Deep Sets

NIPS 2017

Original Authors: Manzil Zaheer, et al.

Presented by: Alisina Bayati

Academic Advisor: Professor Srinivasa Salapaka

Qualifying Exam Presentation

Department of Mechanical Science and Engineering,
University of Illinois Urbana Champaign

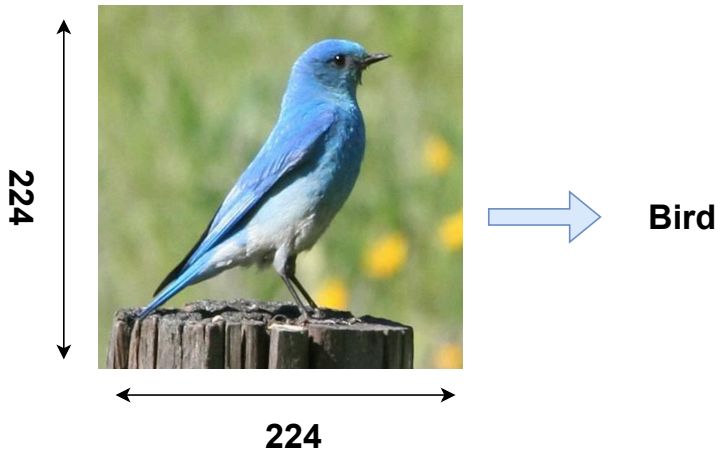
03/27/2024

Outline

1. Introduction
2. Problem Definition
   - Permutation Invariance
   - Permutation Equivariance
3. Applications and Results
4. Assessment: Contributions and Areas of Improvement

Introduction

- Traditionally, machine learning methodologies primarily focused on data of the forms of:
  - Fixed dimensional vectors: images, etc.
  - Ordered sequences: texts, etc.



- What happens if inputs are sets, where the data is
  - Unordered collection of objects
  - Number of objects can vary

Introduction

- Examples:
  - Supervised learning:

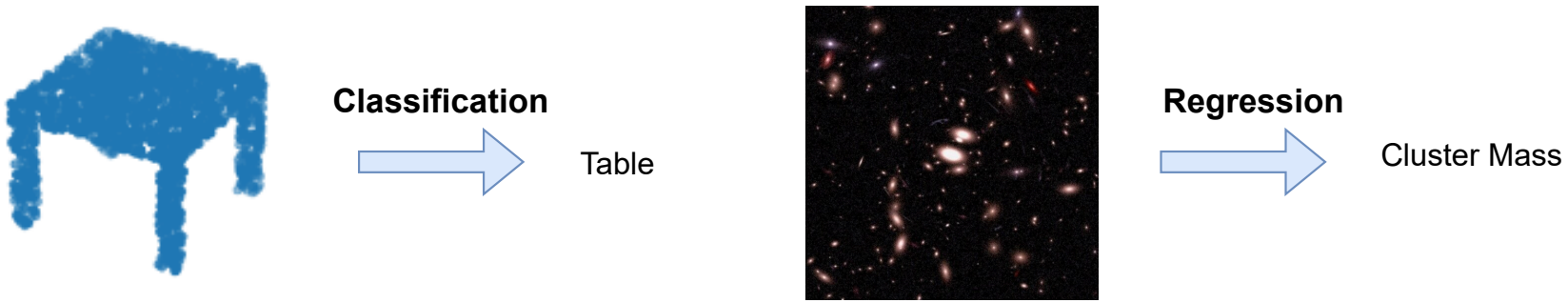


Figure: Point cloud classification [1] and red-shift estimation of galaxy clusters[2].

  - Unsupervised learning:
    - Set expansion: e.g., $\{lion, tiger, leopard\} \Rightarrow \{lion, tiger, leopard, cheetah, jaguar\}$
    - Anomaly detection
    - ...

[1] David Griffiths, *Point Cloud Classification with PointNet*.

[2] Massimo Meneghetti, *Weighing Simulated Galaxy Clusters Using Lensing and X-Ray*, Astronomy and Astrophysics.

Problem Definition

Permutation Invariance

A function $f : 2^{\mathfrak{X}} \to \mathcal{Y}$ acting on sets is **permutation invariant** if the output does not change under any permutation of the input set. For any permutation π:

$$f(\{x_1, \ldots, x_M\}) = f(\{x_{\pi(1)}, \ldots, x_{\pi(M)}\})$$

Permutation Equivariance

A function $f : \mathcal{X}^M \to \mathcal{Y}^M$ is **permutation equivariant** if upon permuting the input instances, the output labels are permuted in the same way. For any permutation π:

$$f([x_{\pi(1)}, \ldots, x_{\pi(M)}]) = [f_{\pi(1)}(x), \ldots, f_{\pi(M)}(x)]$$

Theorem 2: Permutation Invariance

Theorem 2.

A function $f(X)$ operating on a set X having elements from a countable universe, is a valid set function, i.e., invariant to the permutation of instances in X, **iff** it can be decomposed in the form

$$\rho\left(\sum_{x \in X} \phi(x)\right),$$

for suitable transformations ϕ and ρ.

Remarks:

- The Universal Approximation Theorem ensures that neural networks can closely approximate any continuous function and thus, the problem reduces to approximating ϕ and ρ functions.
- The model handles variable-length inputs, as set size is influencing only through $\sum_{x \in X} \phi(x)$.

Proof of Theorem 2: Countable Case

Sufficiency ($\Leftarrow$):

- For $\rho\left(\sum_{x \in X} \phi(x)\right)$, changing the order of summation does not affect the result, confirming invariance to permutation of elements of the input set $X \subseteq \mathfrak{X}$.

Necessity ($\Rightarrow$):

- Given countable $\mathfrak{X}$, there exists a bijective mapping $c : \mathfrak{X} \to \mathbb{N}$.
- By defining $\phi(x) = 4^{-c(x)}$, we ensure that $\sum_{x \in X} \phi(x)$ assigns a unique representation to each subset $X \subseteq \mathfrak{X}$.
- Construct ρ such that $f(X) = \rho\left(\sum_{x \in X} \phi(x)\right)$, encapsulating the set function f.

Extension to The Uncountable Case

- The extension to when $\mathfrak{X}$ is uncountable, e.g., $\mathfrak{X} = \mathbb{R}$, is not so trivial.
- Only proved for the case of fixed set size, e.g., $\mathcal{X} = \mathbb{R}^M$, instead of $\mathcal{X} = 2^{\mathfrak{X}} = 2^{\mathbb{R}}$.
- Without loss of generality, assume $\mathfrak{X} = [0, 1]$ and thus, $\mathcal{X} = [0, 1]^M$.
- To handle ambiguity due to permutation, we often define the domain to be $\mathcal{X} = \{(x_1, \ldots, x_M) \in [0, 1]^M : x_1 \leq x_2 \leq \ldots \leq x_M\}$.

Theorem 7.

$f : [0, 1]^M \to \mathbb{R}$ is a permutation invariant continuous function iff it has the representation

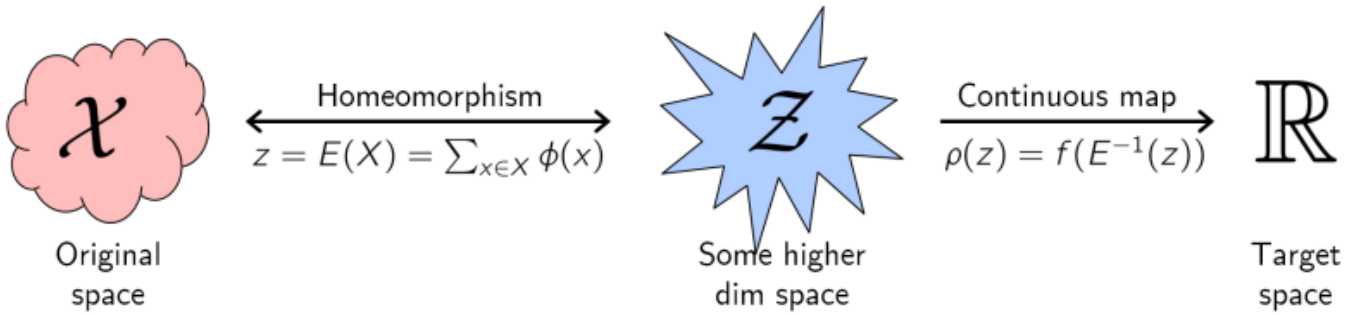
$$f(x_1, \ldots, x_M) = \rho\left(\sum_{m=1}^{M} \phi(x_m)\right)$$

for some independent and continuous outer and inner functions $\rho : \mathbb{R}^{M+1} \to \mathbb{R}$ and $\phi : \mathbb{R} \to \mathbb{R}^{M+1}$ respectively.

Theorem 7 - Proof

Sketch of the Proof. Sufficiency is straightforward. To prove necessity,

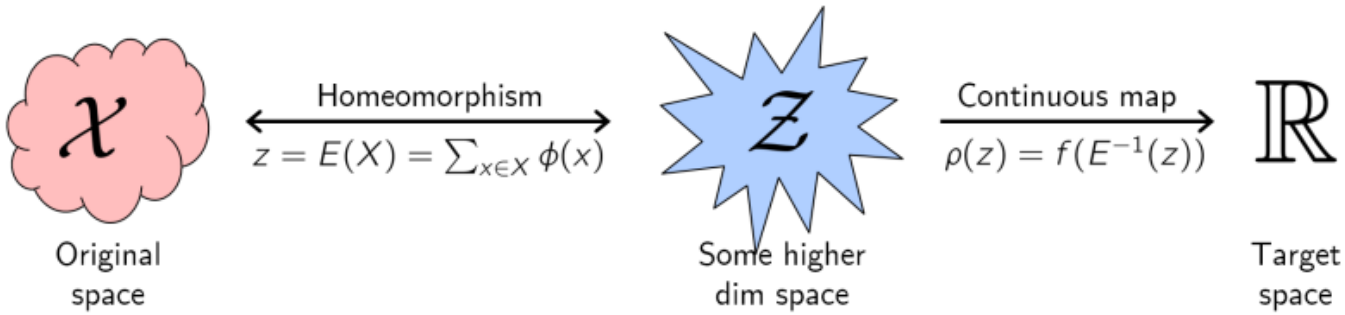
- We establish unique embeddings E for each $X \in [0,1]^M$ defined as $E(X) = \left(\sum_{m=1}^{M} \phi(x_m)\right)$, where $\phi : \mathbb{R} \to \mathbb{R}^{M+1}$ is $\phi(x) = [1, x, x^2, \ldots, x^M]$.



Theorem 7 - Proof

Sketch of the Proof. Sufficiency is straightforward. To prove necessity,

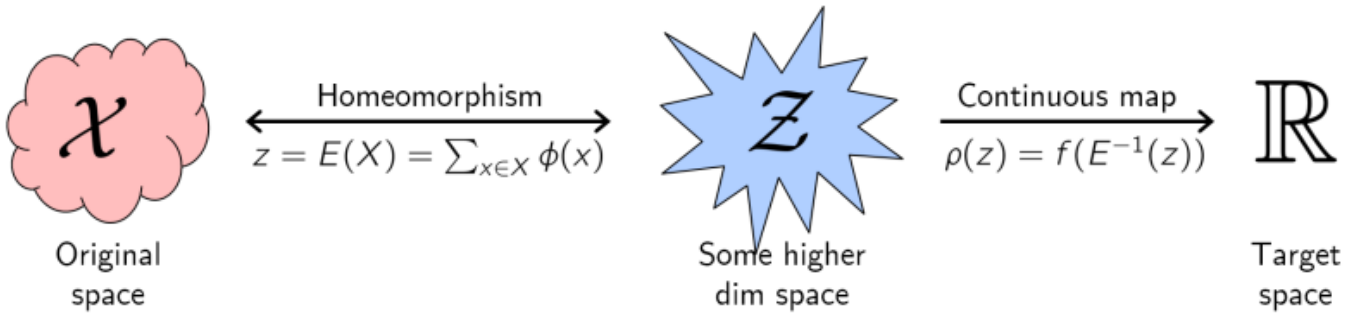
- We establish unique embeddings E for each $X \in [0,1]^M$ defined as $E(X) = \left(\sum_{m=1}^{M} \phi(x_m)\right)$, where $\phi : \mathbb{R} \to \mathbb{R}^{M+1}$ is $\phi(x) = [1, x, x^2, \ldots, x^M]$.
- Let $\mathcal{Z}$ be the image of $[0,1]^M$ under E, and thus compact.



Theorem 7 - Proof

Sketch of the Proof. Sufficiency is straightforward. To prove necessity,

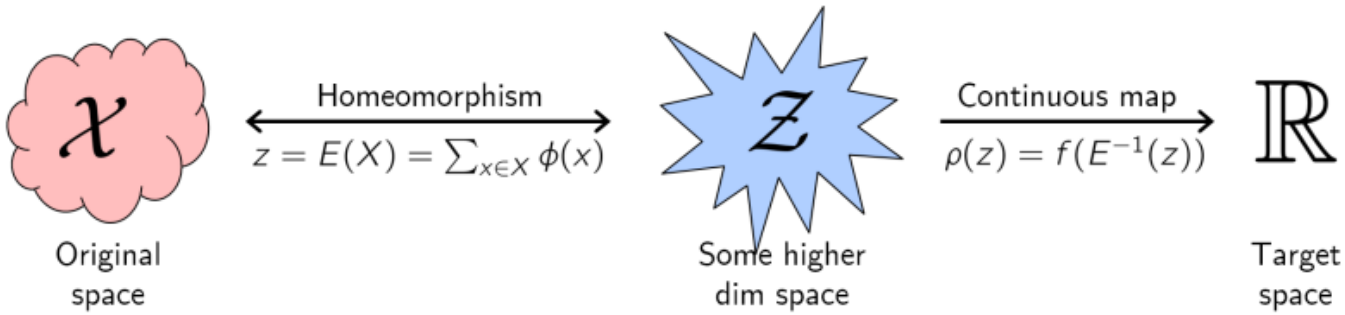
- We establish unique embeddings E for each $X \in [0,1]^M$ defined as $E(X) = \left(\sum_{m=1}^{M} \phi(x_m)\right)$, where $\phi : \mathbb{R} \to \mathbb{R}^{M+1}$ is $\phi(x) = [1, x, x^2, \ldots, x^M]$.
- Let $\mathcal{Z}$ be the image of $[0,1]^M$ under E, and thus compact.
- We **claim** that $E : \mathcal{X} \to \mathcal{Z}$ is a homeomorphism, where $\mathcal{X} = \{(x_1, \ldots, x_M) \in [0,1]^M : x_1 \leq x_2 \leq \ldots \leq x_M\}$



Theorem 7 - Proof

Sketch of the Proof. Sufficiency is straightforward. To prove necessity,

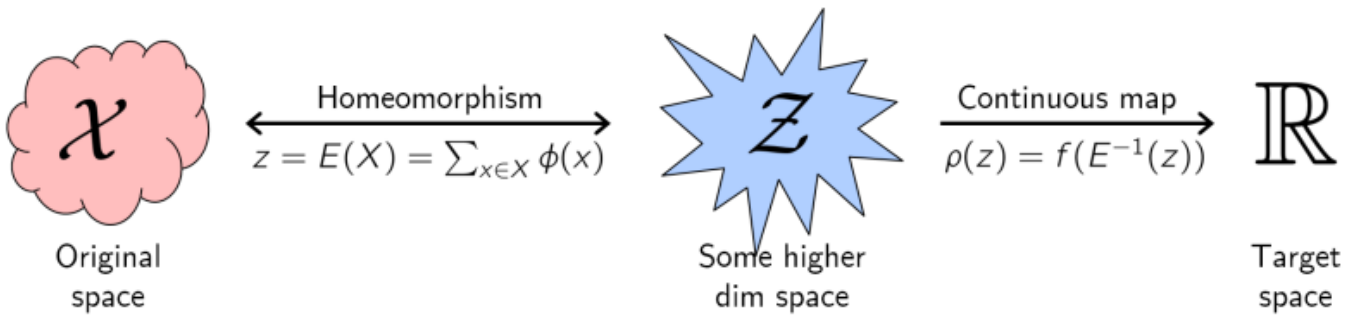
- We establish unique embeddings E for each $X \in [0,1]^M$ defined as $E(X) = \left(\sum_{m=1}^{M} \phi(x_m)\right)$, where $\phi : \mathbb{R} \to \mathbb{R}^{M+1}$ is $\phi(x) = [1, x, x^2, \ldots, x^M]$.
- Let $\mathcal{Z}$ be the image of $[0,1]^M$ under E, and thus compact.
- We **claim** that $E : \mathcal{X} \to \mathcal{Z}$ is a homeomorphism, where $\mathcal{X} = \{(x_1, \ldots, x_M) \in [0,1]^M : x_1 \leq x_2 \leq \ldots \leq x_M\}$
- We exhibit a continuous map $\rho : \mathcal{Z} \to \mathbb{R}$ that recovers f from $E(X)$ via $\rho(z) = f(E^{-1}(z))$. (since f and E^{-1} are both continuous)



Theorem 7 - Proof

Sketch of the Proof. Sufficiency is straightforward. To prove necessity,

- We establish unique embeddings E for each $X \in [0,1]^M$ defined as $E(X) = \left(\sum_{m=1}^{M} \phi(x_m)\right)$, where $\phi : \mathbb{R} \to \mathbb{R}^{M+1}$ is $\phi(x) = [1, x, x^2, \ldots, x^M]$.
- Let $\mathcal{Z}$ be the image of $[0,1]^M$ under E, and thus compact.
- We **claim** that $E : \mathcal{X} \to \mathcal{Z}$ is a homeomorphism, where $\mathcal{X} = \{(x_1, \ldots, x_M) \in [0,1]^M : x_1 \leq x_2 \leq \ldots \leq x_M\}$
- We exhibit a continuous map $\rho : \mathcal{Z} \to \mathbb{R}$ that recovers f from $E(X)$ via $\rho(z) = f(E^{-1}(z))$. (since f and E^{-1} are both continuous)
- We conclude that $\forall X \in \mathcal{X}$, $f(X) = \rho(E(X)) = \rho\left(\sum_{m=1}^{M} \phi(x_m)\right)$, where ρ and ϕ are independent and continuous.



Proof of the Claim: Lemma 4.

Lemma 4.

Let $\mathcal{X} = \{(x_1, \ldots, x_M) \in [0,1]^M : x_1 \leq x_2 \leq \ldots \leq x_M\}$. The sum-of-power mapping $E : \mathcal{X} \to \mathbb{R}^{M+1}$ defined by the following coordinate functions is **injective**.

$$Z_q = E_q(\mathcal{X}) := \sum_{m=1}^{M} (x_m)^q, \quad q = 0, \ldots, M.$$

Proof. Let $u, v \in \mathcal{X}$, and assume $E(u) = E(v)$. Define polynomials:

$$P_u(x) = \prod_{m=1}^{M} (x - u_m), \quad P_v(x) = \prod_{m=1}^{M} (x - v_m).$$

Expanding these polynomials gives us:

$$P_u(x) = x^M - a_1 x^{M-1} + \cdots + (-1)^M a_M,$$
$$P_v(x) = x^M - b_1 x^{M-1} + \cdots + (-1)^M b_M,$$

Proof of the Claim: Lemma 4.

By Newton-Girard formulae, these coefficients relate to sums of powers:

$$a_m = \frac{1}{m} \sum_{1 \le j_1 < \cdots < j_m \le M} u_{j_1} \cdots u_{j_m}, \quad b_m = \frac{1}{m} \sum_{1 \le j_1 < \cdots < j_m \le M} v_{j_1} \cdots v_{j_m}.$$

Coefficients a_m and b_m can be expressed using determinants involving $E(u)$ and $E(v)$:

$$a_m = \frac{1}{m} \det \begin{pmatrix} E_1(u) & 1 & 0 & \cdots & 0 \\ E_2(u) & E_1(u) & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ E_m(u) & E_{m-1}(u) & E_{m-2}(u) & \cdots & 1 \end{pmatrix},$$

$$b_m = \frac{1}{m} \det \begin{pmatrix} E_1(v) & 1 & 0 & \cdots & 0 \\ E_2(v) & E_1(v) & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ E_m(v) & E_{m-1}(v) & E_{m-2}(v) & \cdots & 1 \end{pmatrix}.$$

Since $E(u) = E(v)$, it follows that $P_u(x) = P_v(x)$, and hence $u = v$.

Proof of the Claim: Theorem 5

Theorem 5. [1]

The function $f : \mathbb{C}^M \to \mathcal{M}$, which associates every $a \in \mathbb{C}^M$ to the multiset of roots, $f(a) \in \mathcal{M}$, of the monic polynomial formed using a as the coefficient i.e., $x^M + a_1 x^{M-1} + \ldots + (-1)^{M-1} a_{M-1} x + (-1)^M a_M$, is a homeomorphism.

Remark: This implies that (complex) roots of a polynomial depend continuously on the coefficients.

[1] Branko Ćurgus, Vania Mascioni, , *Roots and polynomials as Homeomorphic spaces, Expositiones Mathematicae*.

Proof of The Claim: Lemma 6.

Lemma 6

Let $\mathcal{X} = \{(x_1, \ldots, x_M) \in [0,1]^M : x_1 \leq x_2 \leq \ldots \leq x_M\}$. We define the sum-of-power mapping $E : X \to \mathbb{Z}$ by the coordinate functions

$$Z_q := E_q(X) := \sum_{m=1}^{M} (x_m)^q, \quad q = 0, \ldots, M,$$

where $\mathbb{Z}$ is the range of the function. The function E has a continuous inverse mapping.

Proof

- As in Lemma 4, pick a $u \in \mathcal{X}$ and construct the polynomial $P_u(x) = \prod_{m=1}^{M}(x - u_m)$, where u is its root.
- Expanding $P_u(x)$ gives $x^M - a_1 x^{M-1} + \ldots + (-1)^M a_M$, where a_m are the coefficients of the polynomial.

Proof of the Claim: Lemma 6.

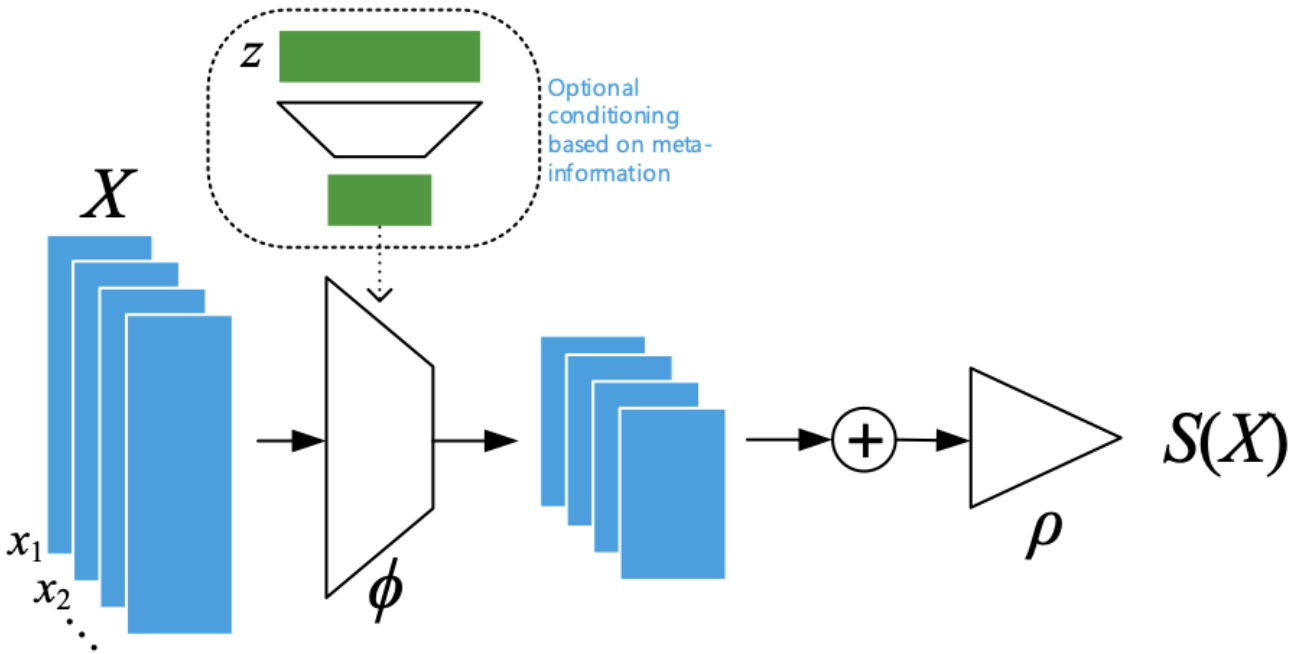
- The coefficients a_m are expressible uniquely as functions of $z = E(u)$ using the Newton-Girard formula, as

$$a_m = \frac{1}{m} \det \begin{pmatrix} z_1 & 1 & 0 & \cdots & 0 \\ z_2 & z_1 & 1 & \cdots & 0 \\ z_3 & z_2 & z_1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ z_{m-1} & z_{m-2} & z_{m-3} & \cdots & 1 \\ z_m & z_{m-1} & z_{m-2} & \cdots & z_1 \end{pmatrix}$$

- Coefficients a_m are given by determinants, which are continuous functions of z.
- By Theorem 5, the roots u of $P_u(x)$ depend continuously on coefficients, and thus on z.
- Since the determinants are polynomials in z, and polynomials are continuous, u (as roots) are continuous in z.
- $u = E^{-1}(z)$ is continuous in z, and thus, E^{-1} is continuous.

Structure of Permutation Invariant Networks

- Each instance x_m, is transformed into a representation $\phi(x_m)$.
- These representations are summed: $\sum_m \phi(x_m)$, and processed with a deep neural network ρ.
- Optionally, with additional meta-information z, ϕ can be conditioned on z to produce $\phi(x_m|z)$, allowing for context-specific representations.



Lemma 3: Permutation Equivariance

Lemma 3.

The function $f_\Theta : \mathbb{R}^M \to \mathbb{R}^M$ defined as $f_\Theta(x) = \sigma(\Theta x)$ where $\Theta \in \mathbb{R}^{M \times M}$, is permutation equivariant iff all the off-diagonal elements of Θ are tied together and all the diagonal elements are equal as well. That is,

$$\Theta = \lambda I + \gamma(11^T), \quad \text{where} \quad \lambda, \gamma \in \mathbb{R},\ 1 = [1, \ldots, 1]^T \in \mathbb{R}^M,$$

and $I \in \mathbb{R}^{M \times M}$ is the identity matrix.

Sketch of the Proof:

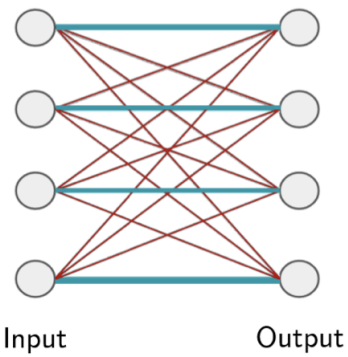
- For each network layer,

  $$f_\Theta(\pi x) = \pi f_\Theta(x) \implies \sigma(\Theta \pi x) = \pi \sigma(\Theta x) = \sigma(\pi \Theta x)$$

  Thus, it is sufficient to derive conditions so that

  $$\Theta \pi = \pi \Theta$$

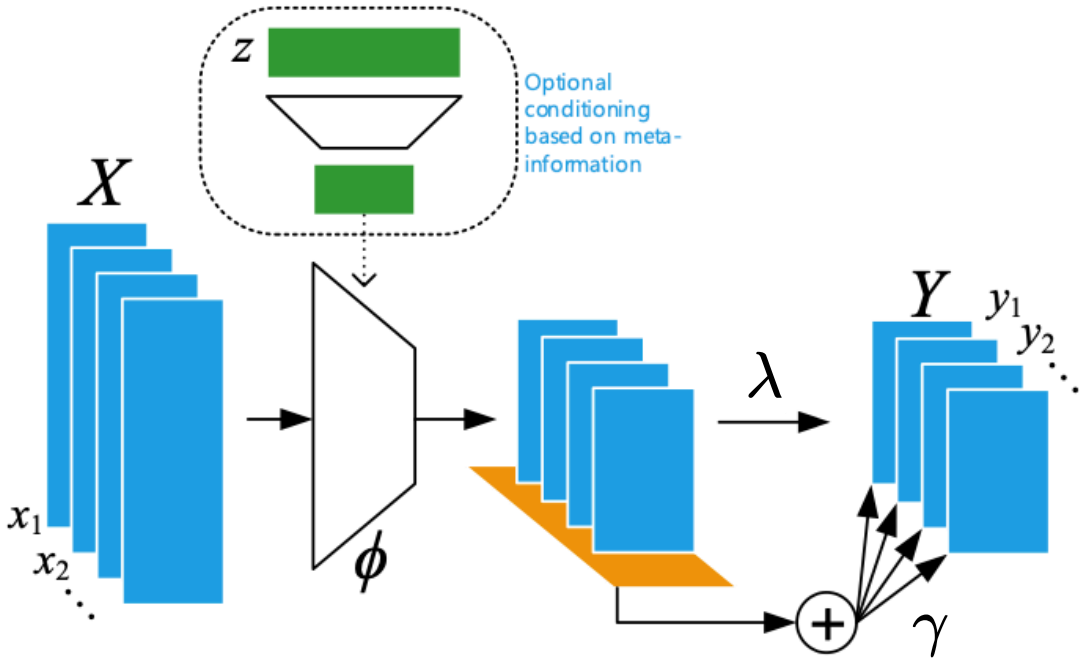
  for all permutation matrices π.



Structure of Permutation Equivariant Networks

Remarks:

- Note that $\Theta x = \lambda x + \gamma(1^T x)1$ where the first term is permutation equivariant and the second term is permutation invariant.
- In practice, sometimes $f(x) = \sigma(\lambda I x + \gamma \text{maxpool}(x)1)$ works better.



Applications and Results

Sum of Digits

- Find sum of a given set of
  - digits, or
  - images of handwritten digits
- Train on sets of size 10 at most, while at test time we use examples of length up to 100
- DeepSets generalize much better than LSTM or GRU

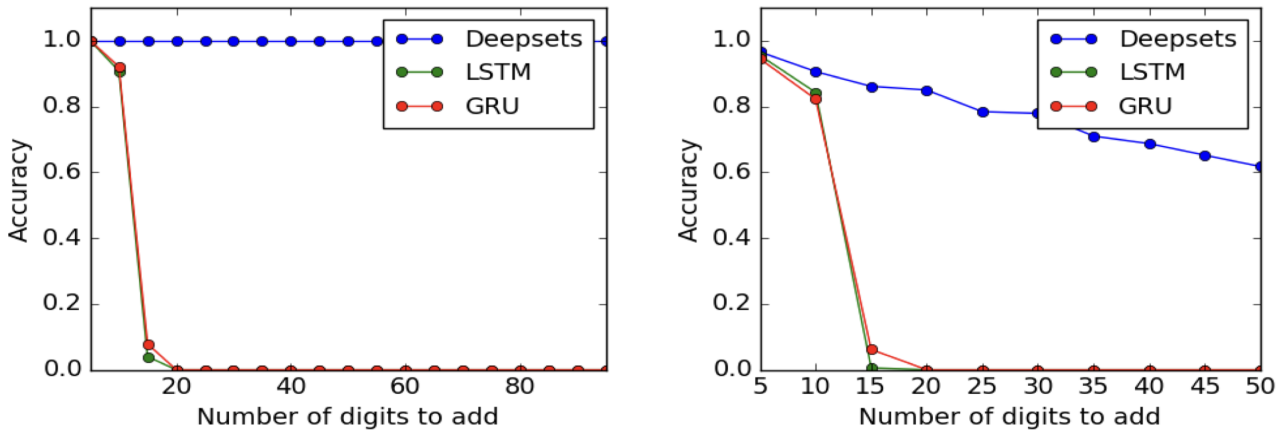
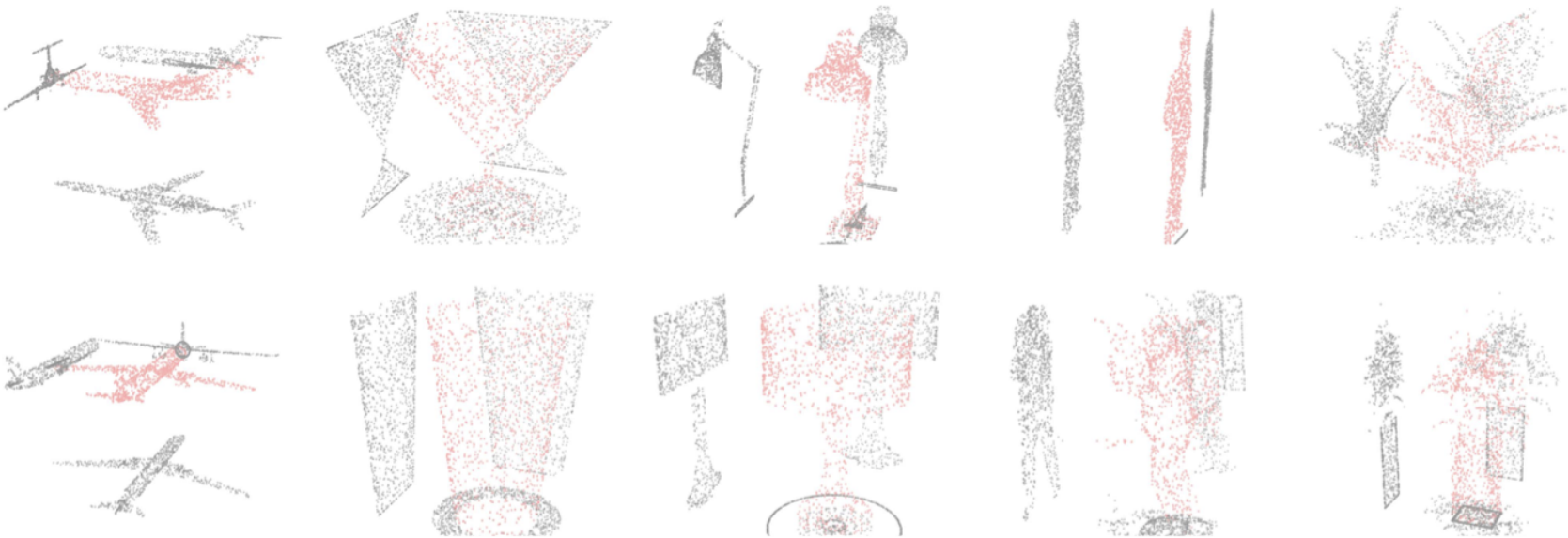


Figure: Accuracy of digit summation with text (left) and image (right) inputs. All approaches are trained on tasks of length 10 at most, tested on examples of length up to 100. We see that DeepSets generalizes better.

Applications and Results

Point Cloud Classification

- A point-cloud is a set of 3D coordinates of an underlying sampled surface.
- Note the point cloud will be permutation invariant.
- Application in face regognition of cameras, autonomous vehicles, gesture estimation devices (Xbox Kinect, etc.)



Applications and Results

Point Cloud Classification

- Benchmarking DeepSets on ModelNet40 which containts $>$ 10000 3D objects belonging to 40 classes by trating point clouds as set of points.

Model	Instance Size	Representation	Accuracy
3DShapeNets [25]	30^3	voxels (using convolutional deep belief net)	77%
VoxNet [26]	32^3	voxels (voxels from point-cloud + 3D CNN)	83.10%
MVCNN [21]	$164 \times 164 \times 12$	multi-vew images (2D CNN + view-pooling)	90.1%
VRN Ensemble [27]	32^3	voxels (3D CNN, variational autoencoder)	95.54%
3D GAN [28]	64^3	voxels (3D CNN, generative adversarial training)	83.3%
DeepSets	$\mathbf{5000 \times 3}$	point-cloud	$90 \pm .3\%$
DeepSets	$\mathbf{100 \times 3}$	point-cloud	$82 \pm 2\%$

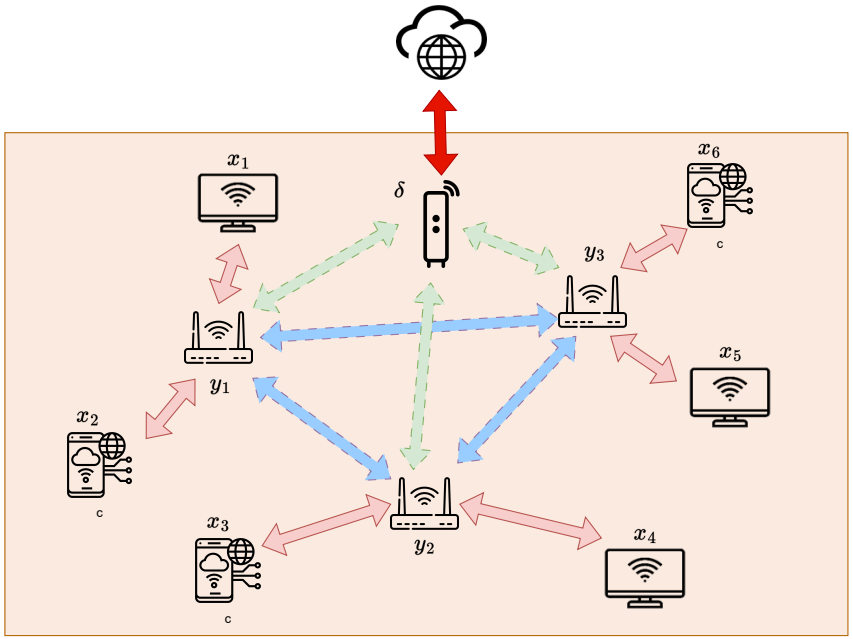
Applications and Results

Other Applications

- Improved red-shift estimation for predicting the mass of galaxy clusters from photometric data
- Set anomaly detection and set expansion, image tagging, ...

Where we used the idea

- Estimating performance metrics of a WIFI mesh network with multiple clients, fixed number of routers and the internet gateway.



Assessment: Strengths and Areas for Improvement

Strengths

- Introduced a method for integrating system knowledge (i.e.,, permutation invariance or equivariance properties) into the learning process, which can be applied in various real world systems.
- Demonstrated broad applicability in various areas with minimal modifications and showed competitive performance in all of them.
- Established solid theoretical proofs combined with a variety of applications in different areas.

Areas for Improvement

- Expansion of theoretical proofs to cover variable input sizes in the uncountable case.

Thank You for your attention!