How to AI (Almost) Anything

Lecture 3 – Common model architectures

Paul Liang
Assistant Professor
MIT Media Lab & MIT EECS

https://pliang279.github.io
ppliang@mit.edu
@pliang279

Assignments for This Coming Week

For project:

- 4-page project proposal due tonight (2/25), email to me.
- Meet with me after class 2-3pm if need feedback about proposal ideas.

Reading assignment due tomorrow Wednesday (2/26).

This Thursday (2/27): first reading discussion on **data and learning**.

Bitter lesson

Grokking/double descent

Logistics – Reading Assignments

Roles and Grading

Role assignments for every reading are linked here

🧑‍🔬 Scientific Peer Reviewer

Task: Complete a full review of the paper, recommending acceptance or rejection, and address all prompts in the review form (e.g., technical soundness, clarity, originality, significance). See an example of review instructions here (navigate down to 'review form').

Grading

1️⃣ (Unacceptable)

- Review is incomplete or missing major components.
- Minimal effort with vague or unsubstantiated comments.
- No clear recommendation or justification provided.

2️⃣ (Needs Improvement)

- Review touches on a few relevant points but lacks depth and critical evaluation.
- Feedback is superficial and misses key aspects of the paper.

3️⃣ (Adequate)

- Provides a basic review covering most areas.
- Offers some useful feedback but misses deeper analysis or constructive suggestions.
- The recommendation is justified but could be stronger.

4️⃣ (Good)

- Thoughtful, thorough review with balanced criticism and praise.
- Clearly explains the paper's strengths and weaknesses.
- Provides helpful suggestions for improvement.

multisensory intelligence

Logistics – Reading Assignments

Roles per Week

File Edit View Insert Format Data Tools Extensions Help

Name	Email	reading 1 (2/27)	reading 2 (3/6)	reading 3 (3/13)	reading 4 (3/20)	reading 5 (4/17)	reading 6 (4/24)	reading 7 (5/1)
Chenyu Zhang	chenyu_zhang@gse.harvard.edu	1		5		9		
Gonzalo Minuto	gminuto@mit.edu	2		6		10		
Valdemar Danry	vdanry@media.mit.edu	3		7			1	
Jiao Zhao	jiaoz17@mit.edu	4		8			2	
Kida Huang	chungtah@mit.edu	5		9			3	
(Michelle) Minsol Kim	minsol@mit.edu	6		10			4	
Ozgun Kilic Afsar	ozgun@media.mit.edu	7			1		5	
Cathy Fang	catfang@media.mit.edu	8			2		6	
Kumar Tanmay	kumartanmay@fas.harvard.edu	9			3		7	
Chenjie Xiong	xiongc@mit.edu	10			4		8	
Mike Jiang	mhjiang@mit.edu		1		5		9	
Mingyang Sun	msun14@mit.edu		2		6		10	
Yujia Qian	yjqian19@mit.edu		3		7			1
Connie Cheng	conniec@mit.edu		4		8			2
Sam Chin	chins@mit.edu		5		9			3
Nikhil Behari	nbehari@mit.edu		6		10			4
Qingyun Liu	qingyun_liu@gsd.harvard.edu		7			1		5
Richa Gupta	richag@mit.edu		8			2		6
Chance Jiajie Li	jiajie@mit.edu		9			3		7
Nelson Hidalgo	nelsonh@mit.edu		10			4		8
Nomy Yu	nomyyu@mit.edu			1		5		9
Nicolas Stas	nstas@mit.edu			2		6		10
Kushagra Tiwary	ktiwary@mit.edu			3		7		
Anku Rani	ankurani@mit.edu			4		8		

Legend

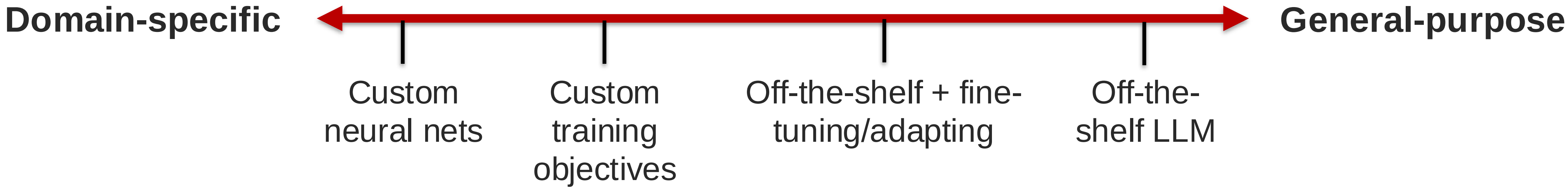
Index	Paper	Role
1	First	Scientific Peer Reviewer
2	First	Research Detective
3	First	Academic Researcher
4	First	Implementation Explainer
5	First	Impact Strategist
6	Second	Scientific Peer Reviewer
7	Second	Research Detective
8	Second	Academic Researcher
9	Second	Implementation Explainer
10	Second	Impact Strategist

reminder of role duties and grading

Lecture Outline

1. A unifying paradigm of model architectures
2. Temporal sequence models
3. Spatial convolution models
4. Models for sets and graphs

Two General Modeling Paradigms



Your decision will depend on many factors.

multisensory intelligence

Designing Models for Data

What is a good model?

One that captures the:
- right semantic information
- at the right granularity
- using an appropriate amount of data and labels
- with the right resource constraints
- with the right level of usability (explainability, accessibility, etc.)
- and more…

Domain-specific ⟷ **General-purpose**

Lecture Topics

(subject to change, based on student interests and course discussions)

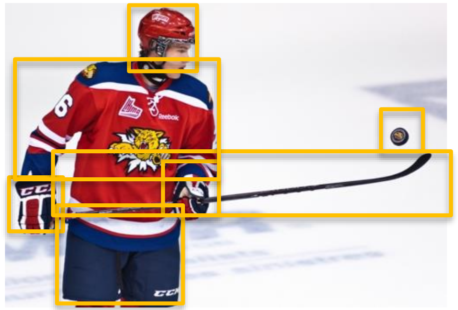
Domain-specific/custom models

Week 4 (2/25): Common model architectures

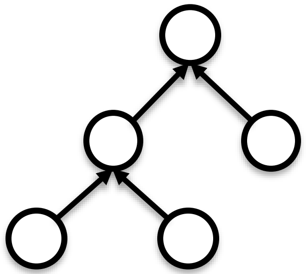
Week 5 (3/4): Multimodal connections and alignment

Week 6 (3/11): Multimodal interactions and fusion

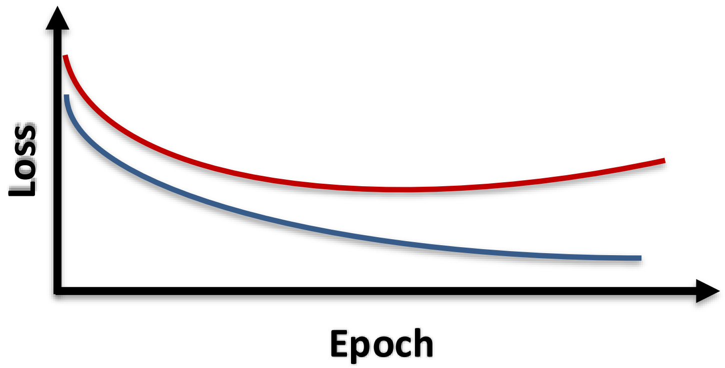
Week 7 (3/18): Cross-modal transfer



Spatial



Hierarchical



multisensory intelligence

Lecture Topics

(subject to change, based on student interests and course discussions)

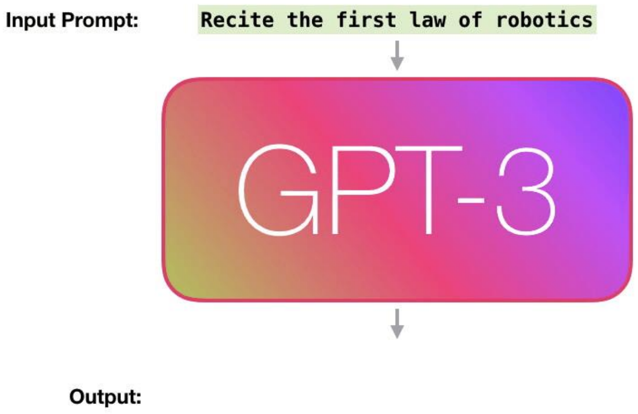
General architectures and adapting pre-trained models

Week 9 (4/1): Pre-training, scaling, fine-tuning LLMs

Week 10 – No class, member's week

Week 11 (4/15): Large multimodal models

Week 12 (4/22): Modern generative AI

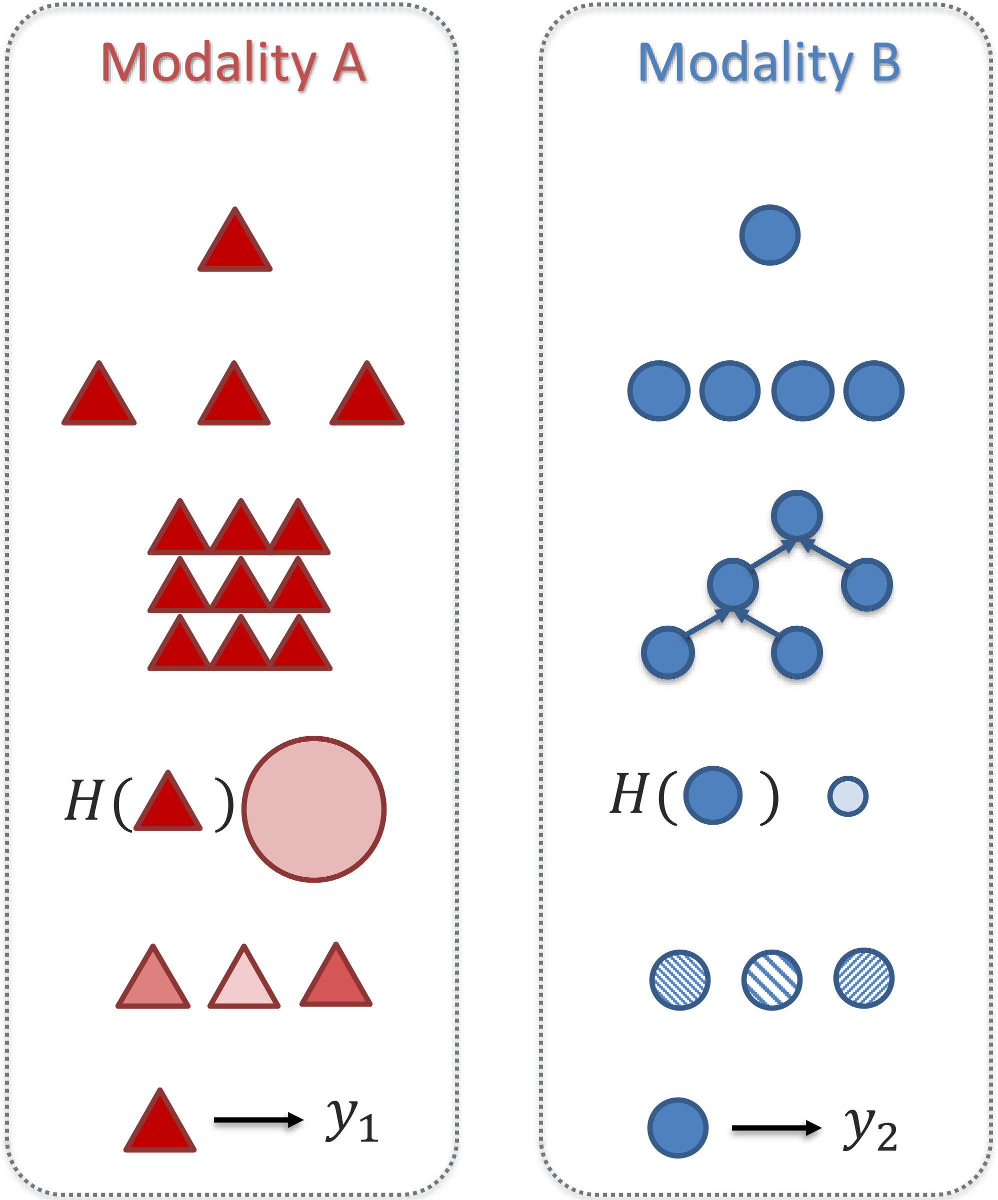


An armchair in the shape of an avocado

multisensory intelligence

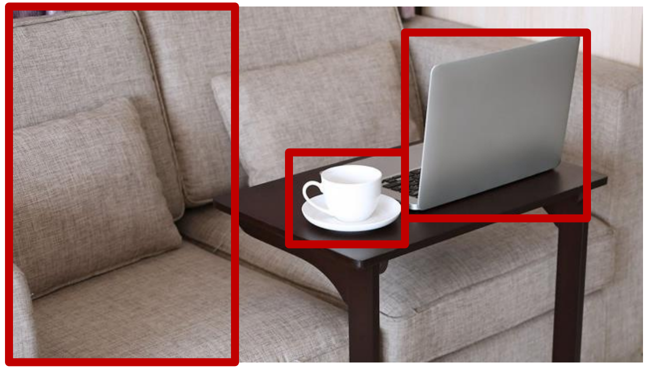
Modality Profile

1. **Element representations:**
   Discrete, continuous, granularity
2. **Element distributions:**
   Density, frequency
3. **Structure:**
   Temporal, spatial, latent, explicit
4. **Information:**
   Abstraction, entropy
5. **Noise:**
   Uncertainty, noise, missing data
6. **Relevance:**
   Task, context dependence



Modality Profile

The distribution of individual elements within that modality.



*A **teacup** on the **right** of a **laptop** in a **clean room**.*

1 **Distribution:** discrete or continuous, support

{  ,  ,  }

{teacup, right, laptop, clean, room}

Modality Profile

The frequency at which elements appear or are sampled.

A teacup on the right of a laptop in a clean room.

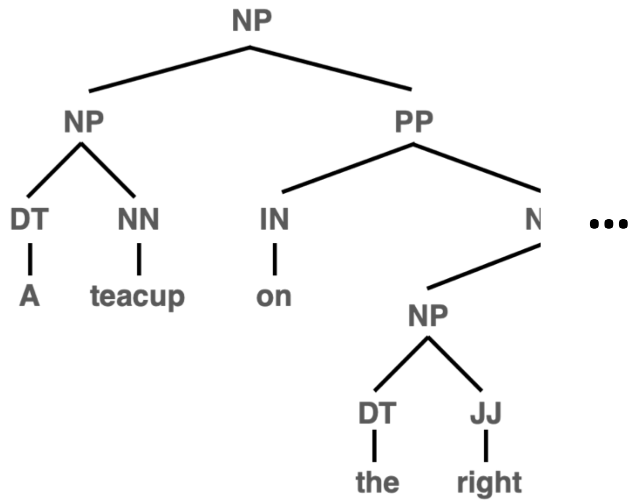
2 **Granularity:** sampling rate and frequency

objects per image

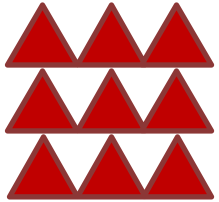
words per minute

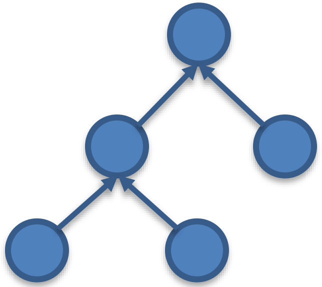
Modality Profile

The way elements compose with each other to form entire data.



3 **Structure:** static, temporal, spatial, hierarchical





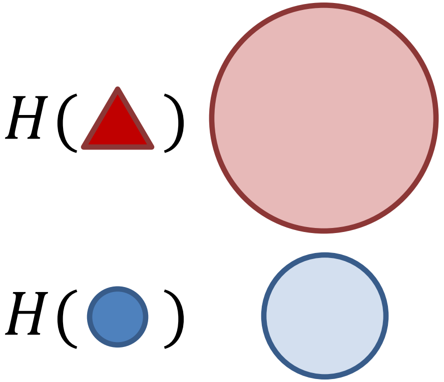
multisensory intelligence

Modality Profile

The total information contained in the elements and their composition.

A teacup on the right of a laptop
in a clean room.

4 **Information:** entropy and density



multisensory intelligence

Modality Profile

The natural imperfections in the data modality.

A teacup on the right of a laptop in a clean room.

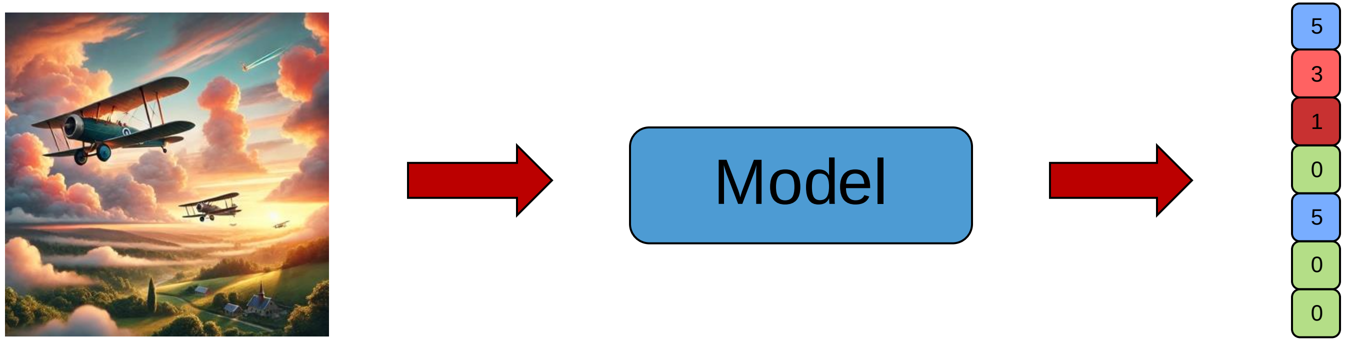
5 **Noise:** uncertainty, signal-to-noise ratio, missing data

teacup → teacip

right → rihjt

Unified View of Deep Learning Models

1. Learning representations



2. Combining representations (information aggregation)

multisensory intelligence

Unified View of Deep Learning Models

Composing differentiable functions and training objectives.

1. Basic representation building blocks for each element

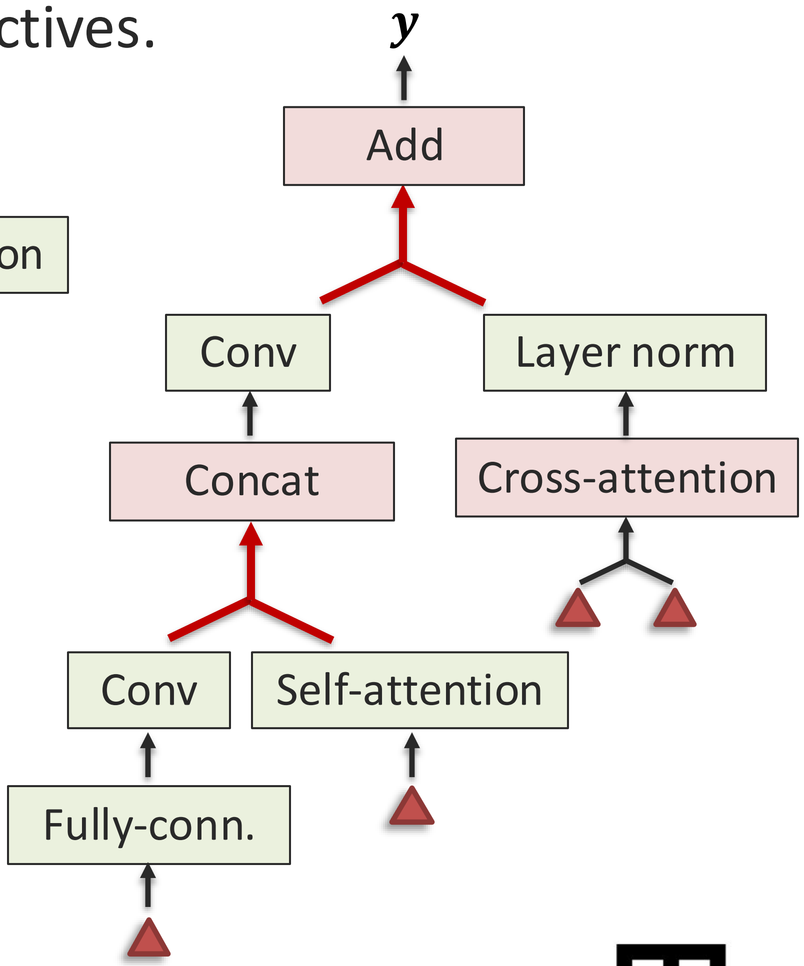
Fully-connected | ReLU | Layer norm | Conv | Self-attention

2. Basic information aggregation blocks

Concat | Cross-attention | Add | Max

3. Compute loss function

4. Take gradients, update with stochastic gradient descent

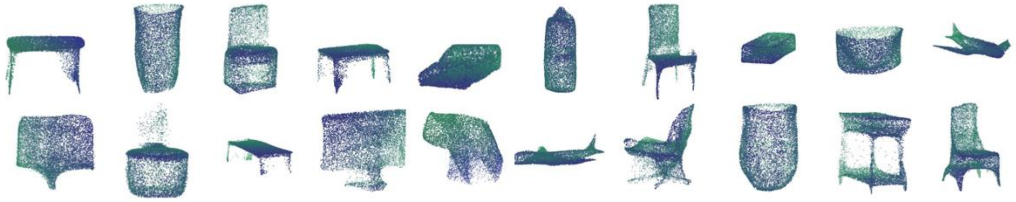


multisensory intelligence

A Simple Classification Example

Sets and point clouds

Sets



Point clouds

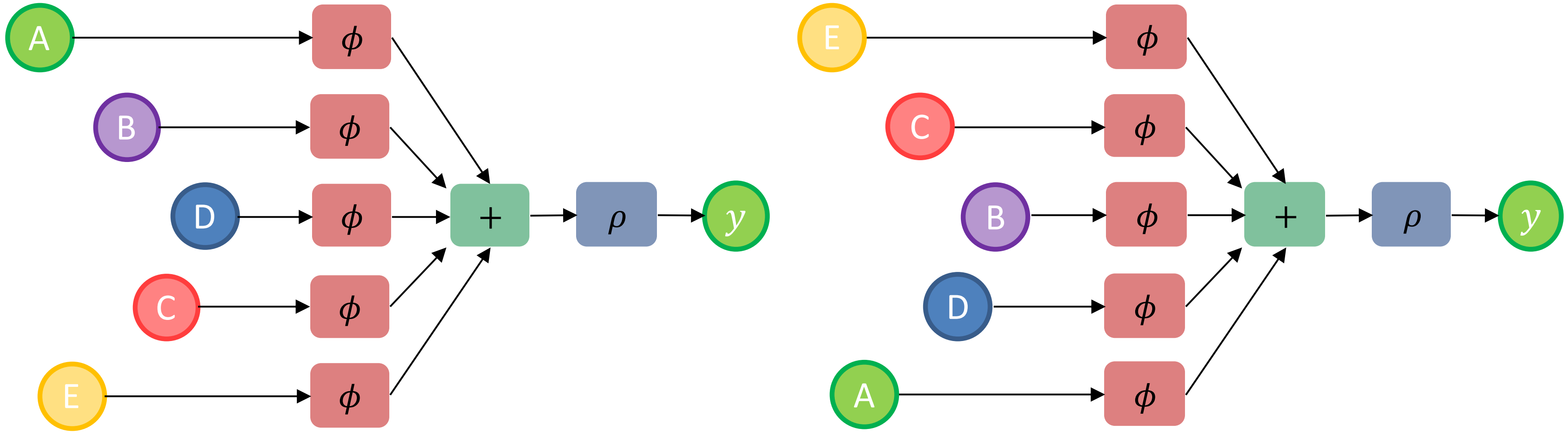
Set anomaly detection
Set expansion
Set completion
Point cloud classification
Point cloud generation

[Zaheer et al., DeepSets. NeurIPS 2017]
[Qi et al., PointNet: Deep Learning on Point Sets for 3D Classification and Segmentation. CVPR 2017]

A Simple Classification Example

Models for set-based data must be invariant to element order.

1. Parameter sharing for each set element
2. Permutation invariant aggregation function



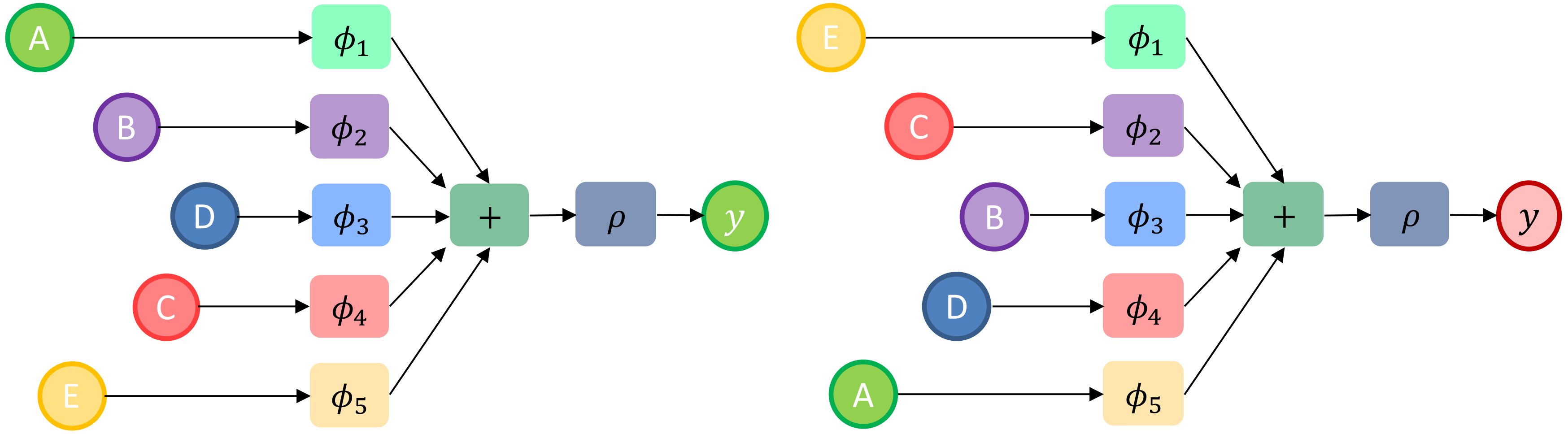
Giving ABDCE also gives ECBDA, BCAED etc…

[Zaheer et al., DeepSets. NeurIPS 2017]
[Qi et al., PointNet: Deep Learning on Point Sets for 3D Classification and Segmentation. CVPR 2017]

A Simple Classification Example

Models for set-based data must be invariant to element order.

1. No parameter sharing for each set element
2. Permutation invariant aggregation function



Need to give ABDCE, then ECBDA, and more…. 5! more samples needed

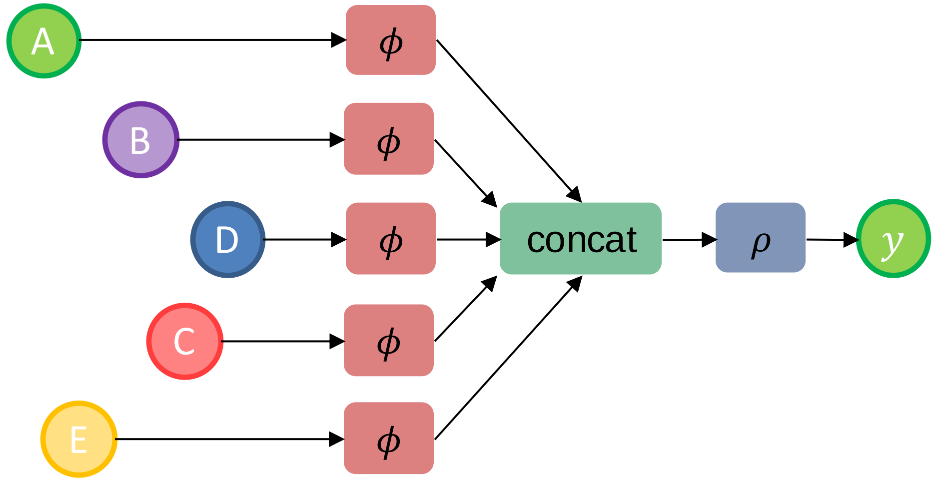
[Zaheer et al., DeepSets. NeurIPS 2017]
[Qi et al., PointNet: Deep Learning on Point Sets for 3D Classification and Segmentation. CVPR 2017]

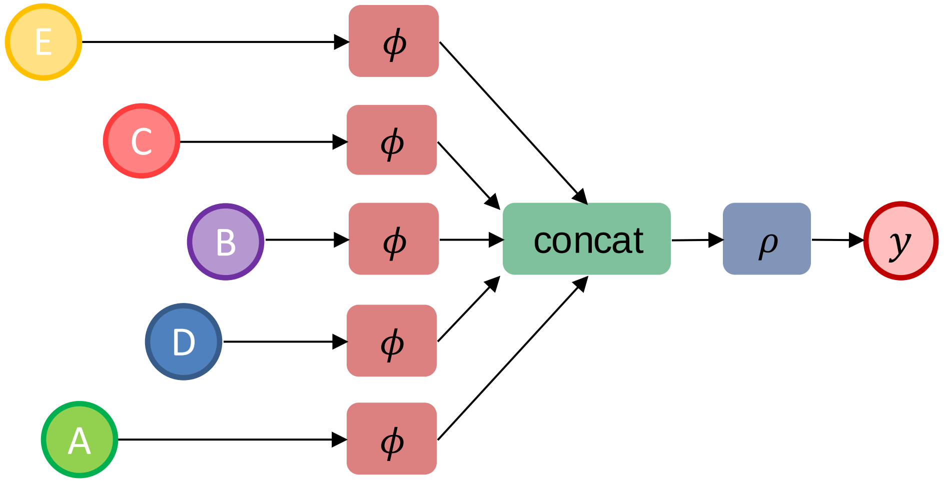
multisensory intelligence

A Simple Classification Example

Models for set-based data must be invariant to element order.

1. Parameter sharing for each set element
2. Not permutation invariant aggregation function





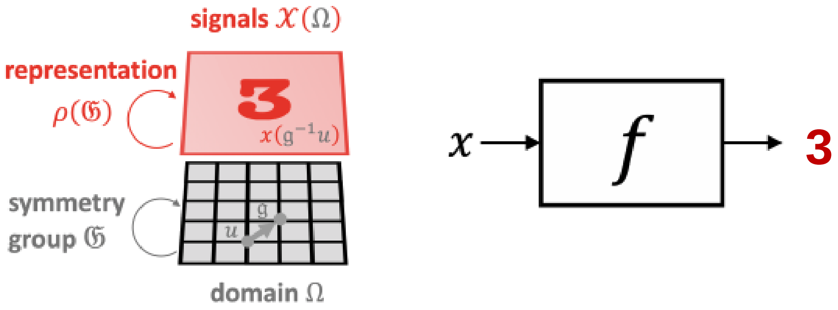
Need to give ABDCE, then ECBDA, and more…. 5! more samples needed

[Zaheer et al., DeepSets. NeurIPS 2017]
[Qi et al., PointNet: Deep Learning on Point Sets for 3D Classification and Segmentation. CVPR 2017]

Structure

Data invariances – example of image classification

A function $f : \mathcal{X}(\Omega) \to \mathcal{Y}$ is $\mathfrak{G}$-*invariant* if $f(\rho(\mathfrak{g})x) = f(x)$ for all $\mathfrak{g} \in \mathfrak{G}$ and $x \in \mathcal{X}(\Omega)$, i.e., its output is unaffected by the group action on the input.



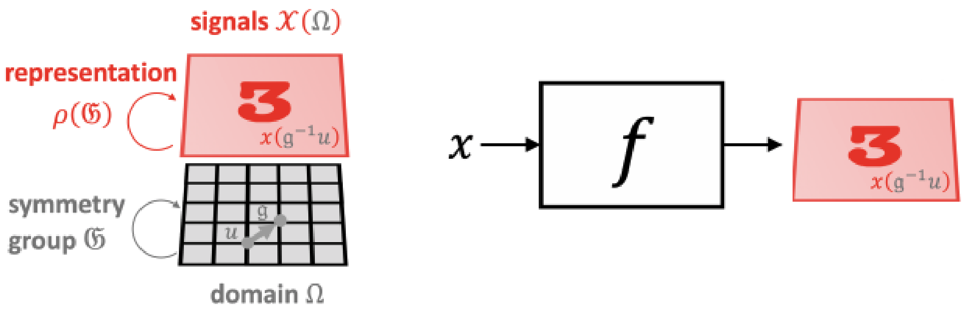
sunset

[Bronstein et al., Geometric Deep Learning: Grids, Groups, Graphs, Geodisics, and Gauges. arXiv 2021]

Structure

Data equivariances – example of image segmentation

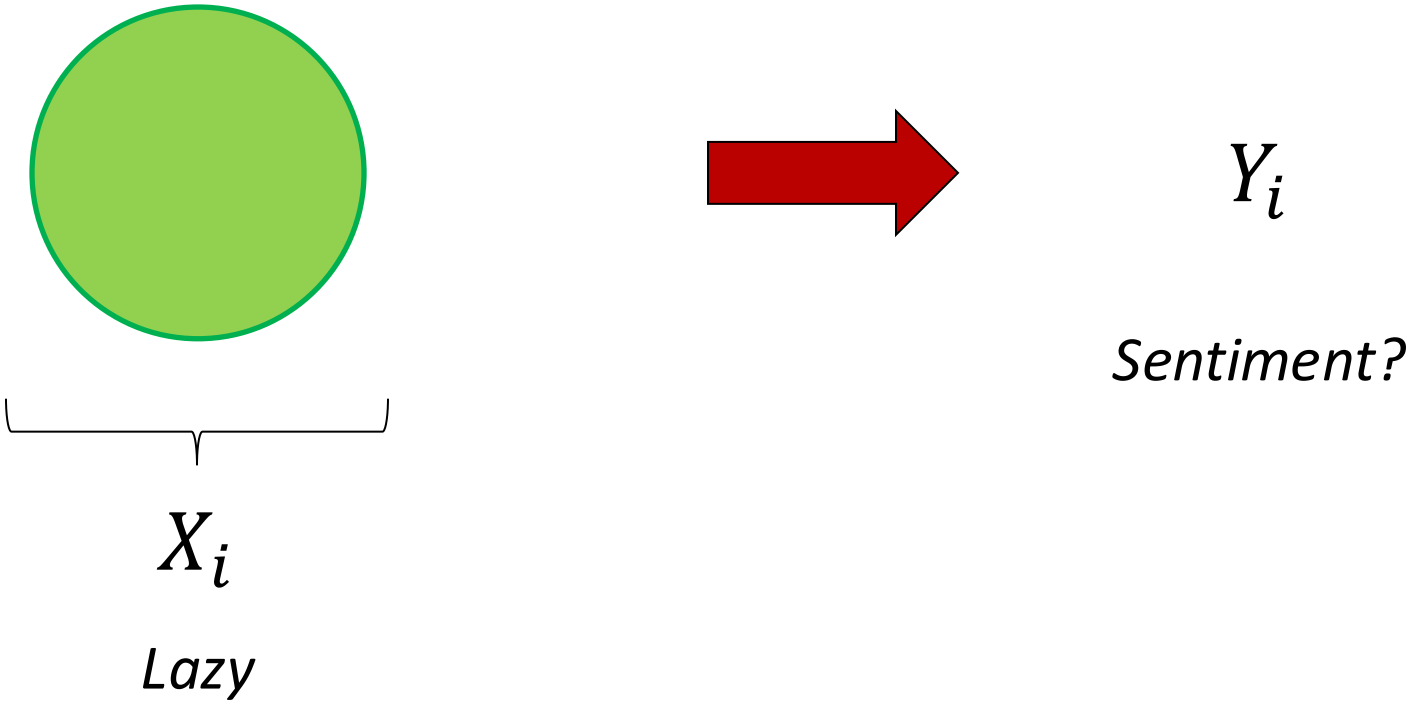
A function $f : \mathcal{X}(\Omega) \to \mathcal{X}(\Omega)$ is $\mathfrak{G}$-*equivariant* if $f(\rho(\mathfrak{g})x) = \rho(\mathfrak{g})f(x)$ for all $\mathfrak{g} \in \mathfrak{G}$, i.e., group action on the input affects the output in the same way.



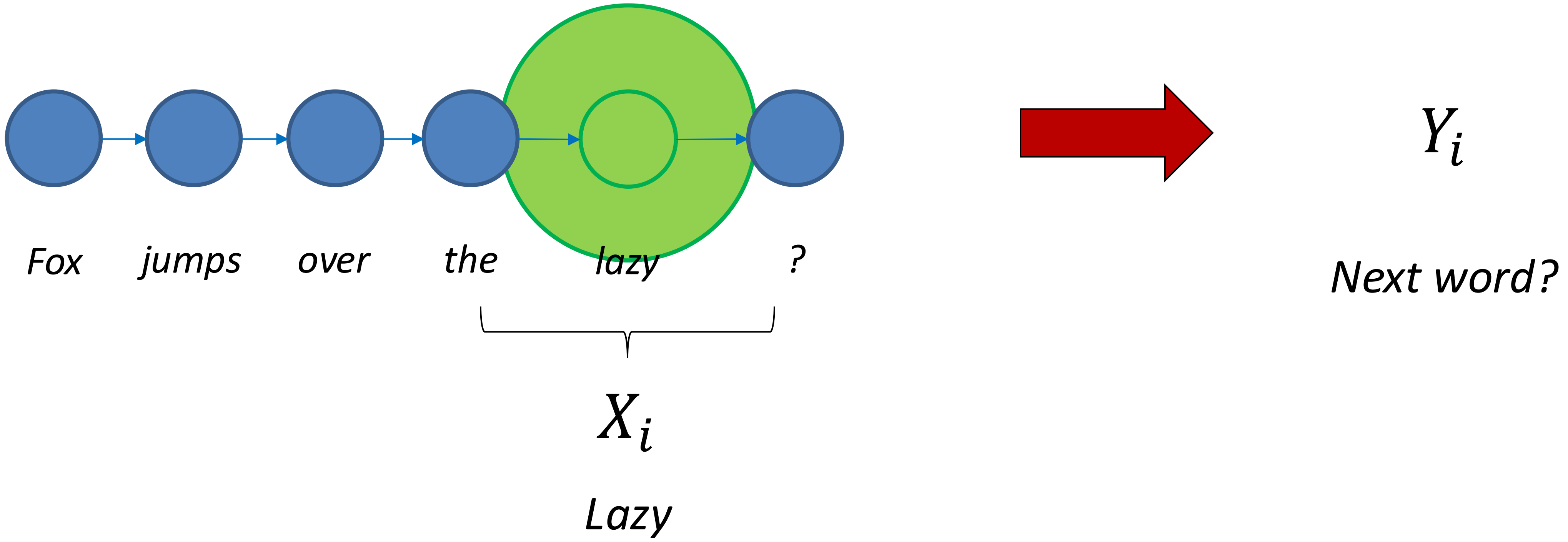
[Bronstein et al., Geometric Deep Learning: Grids, Groups, Graphs, Geodisics, and Gauges. arXiv 2021]

multisensory intelligence

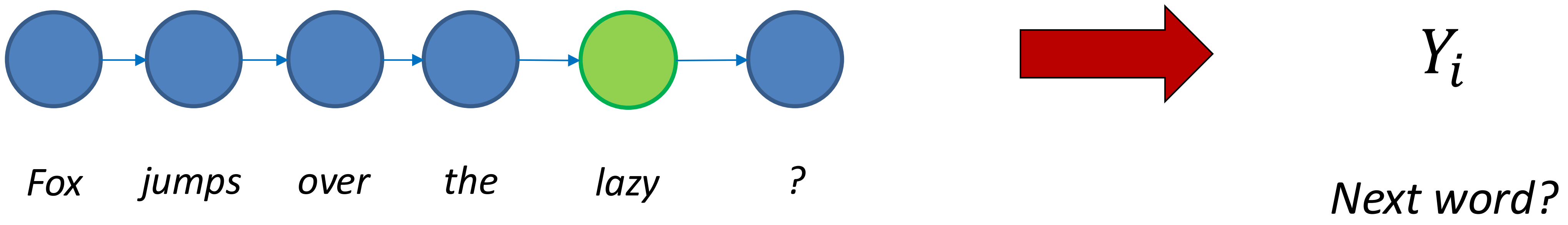
Elements



Sequences

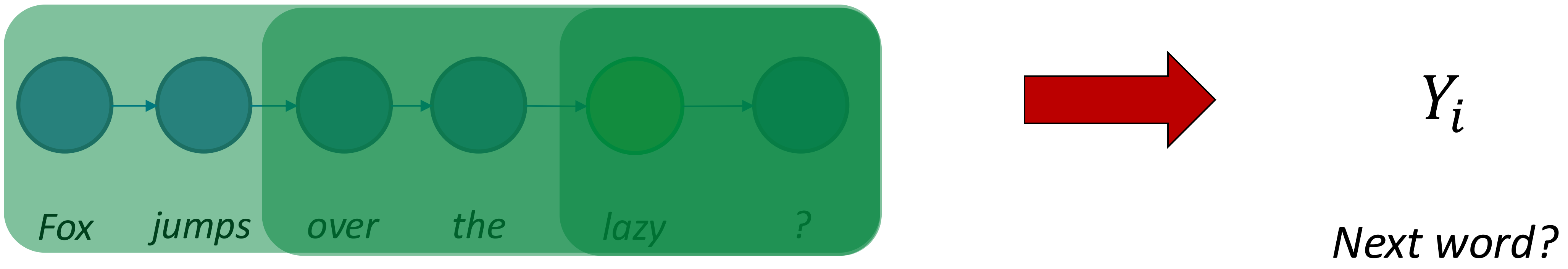


Sequences



How do we aggregate information?

Sequences

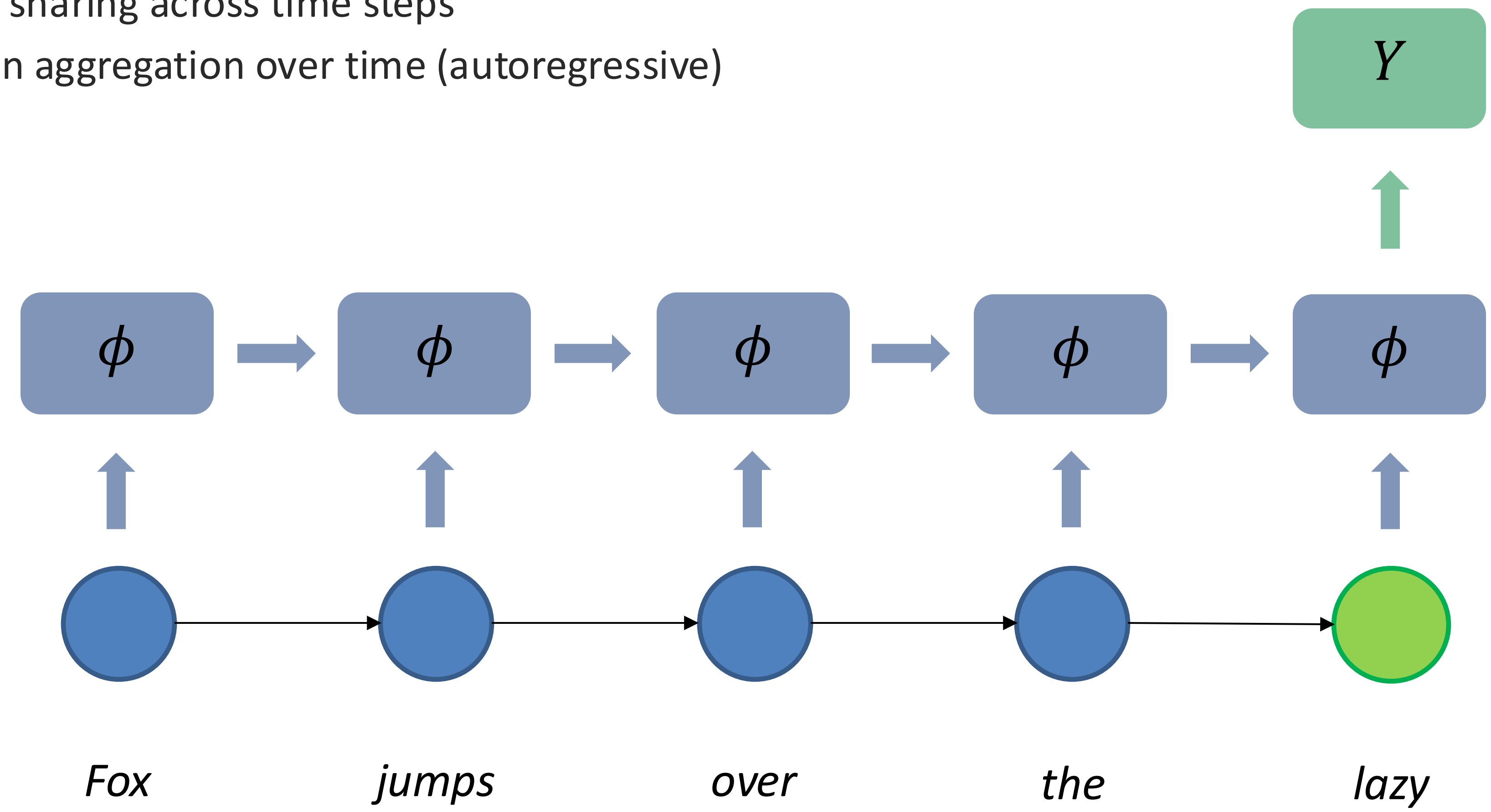


How do we aggregate information?

Sequence Classification

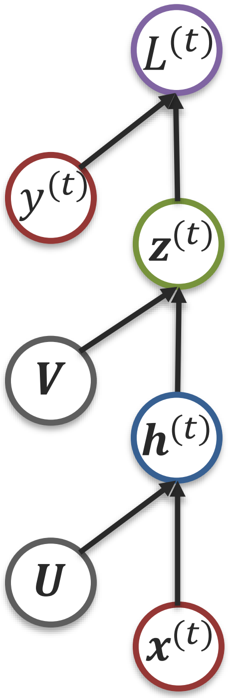
Models for sequential data must be invariant to time, but equivariant to word order.

1. Parameter sharing across time steps
2. Information aggregation over time (autoregressive)



Sequence Classification

Feedforward Neural Network



$$L^{(t)} = -\log P(Y = y^{(t)} | \boldsymbol{z}^{(t)})$$

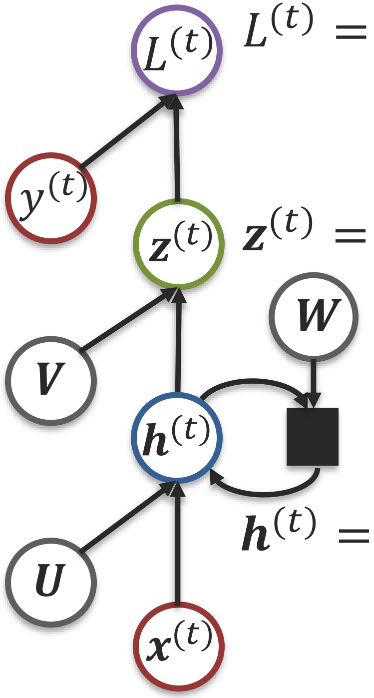
$$\boldsymbol{z}^{(t)} = \boldsymbol{V}\boldsymbol{h}^{(t)}$$

$$\boldsymbol{h}^{(t)} = \tanh(\boldsymbol{U}\boldsymbol{x}^{(t)})$$

multisensory intelligence

Sequence Classification

$$L = \sum_t L^{(t)}$$



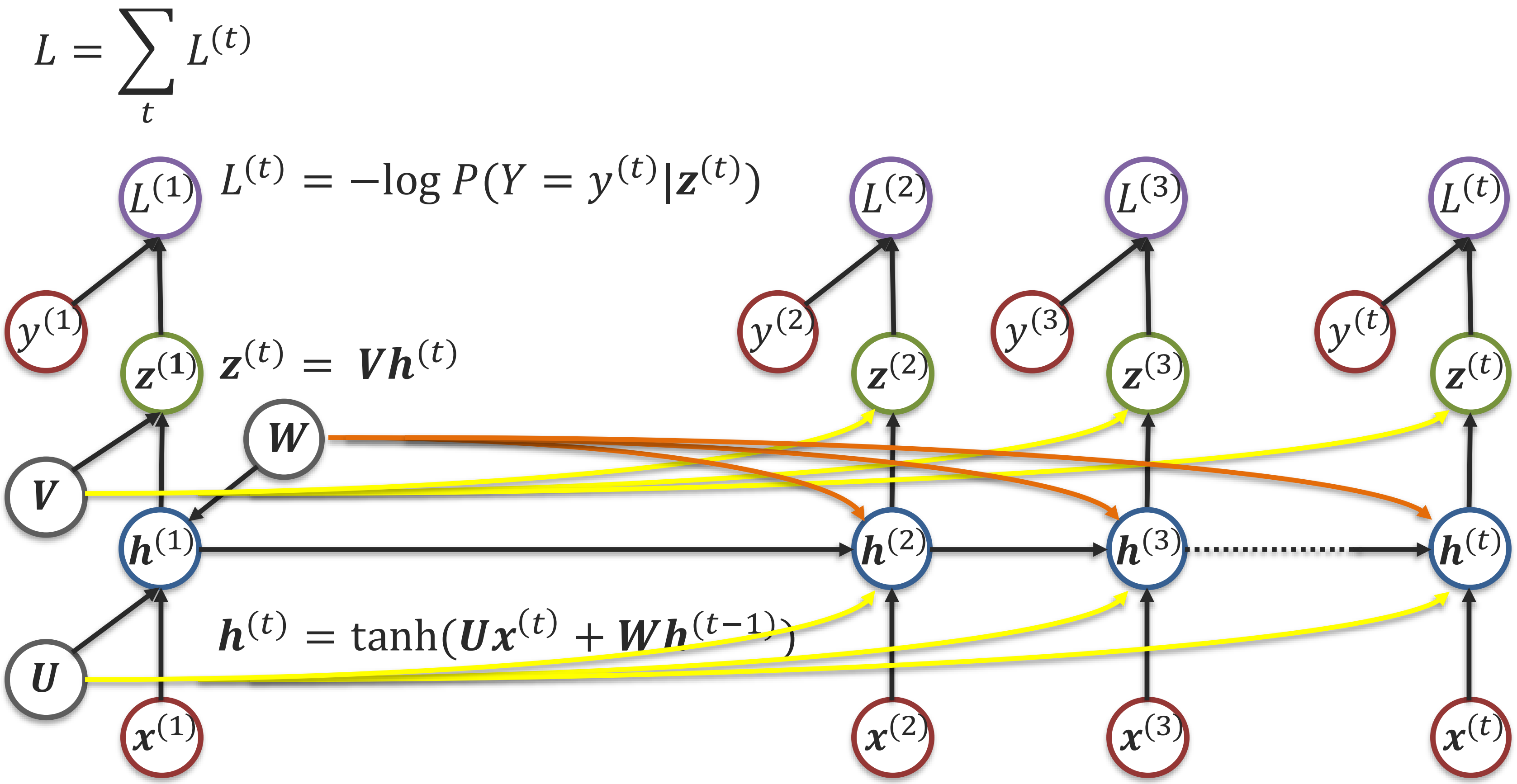
$$L^{(t)} = -\log P(Y = y^{(t)} | \boldsymbol{z}^{(t)})$$

$$\boldsymbol{z}^{(t)} = \boldsymbol{V}\boldsymbol{h}^{(t)}$$

$$\boldsymbol{h}^{(t)} = \tanh(\boldsymbol{U}\boldsymbol{x}^{(t)} + \boldsymbol{W}\boldsymbol{h}^{(t-1)})$$

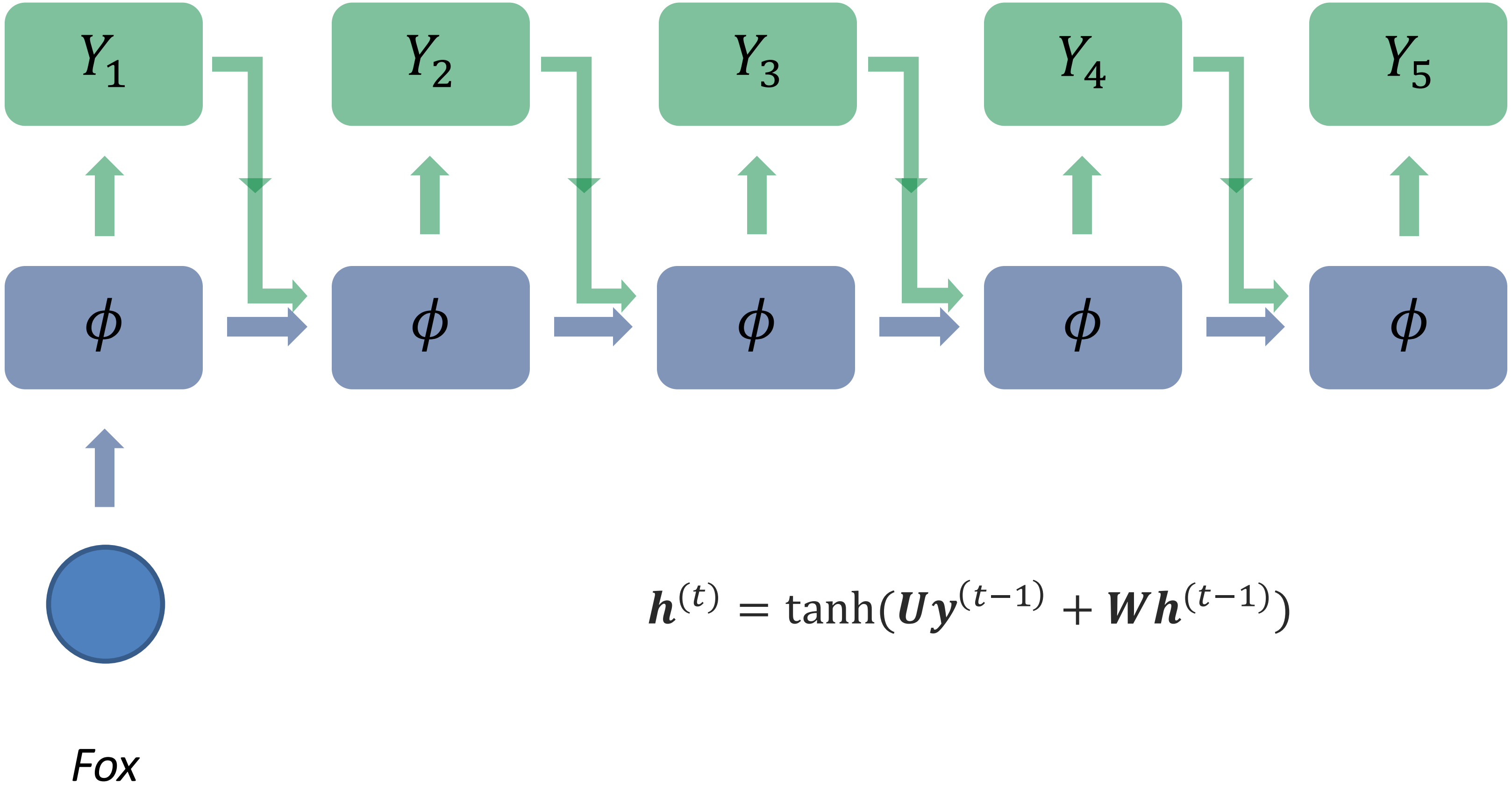
[original slide co-developed with Louis-Philippe Morency for CMU course 11-777]

Sequence Classification



Same model parameters are used for all time steps.

Sequence Generation

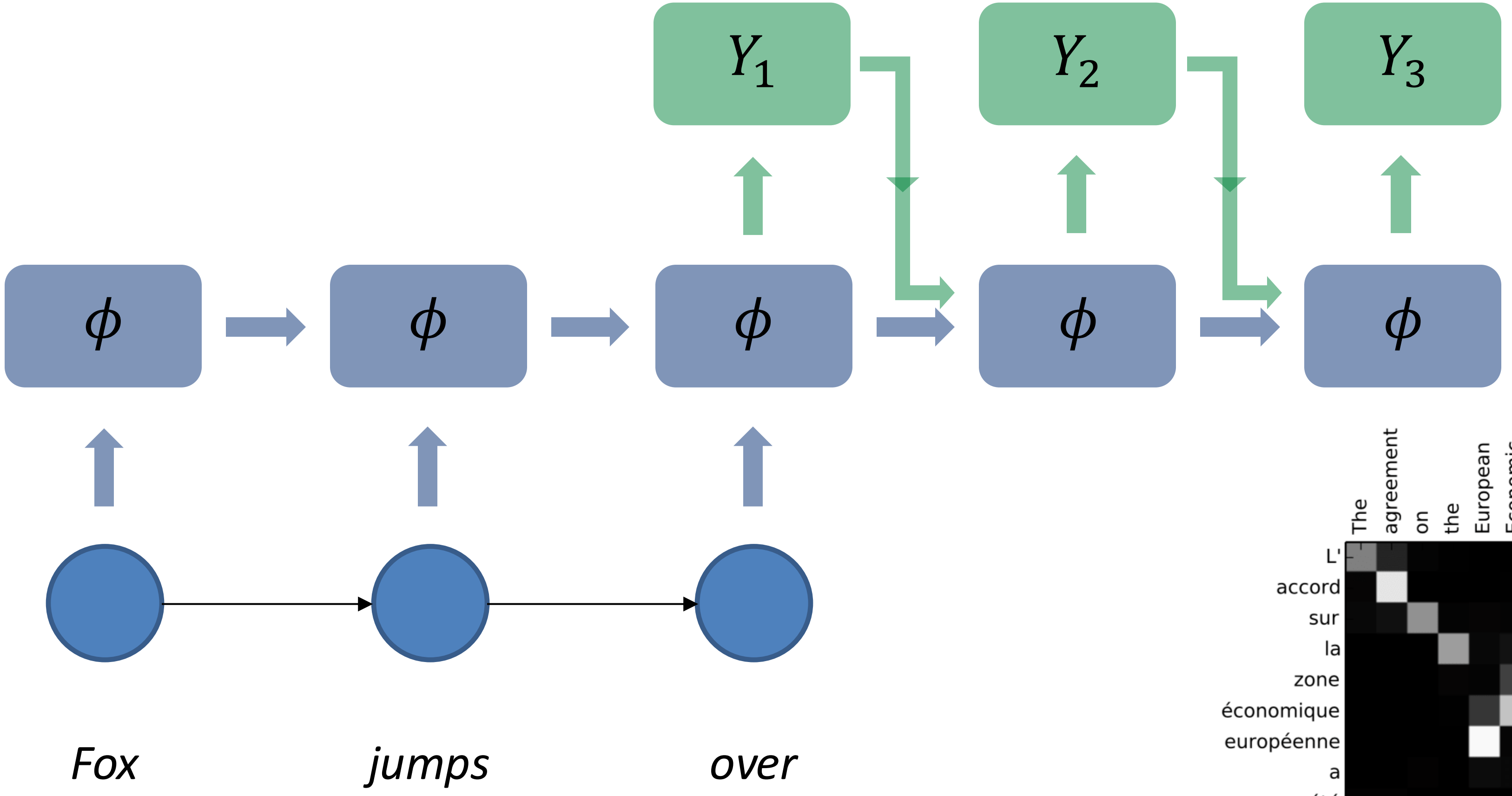


$$\boldsymbol{h}^{(t)} = \tanh(\boldsymbol{U}\boldsymbol{y}^{(t-1)} + \boldsymbol{W}\boldsymbol{h}^{(t-1)})$$

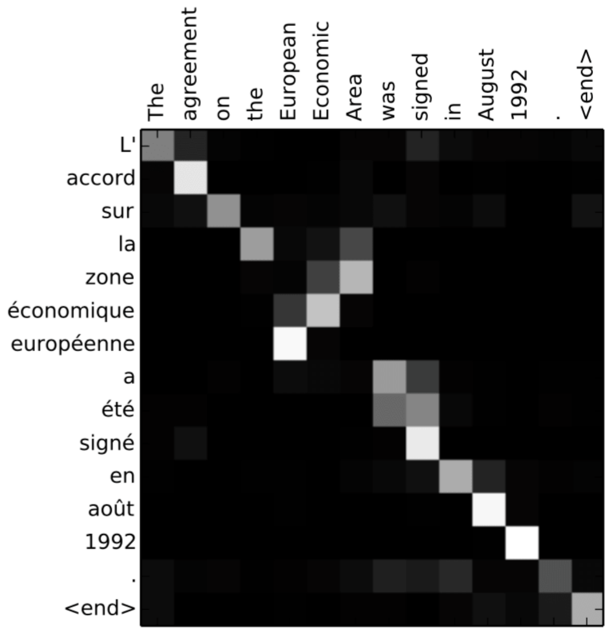
e.g, text or music generation

Modern versions: RNN -> LSTM -> TCN -> State space models

Sequence-to-Sequence Models



e.g, machine translation -> birth of attention-based models

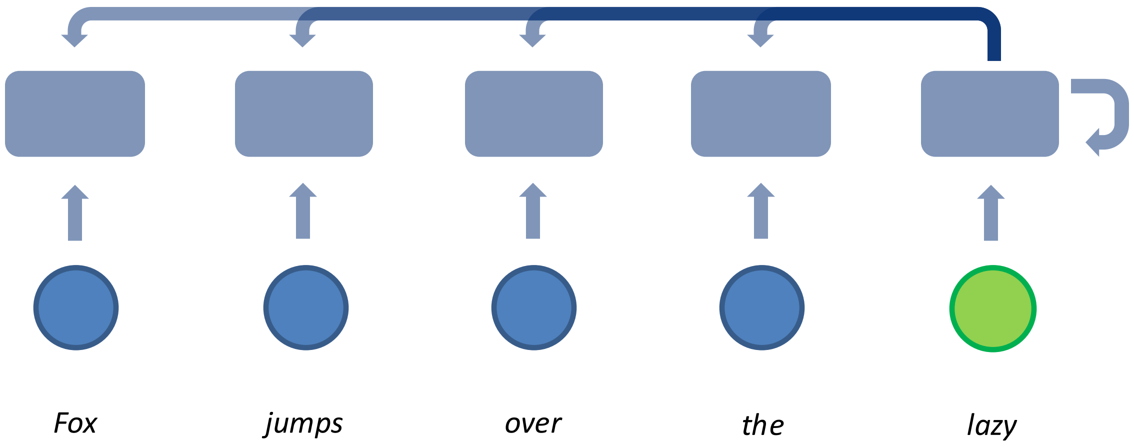


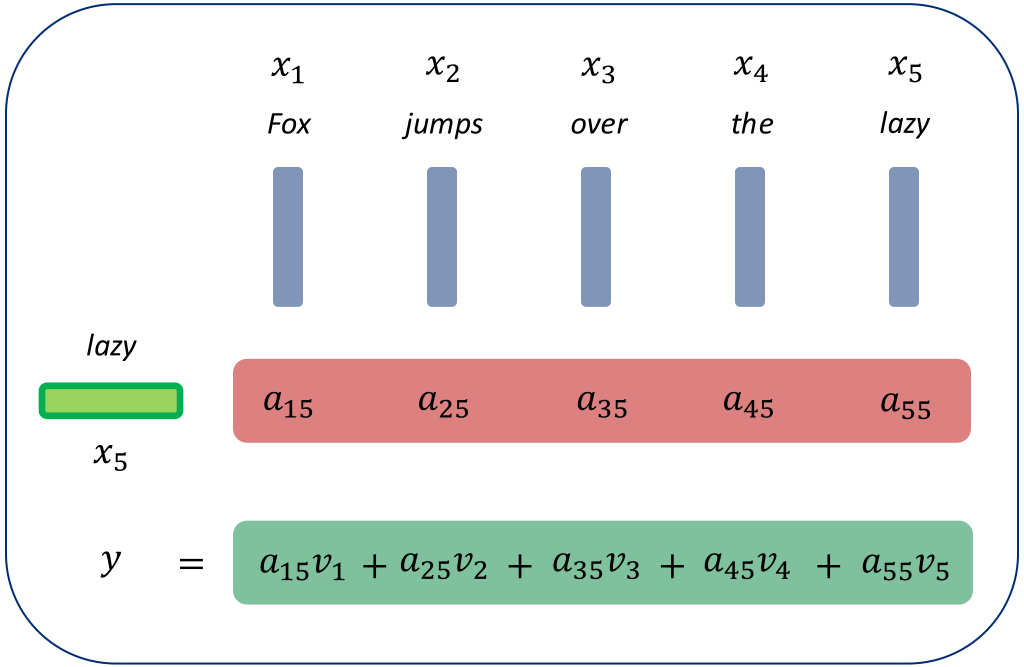
[Bahdanau et al., Neural Machine Translation by Jointly Learning to Align and Translate. ICLR 2015]

Modern Sequence Models

Birth of attention-based models

– Dynamic weights for different elements





[Vaswani et al., Attention is All You Need. NeurIPS 2017]

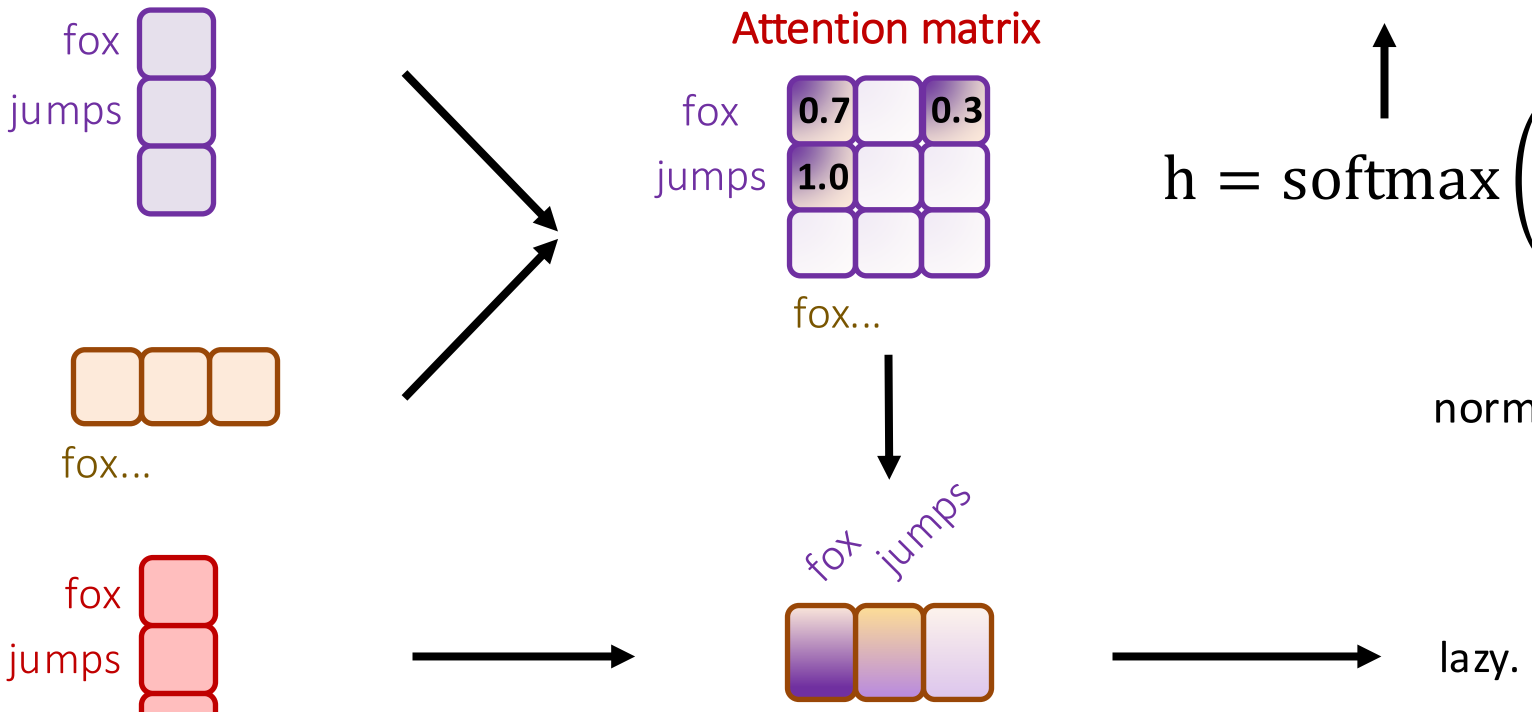
multisensory intelligence

35

Modern Sequence Models

Birth of attention-based models

– Dynamic weights for different elements



New **language** representation

3 x 3

(weighted) outer product

(row) normalize to 0-1

3 x d   d x 3   3 x d

$$h = \mathrm{softmax}\left(\frac{XW_q W_k^T X^T}{\sqrt{d}}\right) XW_v$$

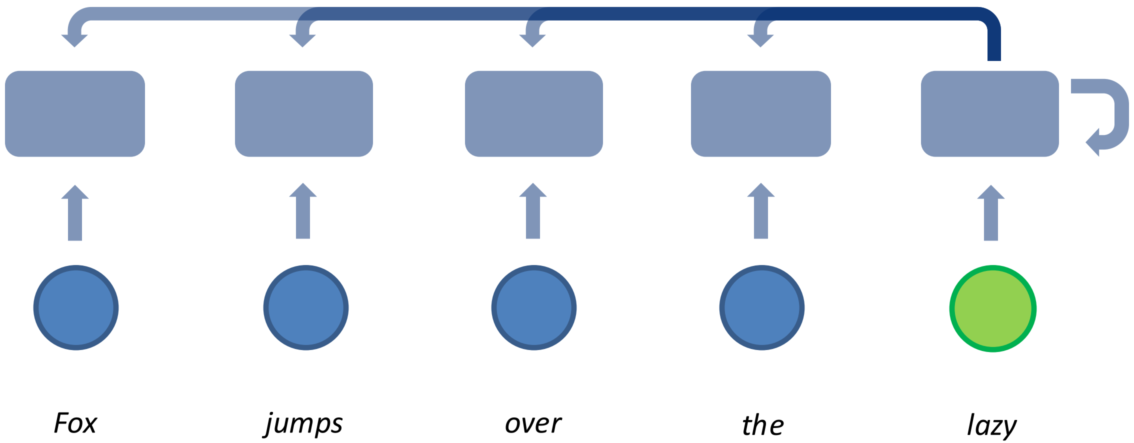
normalize wrt dimension *d*

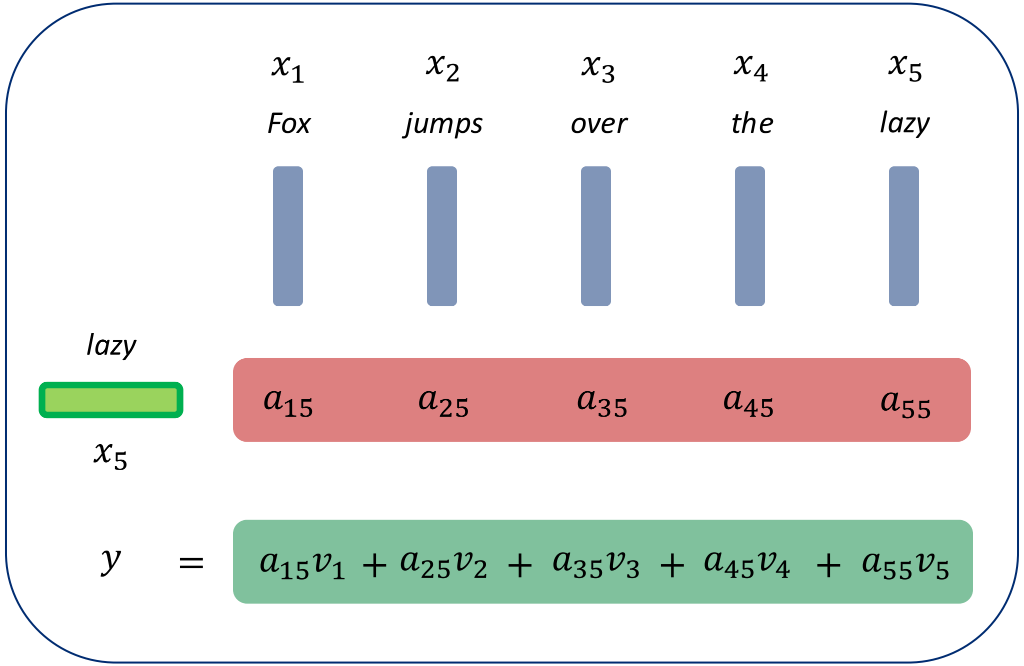
[Vaswani et al., Attention is All You Need. NeurIPS 2017]

Modern Sequence Models

Models for sequential data must be invariant to time, but equivariant to word order.

1. Parameter sharing across time steps
2. Information aggregation over time (in parallel)





$$h = \mathrm{softmax}\left(\frac{\overbrace{XW_q}^{T \times d}\overbrace{W_k^T X^T}^{d \times T}}{\sqrt{d}}\right)\overbrace{XW_v}^{T \times d}$$

[Vaswani et al., Attention is All You Need. NeurIPS 2017]

multisensory intelligence

Spatial Data

$\rightarrow$ Y_i

Is there a fox?

How do we aggregate information?

Spatial Data

Y_i

Is there a fox?

How do we aggregate information?

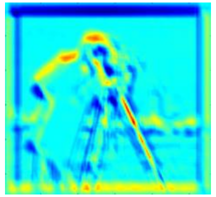
Convolutional Neural Networks

Models for spatial data need to be invariant to spatial translations.

Input

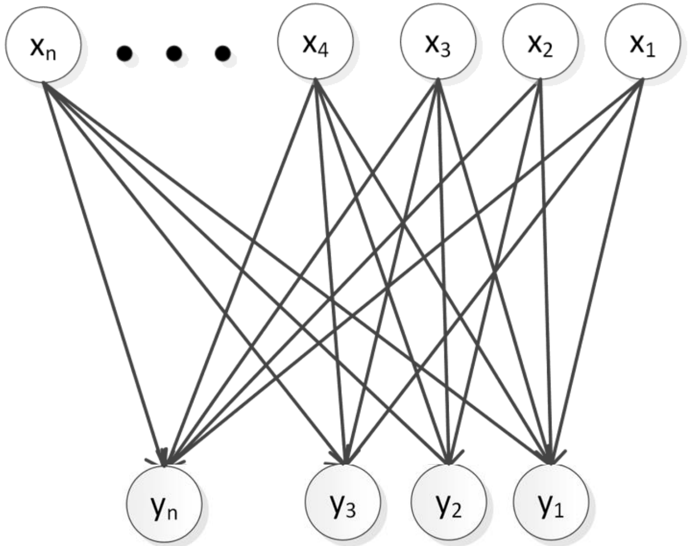
(image)

Output



(response map)

Input: all pixels



Output: responses

Not efficient!

200×200 image requires $40{,}000 \times n$ parameters

(where n is size of output)

And it may learn different outputs for different pixel positions

➡ Not spatial invariant

[original slide co-developed with Louis-Philippe Morency for CMU course 11-777]

multisensory intelligence

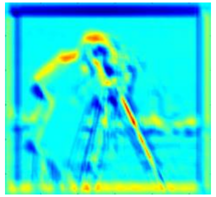
Convolutional Neural Networks

Modification 1: Only apply the filter to a small sliding window
-> for efficiency and locality

Input

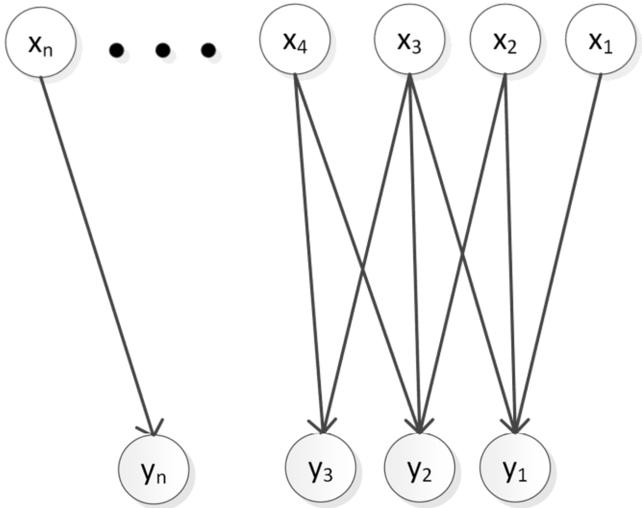
(image)

Output



(response map)

Input: all pixels



Output: responses

Example with 1D filter:

w_1	w_2	w_3

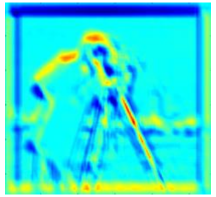
Convolutional Neural Networks

Modification 2: Same filter applied to all sliding windows
-> for spatial invariance

Input

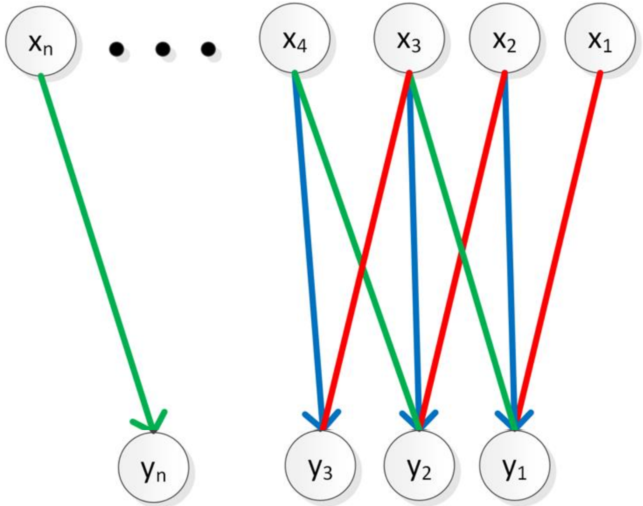
(image)

Output



(response map)

Input: all pixels



Output: responses

Example with 1D filter:

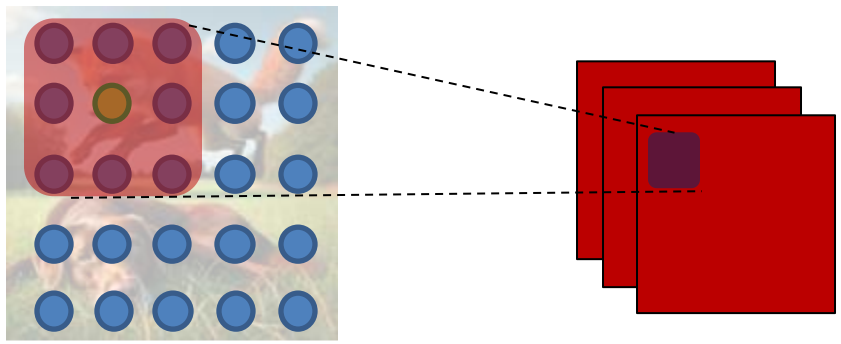
w_1	w_2	w_3

[original slide co-developed with Louis-Philippe Morency for CMU course 11-777]

Convolutional Neural Networks

Models for spatial data need to be invariant to spatial translation

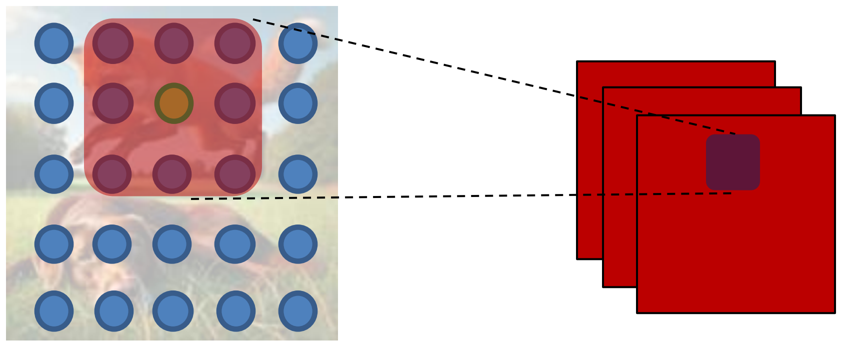
1. Parameter sharing across k x k convolutional filter



Convolutional Neural Networks

Models for spatial data need to be invariant to spatial translation

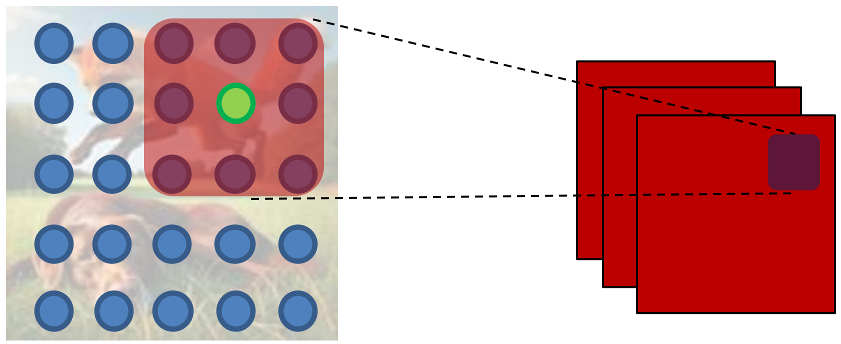
1. Parameter sharing across k x k convolutional filter



Convolutional Neural Networks

Models for spatial data need to be invariant to spatial translation

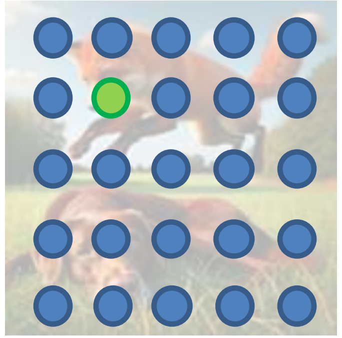
1. Parameter sharing across k x k convolutional filter

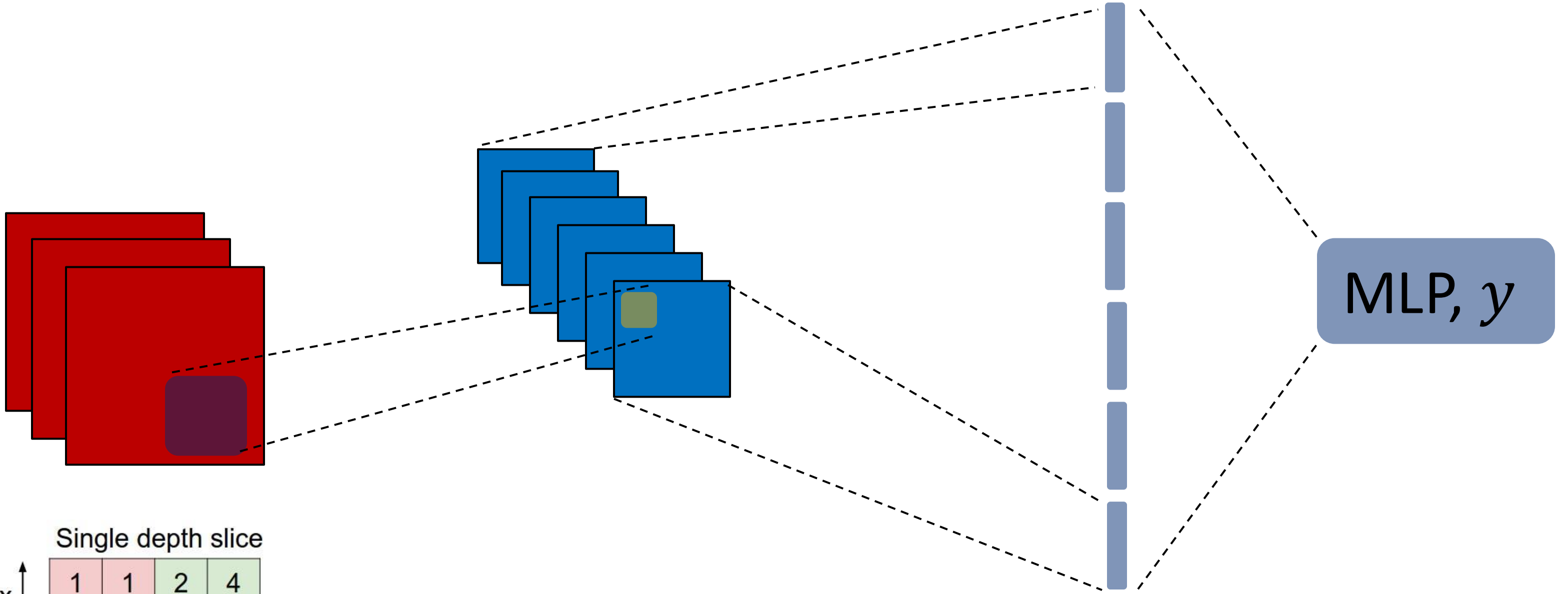


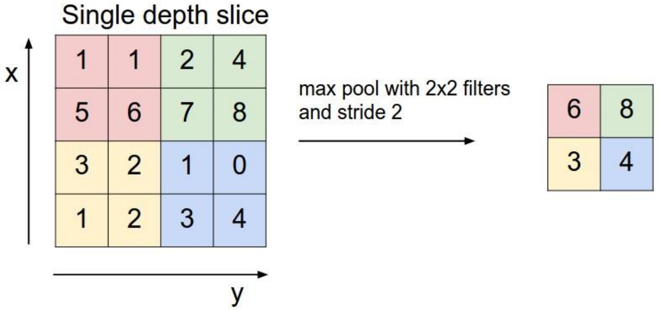
Convolutional Neural Networks

Models for spatial data need to be invariant to spatial translation

1. Parameter sharing across k x k convolutional filter
2. Information aggregation over k x k pooling region







multisensory intelligence

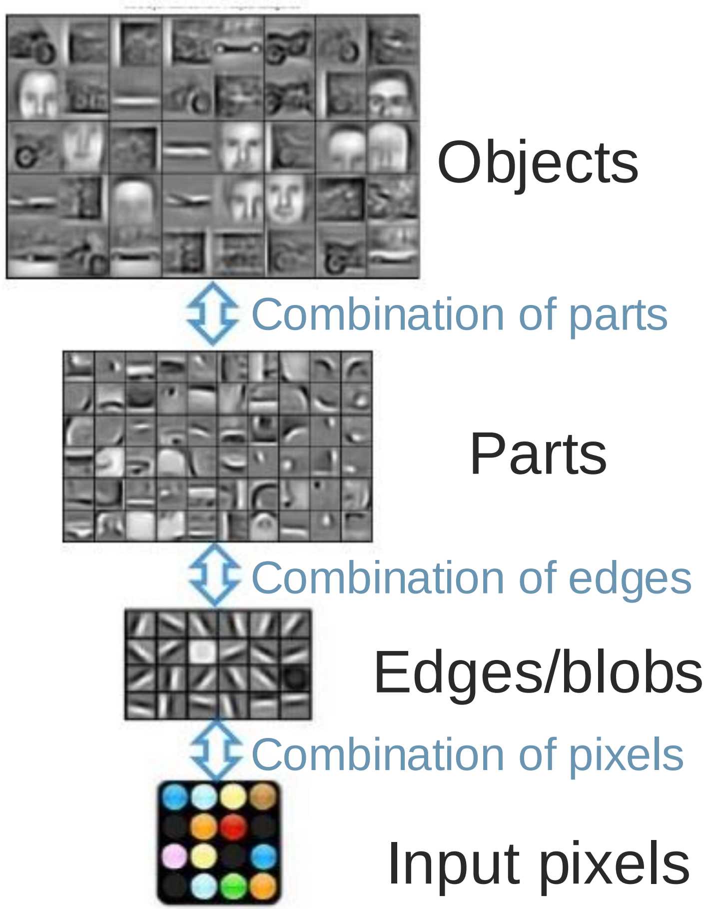
Convolutional Neural Networks

Multiple convolutional layers

- Allows the network to learn combinations of sub-parts, to increase complexity

Multiple pooling layers

- Allows the network to learn increasingly abstract & summarized information

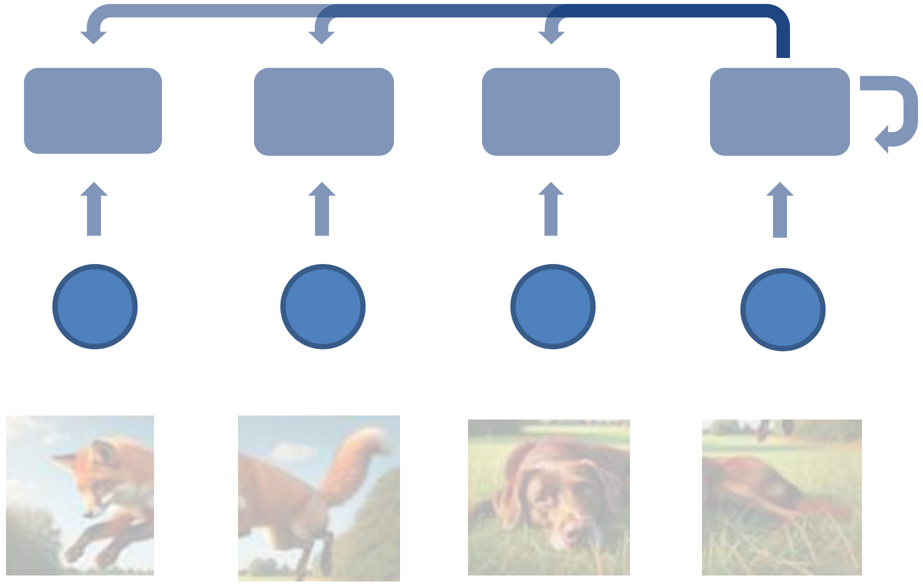


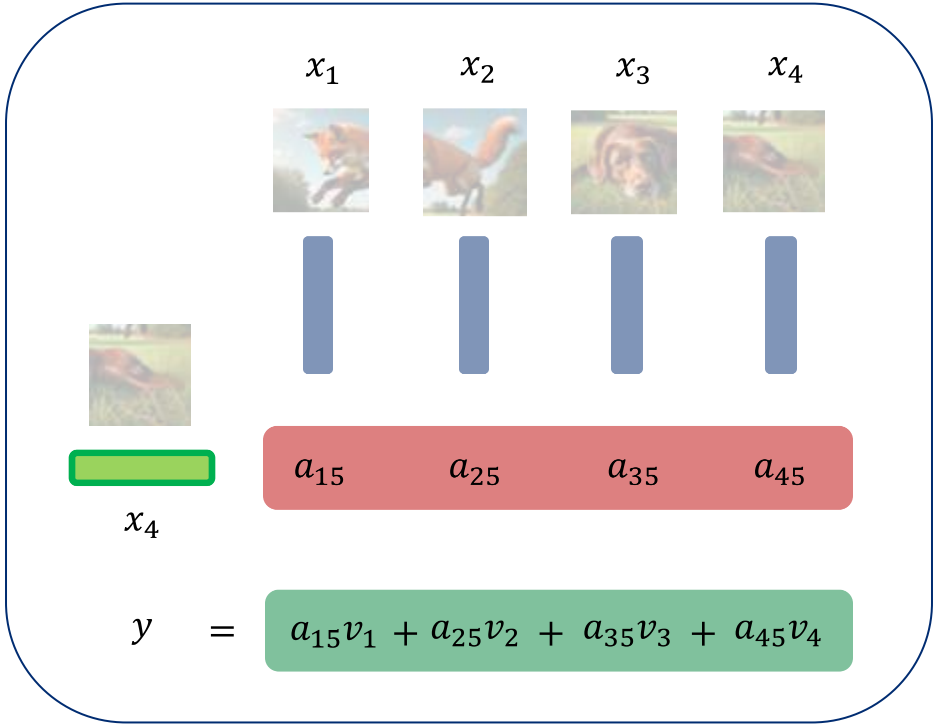
Vision Transformer

$$h = \mathrm{softmax}\left(\frac{\overset{T\,x\,d}{XW_q}\,\overset{d\,x\,T}{W_k^T X^T}}{\sqrt{d}}\right)\overset{T\,x\,d}{XW_v}$$

Models for spatial data need to be invariant to spatial translation

1. Parameter sharing across k x k self-attention region
2. Information aggregation over k x k patch region





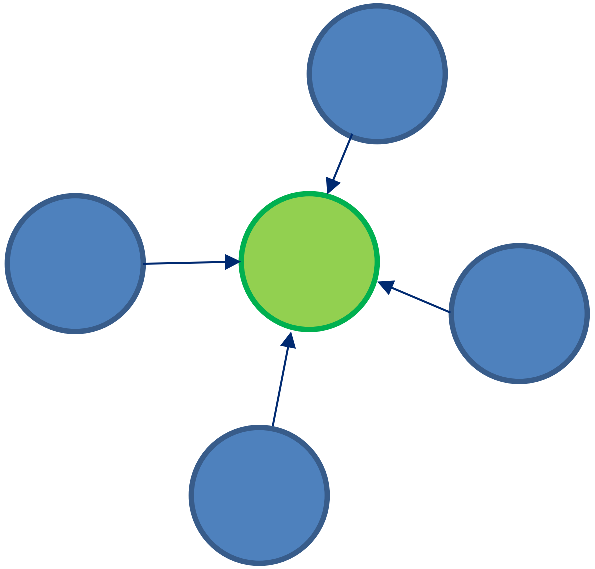
[Dosovitskiy et al. An image is worth 16x16 words: Transformers for image recognition at scale. ICLR 2021]

multisensory intelligence

Vision Transformer

[Dosovitskiy et al. An image is worth 16x16 words: Transformers for image recognition at scale. ICLR 2021]

Graphs



Y_i

What molecule is this?

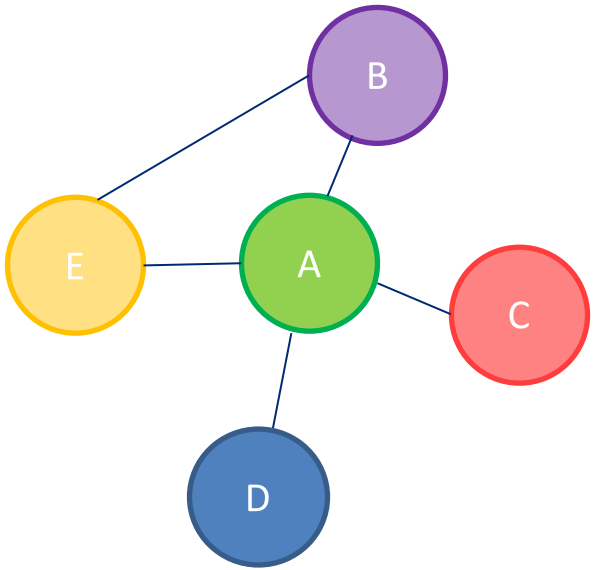
How do we aggregate information?

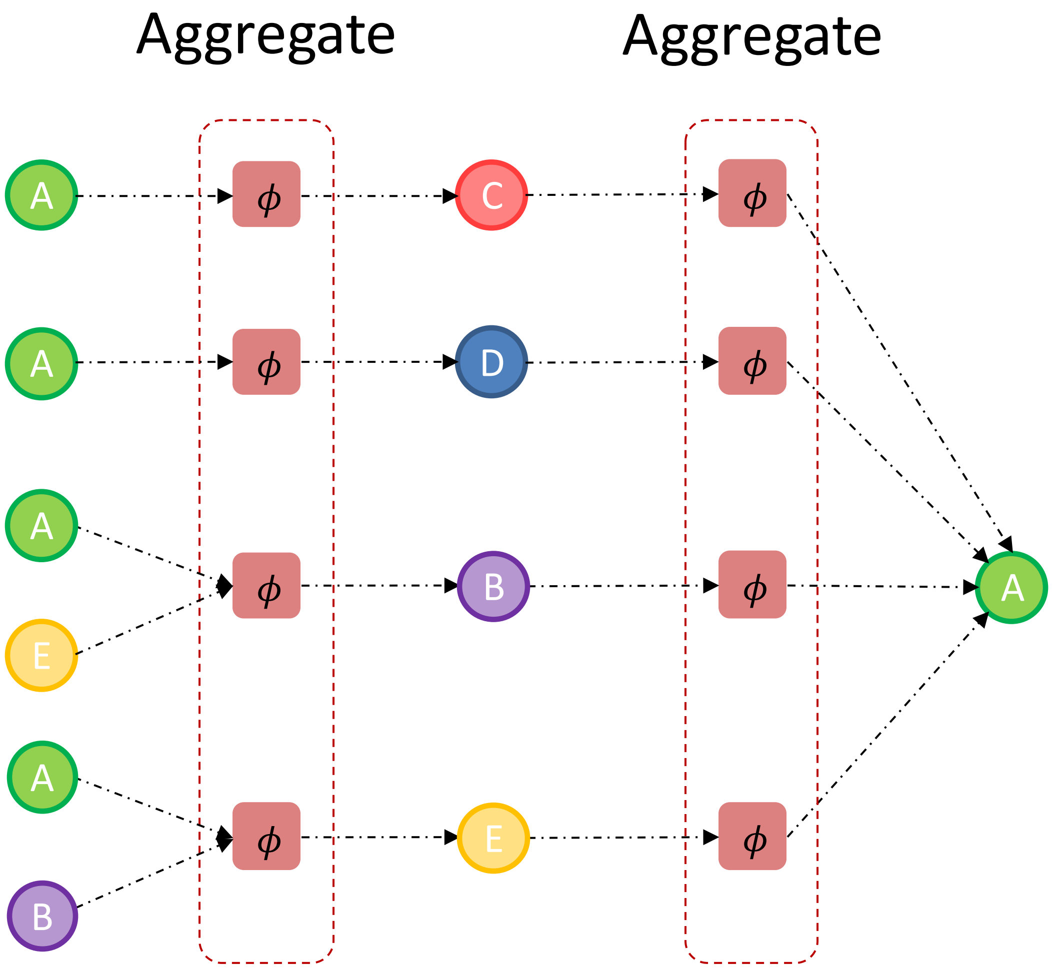
multisensory intelligence

Graph Neural Networks

Models for graph data:

1. Parameter sharing across nodes
2. Information aggregation over neighbors (edges)





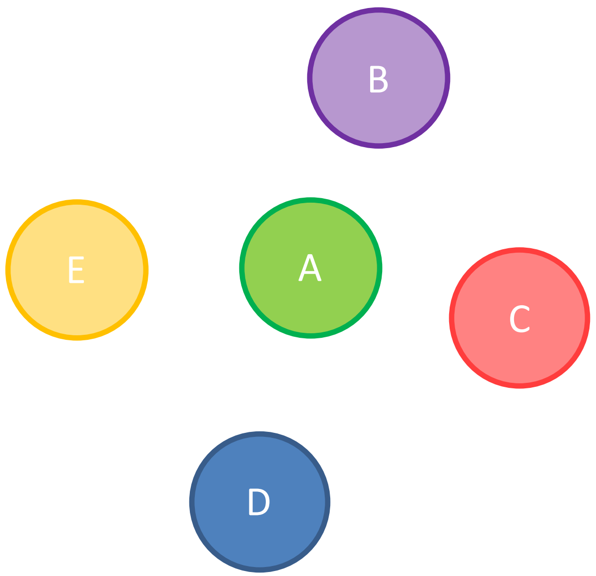
[Velickovic et al., Graph Attention Networks. ICLR 2018]
[Yun et al., Graph Transformer Networks. NeurIPS 2019]

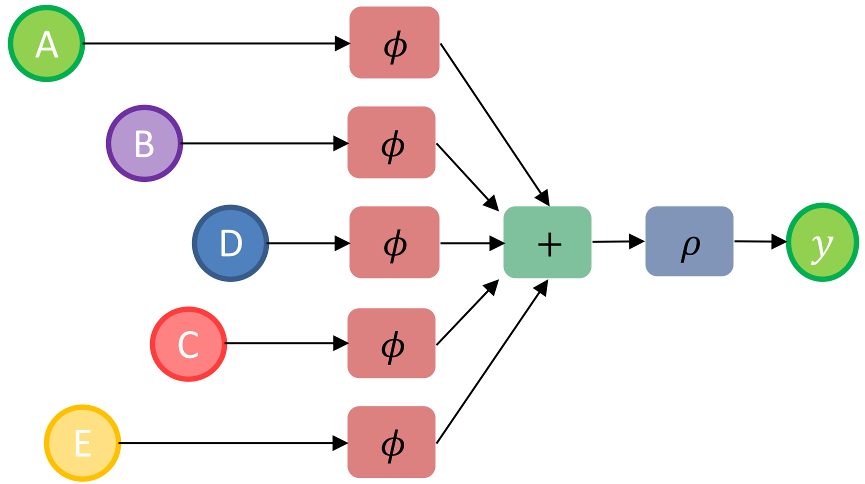
multisensory intelligence

Graph Recover Sets

Sets are graphs with only nodes, no edges

1. Parameter sharing across nodes -> set elements
2. Information aggregation over neighbors -> no neighbors





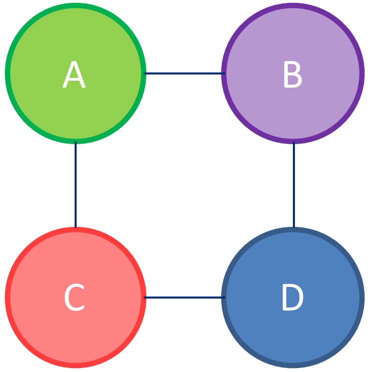
[Zaheer et al., DeepSets. NeurIPS 2017]
[Lee et al., Set Transformer. ICML 2019]

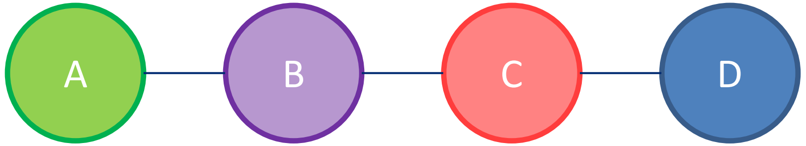
multisensory intelligence

Graph Recover Spatial and Temporal Data

Spatial data and sequential data

1. Parameter sharing across nodes
2. Information aggregation over neighbors

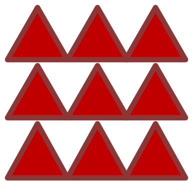


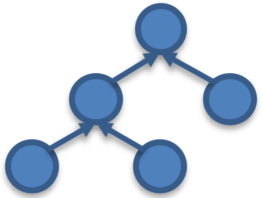


[Bronstein et al., Geometric Deep Learning: Grids, Groups, Graphs, Geodisics, and Gauges. arXiv 2021]

Summary: How To Model

1. Decide how much data to collect, and how much to label (costs and time)
2. Clean data: normalize/standardize, find noisy data, anomaly/outlier detection
3. Visualize data: plot, dimensionality reduction (PCA, t-sne), cluster analysis
4. Decide on evaluation metric (proxy + real, quantitative and qualitative)
5. Choose modeling paradigm - domain-specific vs general-purpose
6. Figure out base elements and their representation
7. Figure out data invariances & equivariances (+other parts of modality profile)
8. Iterate between data collection, model design, model training, hyperparameter tuning etc. until satisfied.





Lecture Summary

1. A unifying paradigm of model architectures
2. Temporal sequence models
3. Spatial convolution models
4. Models for sets and graphs

Assignments for This Coming Week

Reading assignment due tomorrow Wednesday (2/26).

For project:

- Project proposal due tonight (2/25). Email to me.
- Meet with me 2-3pm if need feedback about proposal ideas.

This Thursday (2/27): first reading discussion on **data and learning**.