(C(X)) = r(\overline{\mathcal{N}'}) \subseteq \overline{r(\mathcal{N}')} \subseteq \overline{r(C(X))} = r(C(X))$$

where the first subset results from continuity. Thus $\overline{\mathcal{N}} = \overline{r(\mathcal{N}')} = r(C(X))$.

Finally, to show $r(C(X)) = U(D)$ note that every uniformly continuous function g on D continuously extends to a function on X (because D is dense in X) placing this extension in $C(X)$ and so $g \in r(C(X))$. The reverse inclusion follows as well because restriction preserves uniform continuity. □

Proof of Lemma 2.1. First we show that the r is a linear isometry. Since X is compact for $f \in C(X)$ there is a $p \in X$ so that $|f(p)| = \|f\|_X$. By density of D there is a sequence $p_n \in D$ that limits to p. So

$$\|f\|_X = |f(p)| = \lim_{n\to\infty} |f(p_n)| \leq \sup_{x\in D} |f(x)| = \|r(f)\|_D \leq \sup_{x\in X} |f(x)| = \|f\|_X$$

Next, since $C(X)$ is complete, so is its isometric image $r(C(X))$ and because $C_b(X)$ is complete that means $r(C(X))$ is closed. Thus,

$$r(C(X)) = r(\overline{\mathcal{N}'}) \subseteq \overline{r(\mathcal{N}')} \subseteq \overline{r(C(X))} = r(C(X))$$

where the first subset results from continuity. Thus $\overline{\mathcal{N}} = \overline{r(\mathcal{N}')} = r(C(X))$.

Finally, to show $r(C(X)) = U(D)$ note that every uniformly continuous function g on D continuously extends to a function on X (because D is dense in X) placing this extension in $C(X)$ and so $g \in r(C(X))$. The reverse inclusion follows as well because restriction preserves uniform continuity. □

Proof of Lemma 2.2. First we show that Max_f is d_H-continuous. Let $\epsilon > 0$. Since Ω is compact, f is uniformly continuous and so there is a $\delta > 0$ so that $|f(x) - f(y)| < \epsilon/2$ whenever $d(x, y) < 2\delta$. Now let $A, B \in \mathcal{K}(\Omega)$ and suppose $d_H(A, B) < \delta$. By definition this means $A \subseteq B_\delta$ and $B \subseteq A_\delta$. By the triangle inequality we have

$$|\mathrm{Max}_f(A) - \mathrm{Max}_f(B)| \leq |\mathrm{Max}_f(A) - \mathrm{Max}_f(A_\delta)| + |\mathrm{Max}_f(A_\delta) - \mathrm{Max}_f(B)|$$

Since A_δ is compact there is a $p \in A_\delta$ so that $\mathrm{Max}_f(A_\delta) = f(p)$. Observe that if $q \in K \subseteq A_\delta$ with $d(p, q) < 2\delta$ then $|f(p) - f(q)| < \epsilon/2$ and $f(p) = \mathrm{Max}_f(A_\delta) \geq \mathrm{Max}_f(K) \geq f(q)$. This implies $|\mathrm{Max}_f(A_\delta) - \mathrm{Max}_f(K)| < \epsilon/2$. In particular, since $p \in A_\delta$ there is an $a \in A$ such that $d(p, a) < \delta$, and since B is compact there is a $b \in B$ closest to p and so,

$$d(p, b) = d(p, B) \leq d_H(A_\delta, B) \leq d_H(A_\delta, A) + d_H(A, B) < 2\delta.$$

Thus $|\mathrm{Max}_f(A_\delta) - \mathrm{Max}_f(A)|$ and $|\mathrm{Max}_f(A_\delta) - \mathrm{Max}_f(B)|$ are less than $\epsilon/2$ and so $|\mathrm{Max}_f(A) - \mathrm{Max}_f(B)| < \epsilon$ as desired.

To see why $\mathrm{Ave}_{f,b}$ is d_W-continuous recall that the topology of d_W is the same as the weak-* topology for measures and so the map $\mu \mapsto \int f \,\mathrm{d}\mu$ is by definition continuous whenever $f \in C(\Omega)$.

$$\mathrm{Ave}_{f,b}(i_{\mathcal{P}}(A)) = \mathrm{Ave}_{f,b}\left(\frac{1}{n}\sum_{a\in A} \delta_a\right) = b + \frac{1}{n}\sum_{a\in A}\int f \,\mathrm{d}\delta_a = b + \frac{1}{n}\sum_{a\in A} f(a) = \mathrm{ave}_{f,b}(A)$$

Finally, it's clear that $\mathrm{Max}_f \circ i_{\mathcal{K}} = \max_f$. The other identity follows from the linearity of integration. □