Multimodal Machine Learning

Lecture 2.1: Basic Concepts

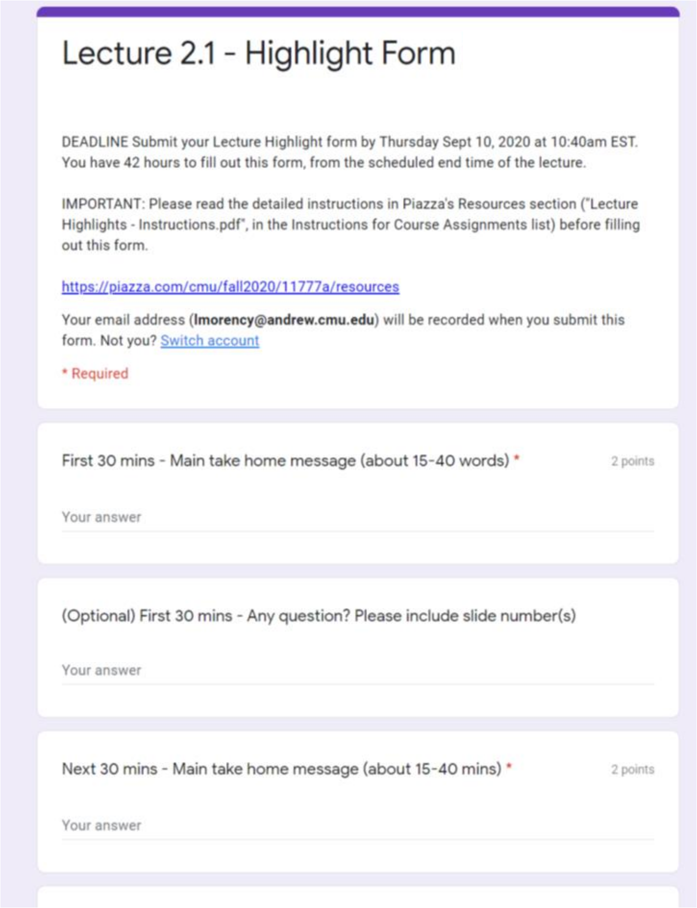
Louis-Philippe Morency

** Co-lecturer: Paul Liang. Original course co-developed with Tadas Baltrusaitis. Spring 2021 and 2022 editions taught by Yonatan Bisk.*

Administrative Stuff

Lecture Highlight Form

https://forms.gle/g6zovyaK2QwUvXW17



Lecture 2.1 - Highlight Form

DEADLINE Submit your Lecture Highlight form by Thursday Sept 10, 2020 at 10:40am EST. You have 42 hours to fill out this form, from the scheduled end time of the lecture.

IMPORTANT: Please read the detailed instructions in Piazza's Resources section ("Lecture Highlights - Instructions.pdf", in the Instructions for Course Assignments list) before filling out this form.

https://piazza.com/cmu/fall2020/11777a/resources

Your email address (**lmorency@andrew.cmu.edu**) will be recorded when you submit this form. Not you? Switch account

* Required

First 30 mins - Main take home message (about 15-40 words) * — 2 points

Your answer

(Optional) First 30 mins - Any question? Please include slide number(s)

Your answer

Next 30 mins - Main take home message (about 15-40 mins) * — 2 points

Your answer

Deadline: Tuesday 11:59pm ET

(for Thursday's lecture, the deadline is Thursday 11:59pm ET)

Use your Andrew CMU email

You will need to login using this address

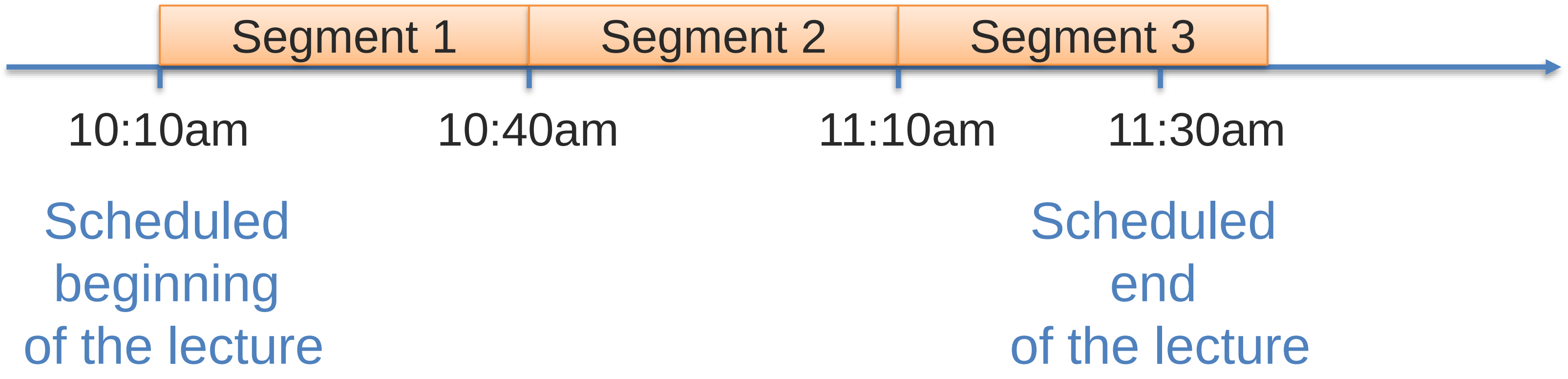
New form for each lecture

Posted on Piazza's Resources section

Ask questions about the lecture

- Will be answered either online or at the next lecture

Lecture Highlight Form - Segments



- Segment 1 starts at 10:10am, even if the lecture starts slightly later.
- Segment 3 ends whenever the lecture ends
- Slides happening around the segment borders (+/- 5min of 10:40am and 11:10am) can be included in either neighboring segment.

Lecture Highlight Form - Grading

For each segment

- Two sentences (10+ words each; complete English sentences) describing two main points described in this segment

For the whole lecture

- Your main two take-aways from the lecture
  - 10+ words each; complete English sentences
- Be as concrete as possible in your take-home messages
  - Avoid generic summaries like: “This is about multimodal”

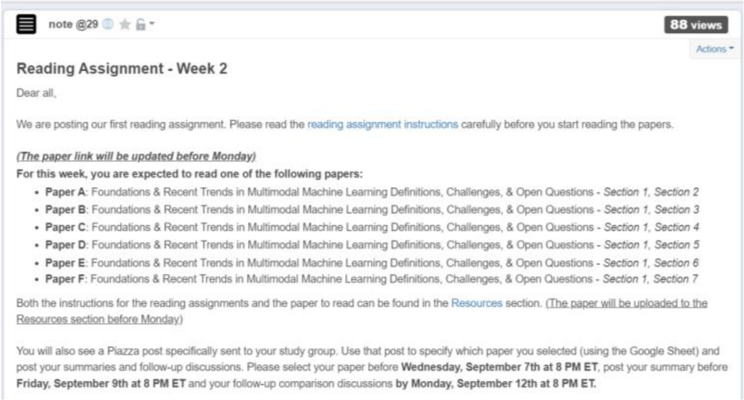
Each submission is worth 1 point

- Final grade is the sum of your top 16 submissions

Reading Assignments – Piazza Posts

For each reading assignment, 2 instruction posts will be created:

1


note @29

88 views

Actions

Reading Assignment - Week 2

Dear all,

We are posting our first reading assignment. Please read the reading assignment instructions carefully before you start reading the papers.

(The paper link will be updated before Monday)

For this week, you are expected to read one of the following papers:

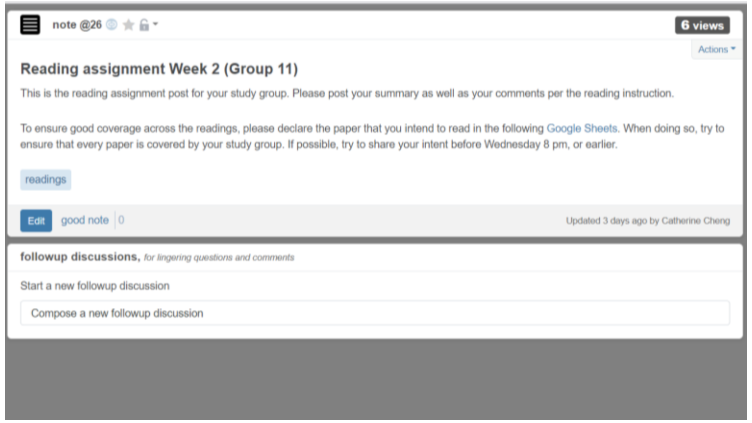
- **Paper A**: Foundations & Recent Trends in Multimodal Machine Learning Definitions, Challenges, & Open Questions - *Section 1, Section 2*
- **Paper B**: Foundations & Recent Trends in Multimodal Machine Learning Definitions, Challenges, & Open Questions - *Section 1, Section 3*
- **Paper C**: Foundations & Recent Trends in Multimodal Machine Learning Definitions, Challenges, & Open Questions - *Section 1, Section 4*
- **Paper D**: Foundations & Recent Trends in Multimodal Machine Learning Definitions, Challenges, & Open Questions - *Section 1, Section 5*
- **Paper E**: Foundations & Recent Trends in Multimodal Machine Learning Definitions, Challenges, & Open Questions - *Section 1, Section 6*
- **Paper F**: Foundations & Recent Trends in Multimodal Machine Learning Definitions, Challenges, & Open Questions - *Section 1, Section 7*

Both the instructions for the reading assignments and the paper to read can be found in the Resources section. (The paper will be uploaded to the Resources section before Monday)

You will also see a Piazza post specifically sent to your study group. Use that post to specify which paper you selected (using the Google Sheet) and post your summaries and follow-up discussions. Please select your paper before **Wednesday, September 7th at 8 PM ET**, post your summary before **Friday, September 9th at 8 PM ET** and your follow-up comparison discussions **by Monday, September 12th at 8 PM ET.**

- Sent to everyone
- Contains list of reading options

2


note @26

6 views

Actions

Reading assignment Week 2 (Group 11)

This is the reading assignment post for your study group. Please post your summary as well as your comments per the reading instruction.

To ensure good coverage across the readings, please declare the paper that you intend to read in the following Google Sheets. When doing so, try to ensure that every paper is covered by your study group. If possible, try to share your intent before Wednesday 8 pm, or earlier.

readings

Edit | good note | 0

Updated 3 days ago by Catherine Cheng

followup discussions, *for lingering questions and comments*

Start a new followup discussion

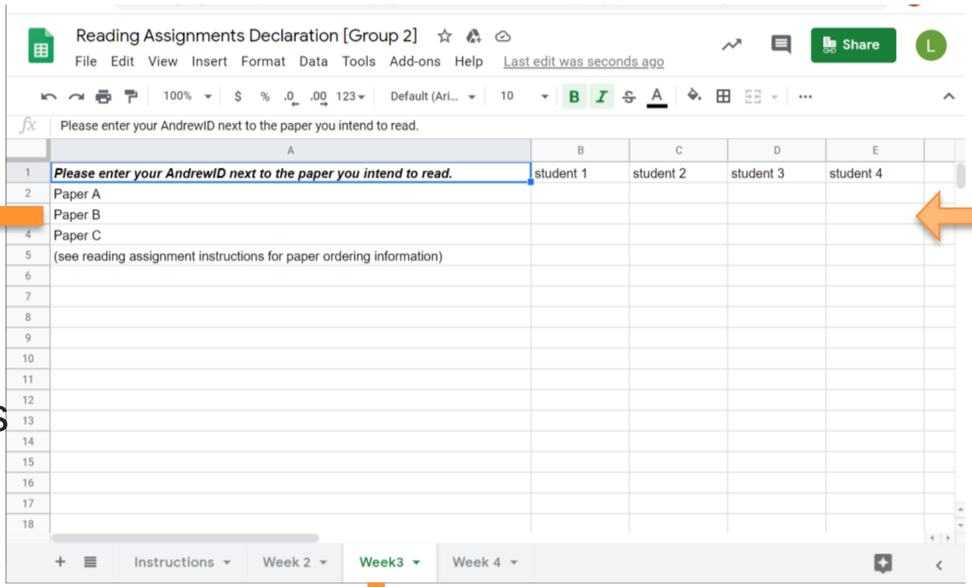
Compose a new followup discussion

- Sent separately to each study group
- Link to personalized signup sheet
- Post your summary as top-level
- Post your follow-up posts

Reading Assignments – Signup Sheet

Each study group has its own signup sheet:

Sign-up here for the paper option you would like to read and summarize



The details for the paper options are in the first Piazza post

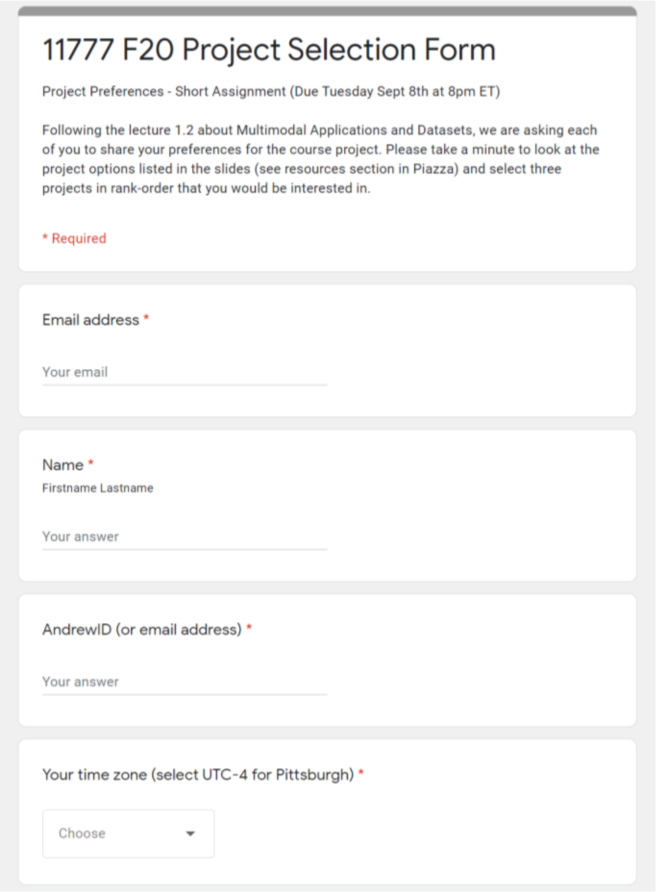
A different tab for each reading assignment

Reading Assignments – Weekly Schedule

Four main steps for the reading assignments

1. **Monday 8pm:** Official start of the assignment
2. **Wednesday 8pm:** Select your paper
3. **Friday 8pm:** Post your summary
4. **Monday 8pm:** End of the reading assignment

Team Matching – Project Preference Form



11777 F20 Project Selection Form

Project Preferences - Short Assignment (Due Tuesday Sept 8th at 8pm ET)

Following the lecture 1.2 about Multimodal Applications and Datasets, we are asking each of you to share your preferences for the course project. Please take a minute to look at the project options listed in the slides (see resources section in Piazza) and select three projects in rank-order that you would be interested in.

* Required

Email address *

Your email

Name *

Firstname Lastname

Your answer

AndrewID (or email address) *

Your answer

Your time zone (select UTC-4 for Pittsburgh) *

Choose

Deadline: Today at 8pm!!

- Every students should submit a form
- Students on the waitlist are also encouraged to submit a form
- A summary will be shared to help you find potential teammates

Team Matching – Thursday Event

Thursday around 11am ET

(later part of the lecture)

- Detailed instructions will be shared during lecture
- Event optional for students who already have a full team

Language Technologies Institute

Carnegie Mellon University

Multimodal Machine Learning

Lecture 2.1: Basic Concepts

Louis-Philippe Morency

** Co-lecturer: Paul Liang. Original course co-developed with Tadas Baltrusaitis. Spring 2021 and 2022 editions taught by Yonatan Bisk.*

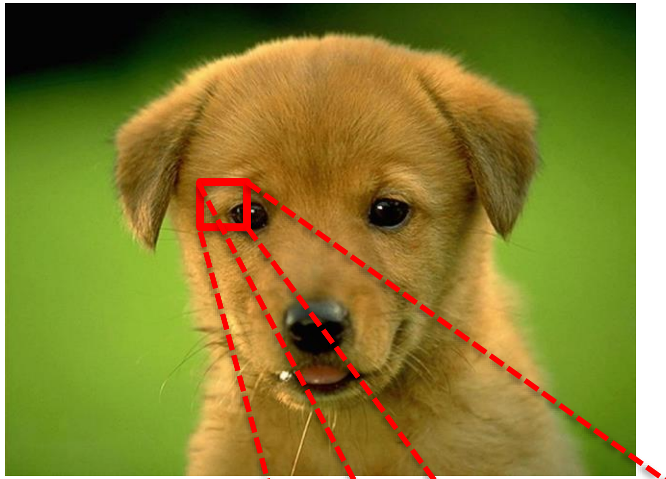
Lecture Objectives

- Unimodal basic representations
  - Visual, language and acoustic modalities
  - Sensors, tables, graphs, sets
- Data-driven machine learning
  - Representation learning
- Neural networks
  - Score and loss functions
  - Parameter optimization
    - Backpropagation and gradient descent
- Optimization – practical guidelines
  - Adaptive learning rate
  - Bias, variance and regularization

Unimodal Basic Representations

Unimodal Representation – Visual Modality

Color image



Each pixel is represented in $\mathcal{R}^d$, d is the number of colors (d=3 for RGB)

88	82	84	88	85	83	80	93	102
88	80	78	80	80	78	73	94	100
85	79	80	78	77	74	65	91	99
38	35	40	35	39	74	77	70	65
20	25	23	28	37	69	64	60	57
22	26	22	28	40	65	64	59	34
24	28	24	30	37	60	58	56	66
21	22	23	27	38	60	67	65	67
23	22	22	25	38	59	64	67	66

Input observation x_i

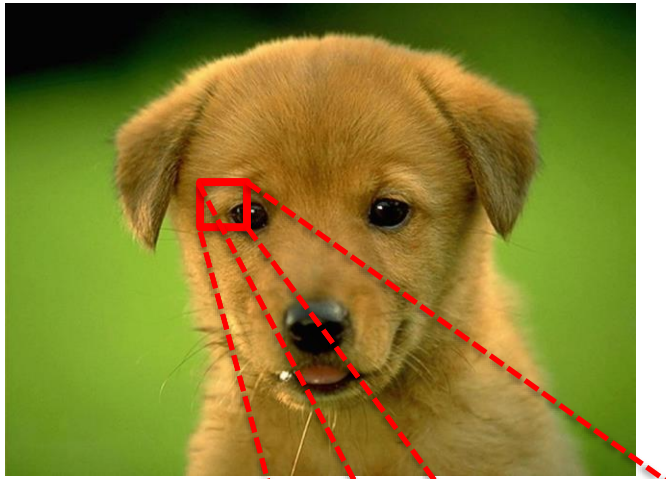
88
88
85
38
20
22
24
21
23
82
80
79
35
25
26
28
22
22
84
78
80
⋮

Binary classification problem

Dog ?

label $y_i \in \mathcal{Y} = \{0,1\}$

Unimodal Representation – Visual Modality



Each pixel is represented in $\mathcal{R}^d$, d is the number of colors (d=3 for RGB)

88	82	84	88	85	83	80	93	102
88	80	78	80	80	78	73	94	100
85	79	80	78	77	74	65	91	99
38	35	40	35	39	74	77	70	65
20	25	23	28	37	69	64	60	57
22	26	22	28	40	65	64	59	34
24	28	24	30	37	60	58	56	66
21	22	23	27	38	60	67	65	67
23	22	22	25	38	59	64	67	66

Input observation x_i

88
88
85
38
20
22
24
21
23
82
80
79
35
25
26
28
22
22
84
78
80
⋮

Multi-class classification problem

Duck

-or-

Cat

-or-

Dog

-or-

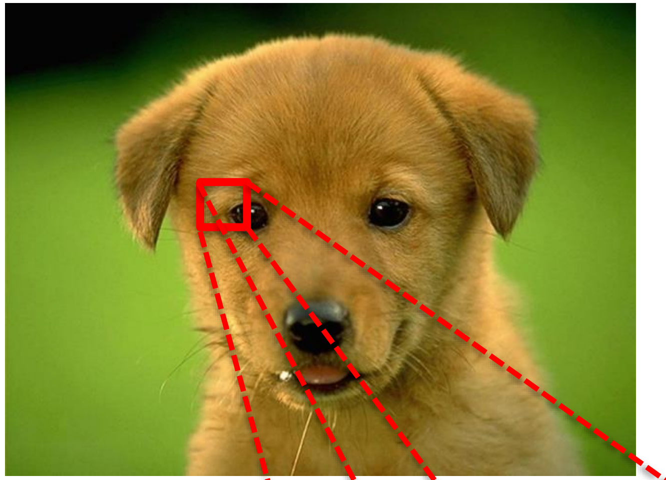
Pig

-or-

Bird ?

label $y_i \in \mathcal{Y} = \{0,1,2,3, \ldots\}$

Unimodal Representation – Visual Modality



88	82	84	88	85	83	80	93	102
88	80	78	80	80	78	73	94	100
85	79	80	78	77	74	65	91	99
38	35	40	35	39	74	77	70	65
20	25	23	28	37	69	64	60	57
22	26	22	28	40	65	64	59	34
24	28	24	30	37	60	58	56	66
21	22	23	27	38	60	67	65	67
23	22	22	25	38	59	64	67	66

Each pixel is represented in $\mathcal{R}^d$, d is the number of colors (d=3 for RGB)

Input observation $\boldsymbol{x_i}$

88
88
85
38
20
22
24
21
23
82
80
79
35
25
26
28
22
22
84
78
80
⋮

Multi-label (or multi-task) classification problem

Duck?

Cat ?

Dog ?

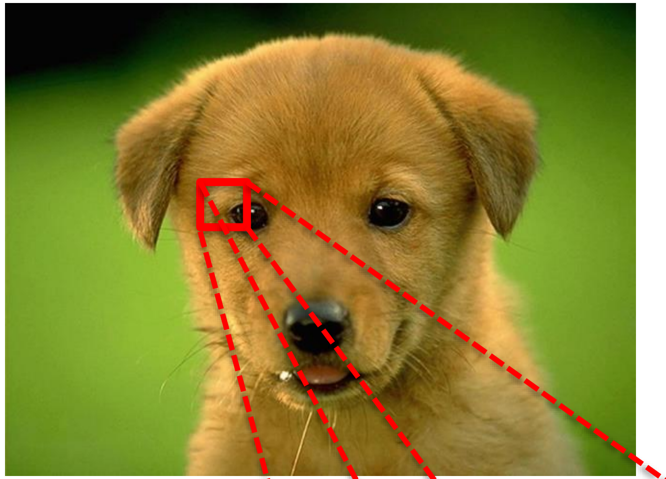
Pig ?

Bird ?

Puppy ?

label vector $\boldsymbol{y_i} \in \mathcal{Y}^m = \{0,1\}^m$

Unimodal Representation – Visual Modality



Each pixel is represented in $\mathbb{R}^d$, d is the number of colors (d=3 for RGB)

88	82	84	88	85	83	80	93	102
88	80	78	80	80	78	73	94	100
85	79	80	78	77	74	65	91	99
38	35	40	35	39	74	77	70	65
20	25	23	28	37	69	64	60	57
22	26	22	28	40	65	64	59	34
24	28	24	30	37	60	58	56	66
21	22	23	27	38	60	67	65	67
23	22	22	25	38	59	64	67	66

Input observation $\boldsymbol{x_i}$

88
88
85
38
20
22
24
21
23
82
80
79
35
25
26
28
22
22
84
78
80
⋮

Multi-label (or multi-task) regression problem

Age ?

Height ?

Weight ?

Distance ?

Happy ?

label vector $\boldsymbol{y_i} \in \mathcal{Y}^m = \mathbb{R}^m$

Unimodal Representation – Language Modality

Written language

★★★★★ **Masterful!**

By Antony Witheyman - January 12, 2006

Ideal for anyone with an interest in disguises who likes to see the subject tackled in a $\boxed{\text{humourous}}$ manner.

0 of 4 people found this review helpful

Spoken language

```
                MARTHA(CON'T)
Look around you. Look at all the
great things you've done and the
people you've helped.

                CLARK
But you've only put up the good
things they say about me.

                MARTHA
Clark, honey. If I were to use the
bad things they say I could cover
the barn, the house and the
outhouse.
```

Input observation x_i

$$\begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ \vdots \end{bmatrix}$$

"one-hot" vector

$|x_i|$ = number of words in dictionary

Word-level classification

Part-of-speech ?
(noun, verb,…)

Sentiment ?
(positive or negative)

Named entity ?
(names of person,…)

Unimodal Representation – Language Modality

Written language

★★★★★ **Masterful!**

By Antony Witheyman - January 12, 2006

Ideal for anyone with an interest in disguises who likes to see the subject tackled in a humourous manner.

0 of 4 people found this review helpful

Spoken language

```
                MARTHA(CON'T)
Look around you. Look at all the
great things you've done and the
people you've helped.

                CLARK
But you've only put up the good
things they say about me.

                MARTHA
Clark, honey. If I were to use the
bad things they say I could cover
the barn, the house and the
outhouse.
```

Input observation x_i

$$x_i = \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \\ 1 \\ 0 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 1 \\ 0 \\ 0 \\ 0 \\ \vdots \end{bmatrix}$$

"bag-of-word" vector

$|x_i|$ = number of words in dictionary

Document-level classification

Sentiment ?

(positive or negative)

Unimodal Representation – Language Modality

Written language

Masterful!

By Antony Witheyman - January 12, 2006

Ideal for anyone with an interest in disguises who likes to see the subject tackled in a humourous manner.

0 of 4 people found this review helpful

Spoken language

```
            MARTHA(CON'T)
Look around you. Look at all the
great things you've done and the
people you've helped.

            CLARK
But you've only put up the good
things they say about me.

            MARTHA
Clark, honey. If I were to use the
bad things they say I could cover
the barn, the house and the
outhouse.
```

Input observation $\boldsymbol{x_i}$

$$\begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \\ 1 \\ 0 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 1 \\ 0 \\ 0 \\ 0 \\ \vdots \end{bmatrix}$$

"bag-of-word" vector

$|\boldsymbol{x_i}|$ = number of words in dictionary

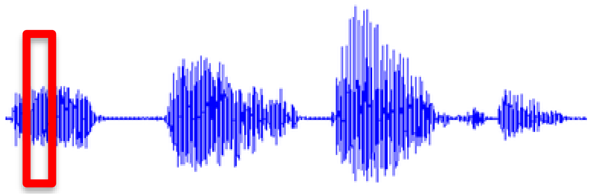
Utterance-level classification

Sentiment ?

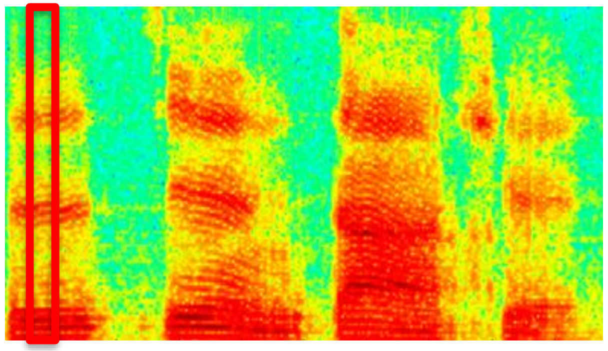
(positive or negative)

Unimodal Representation – Acoustic Modality

Digitalized acoustic signal



- Sampling rates: 8~96kHz
- Bit depth: 8, 16 or 24 bits
- Time window size: 20ms
  - Offset: 10ms



Spectogram

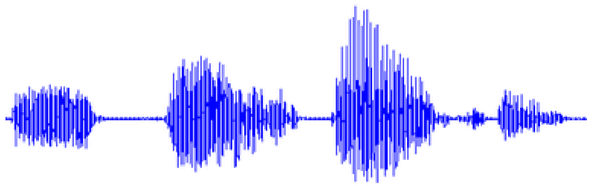
Input observation x_i

0.21
0.14
0.56
0.45
0.9
0.98
0.75
0.34
0.24
0.11
0.02

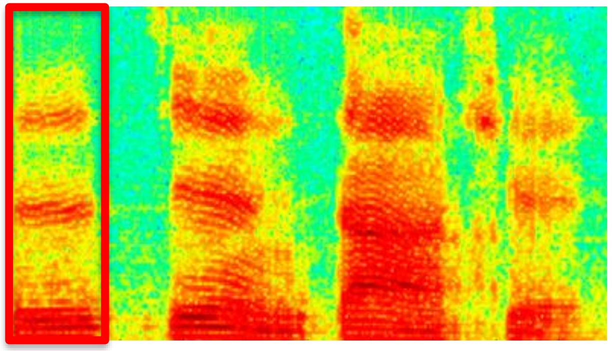
Spoken word ?

Unimodal Representation – Acoustic Modality

Digitalized acoustic signal



- Sampling rates: 8~96kHz
- Bit depth: 8, 16 or 24 bits
- Time window size: 20ms
  - Offset: 10ms



Spectogram

Input observation x_i

0.21
0.14
0.56
0.45
0.9
0.98
0.75
0.34
0.24
0.11
0.02
0.24
0.26
0.58
0.9
0.99
0.79
0.45
0.34
0.24

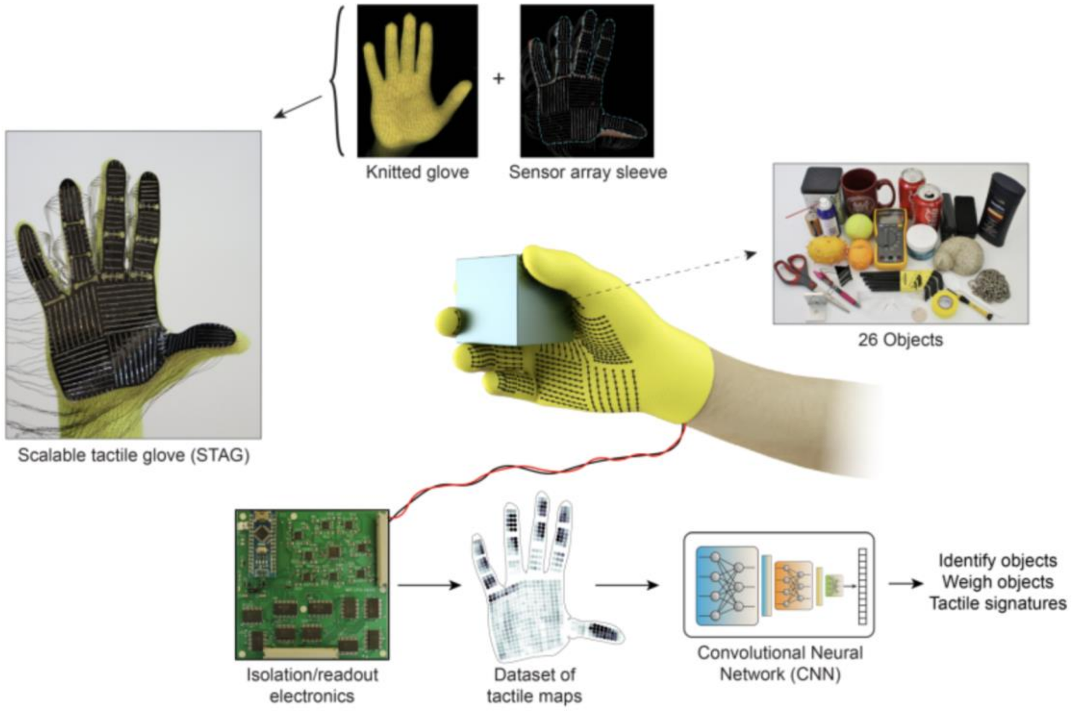
⋮

Emotion ?

Spoken word ?

Voice quality ?

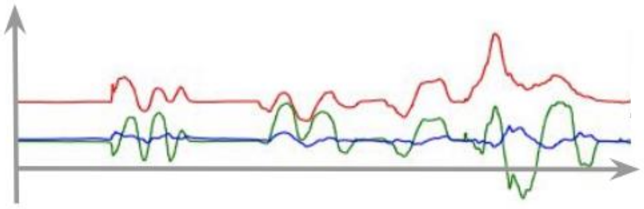
Unimodal Representation – Sensors



The tactile sensor array (548 sensors) is assembled on a knitted glove uniformly distributed over the hand.

Sundaram et al., Learning the signatures of the human grasp using a scalable tactile glove. Nature 2019

Unimodal Representation – Sensors



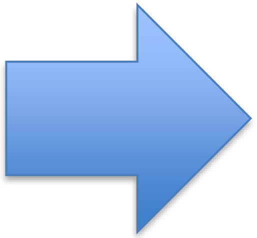
Force-Torque Sensor

Time series data across six-axis Force-Torque sensor:
T × 6 signal.

Proprioception

Measure values internal to the system (robot); e.g. motor speed, wheel load, **robot arm joint angles**, battery voltage.

Time series data across current position and velocity of the end-effector:
T × 2d signal.



Next action

Lee et al., Making Sense of Vision and Touch: Self-Supervised Learning of Multimodal Representations for Contact-Rich Tasks. ICRA 2019

Unimodal Representation – Tables

Caption — Singapore Cup

Attribute / Cell:

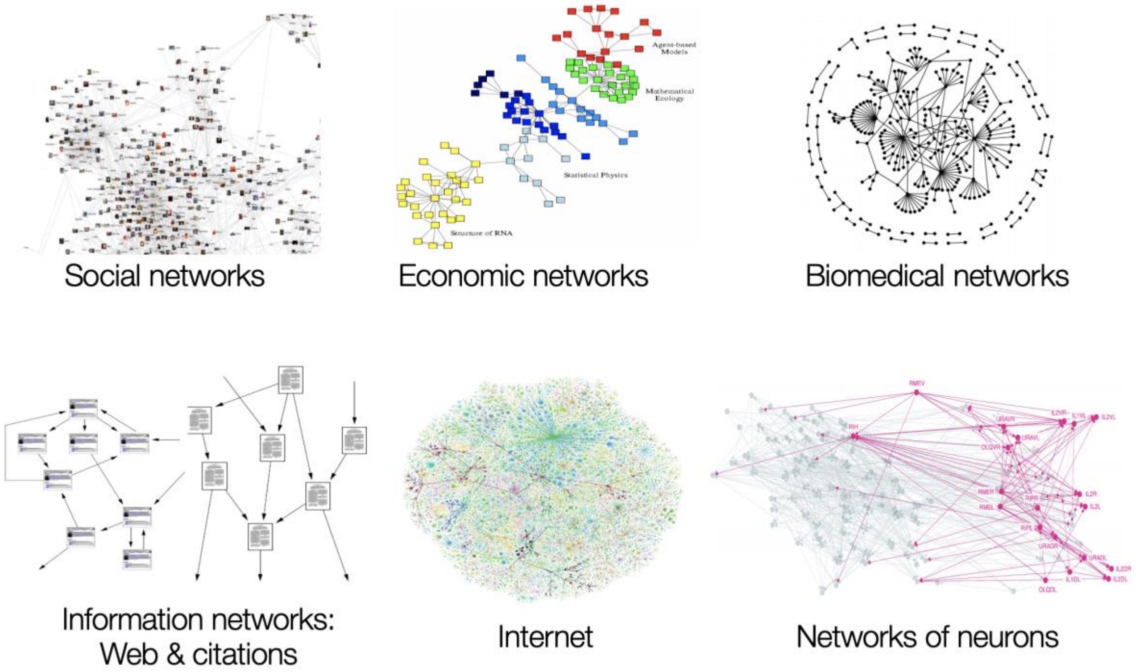
Year	Champions	Runners-up
1996	Geylang United	Singapore Armed Forces
1997	Singapore Armed Forces	Woodlands Wellington
1998	Tanjong Pagar United	Sembawang Rangers

Text — *Singapore Armed forces was the champion of Singapore Cup in 1997.*

Table-to-text generation

Bao et al., Table-to-Text: Describing Table Region with Natural Language. AAAI 2018

Unimodal Representation – Graphs



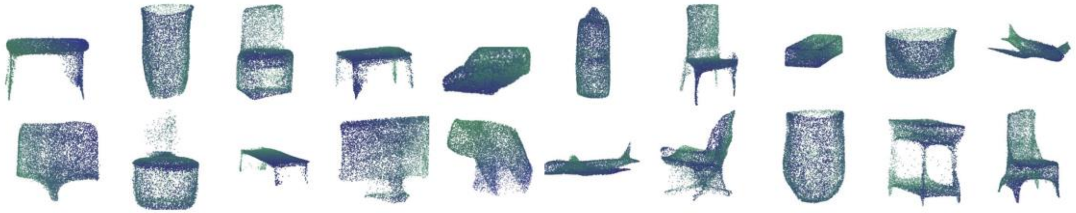
Tasks on graphs:
Node classification
Link prediction
…

Using graphs:
Knowledge graphs
for QA
Social network for
sentiment analysis
…

Hamilton and Tang, Tutorial on Graph Representation Learning. AAAI 2019

Unimodal Representation – Sets

Sets



Point clouds

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

Zaheer et al., DeepSets. NeurIPS 2017, Li et al., Point Cloud GAN. arxiv 2018

Machine Learning – Basic Concepts

Simplest Classifier ?

?

Traffic light

-or-

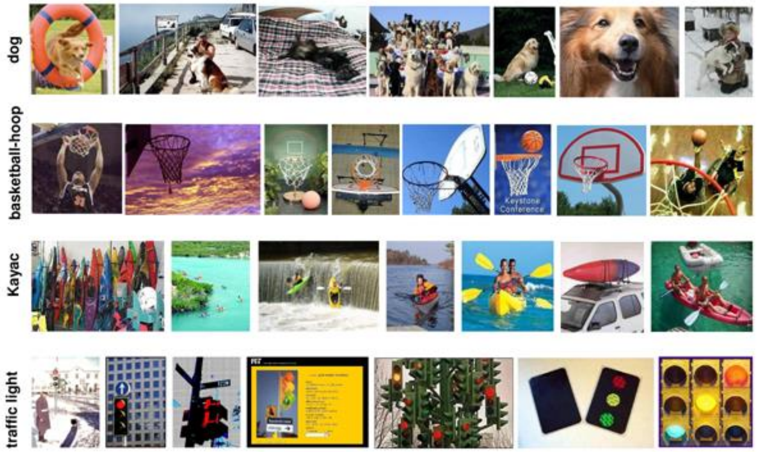
Dog

-or-

Basket

-or-

Kayak ?



Dataset

Simple Classifier: Nearest Neighbor

?

Traffic light

-or-

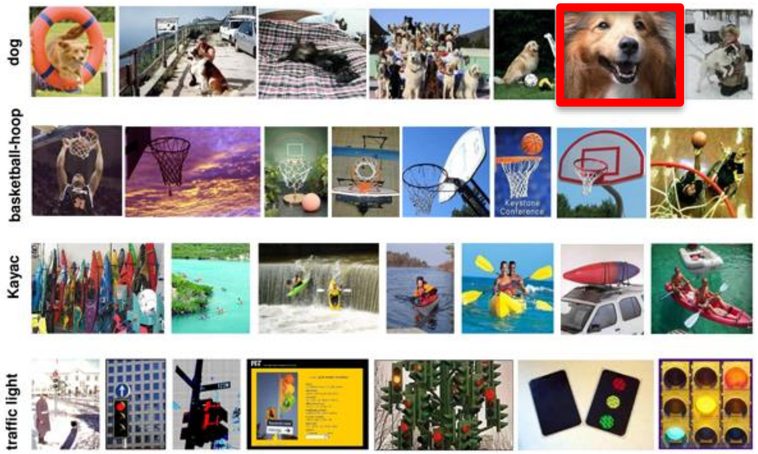
Dog

-or-

Basket

-or-

Kayak ?



Training

Nearest Neighbor Classifier

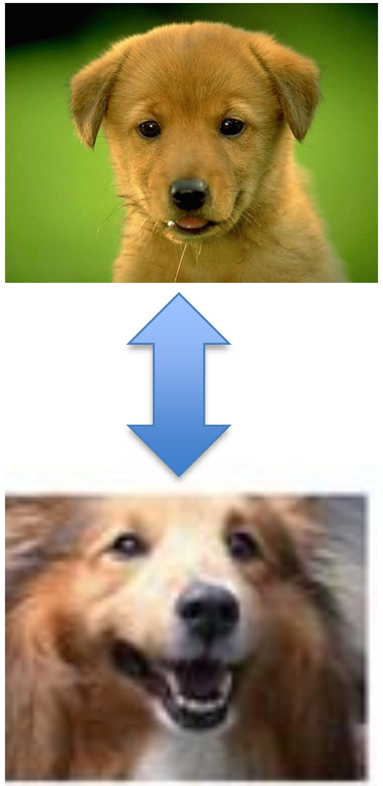
- Non-parametric approaches—key ideas:
  - *“Let the data speak for themselves”*
  - *“Predict new cases based on similar cases”*
  - *“Use multiple local models instead of a single global model”*

- What is the complexity of the NN classifier w.r.t training set of N images and test set of M images?
  - at training time?

    O(1)
  - At test time?

    O(N)

Simple Classifier: Nearest Neighbor



Distance metrics

L1 (Manhattan) distance:

$$d_1(x_1, x_2) = \sum_j \left| x_1^j - x_2^j \right|$$

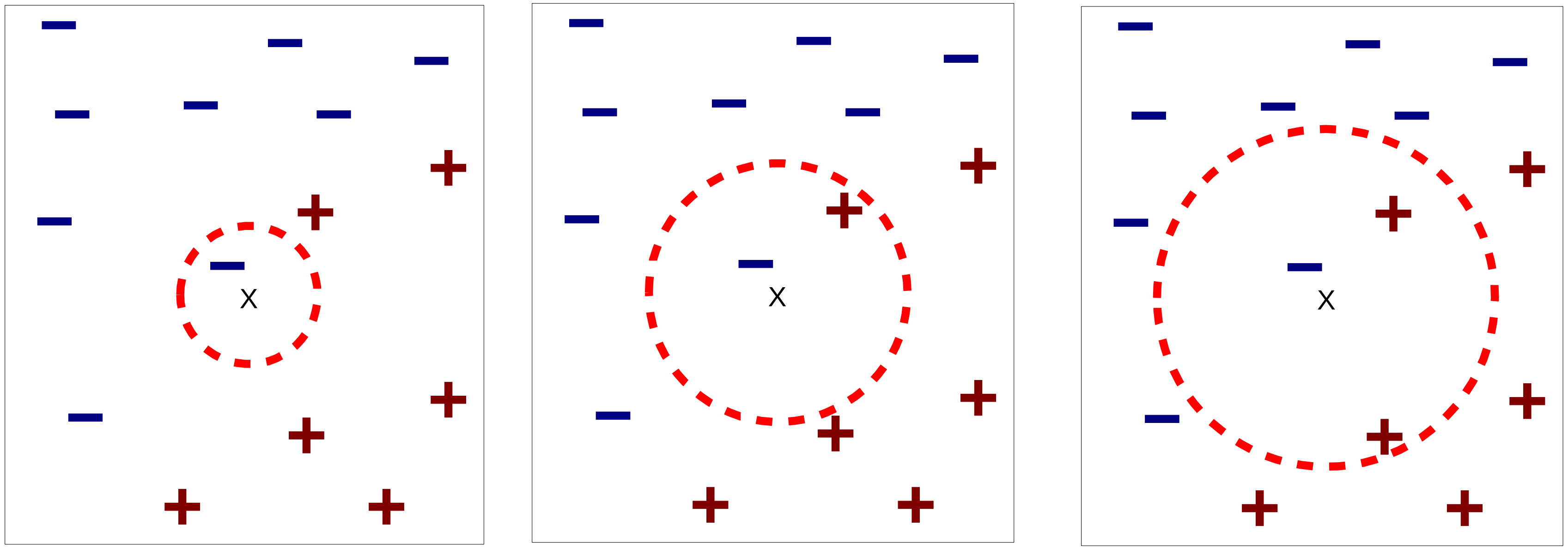
L2 (Eucledian) distance:

$$d_2(x_1, x_2) = \sqrt{\sum_j \left(x_1^j - x_2^j \right)^2}$$

Which distance metric to use?

First hyper-parameter!

Definition of K-Nearest Neighbor



(a) 1-nearest neighbor (b) 2-nearest neighbor (c) 3-nearest neighbor

What value should we set K?

Second hyper-parameter!

Going beyond Classifiers…

Representation Learning: A Review and New Perspectives

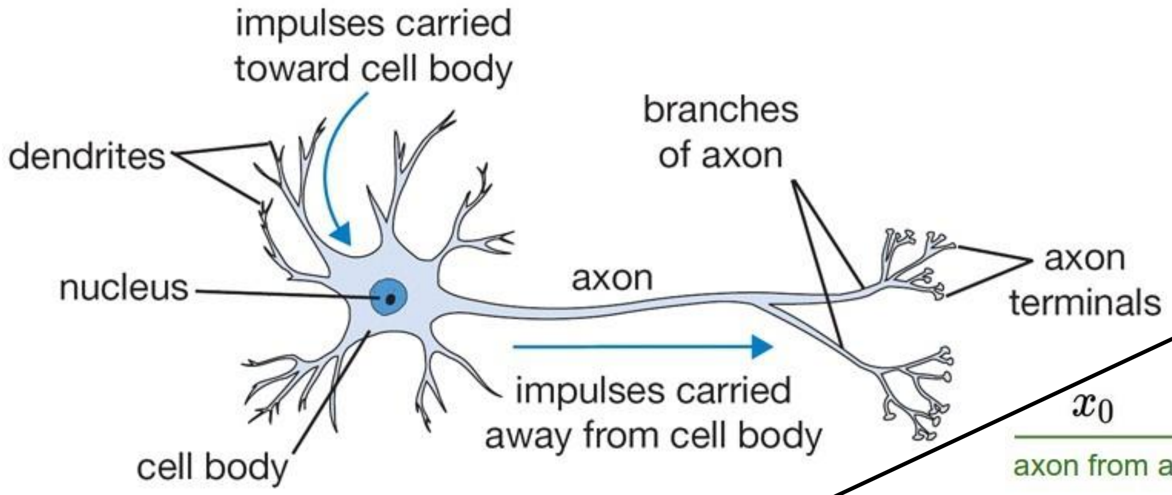
Yoshua Bengio†, Aaron Courville, and Pascal Vincent†
Department of computer science and operations research, U. Montreal
† also, Canadian Institute for Advanced Research (CIFAR)

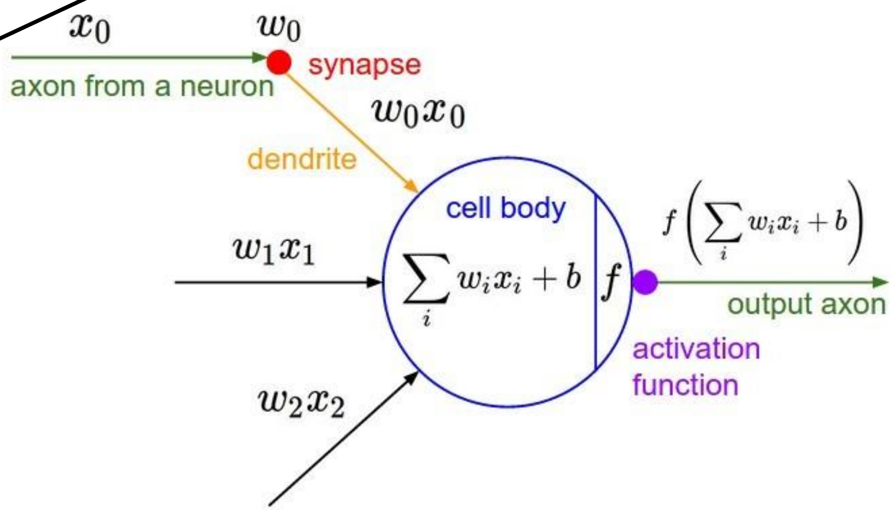
Data-driven feature representation learning: **Deep neural networks**

Basic Concepts: Neural Networks

Neural Networks – inspiration

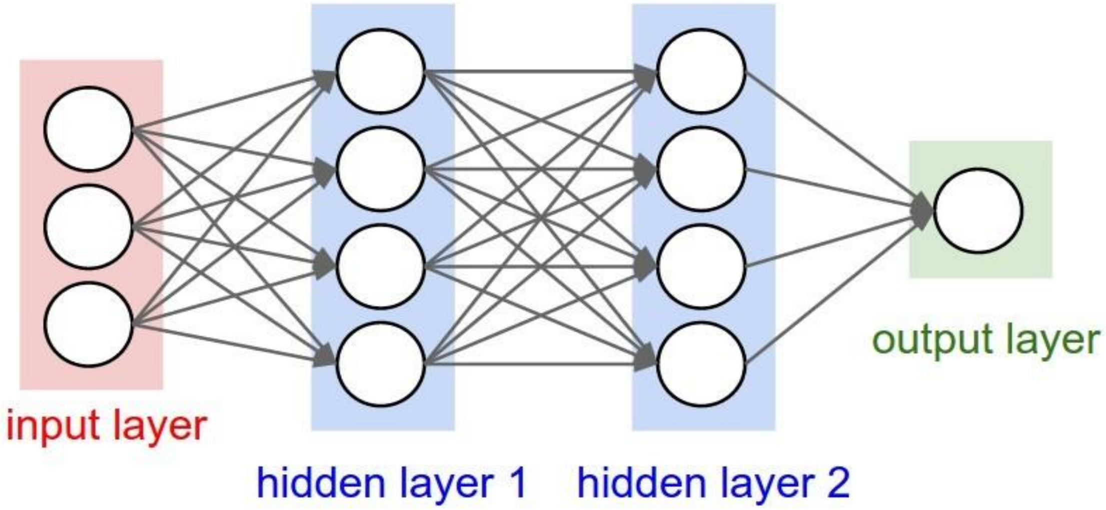
- Made up of artificial neurons





Neural Networks – score function

- Made up of artificial neurons
  - Linear function (dot product) followed by a nonlinear activation function
- Example a Multi Layer Perceptron



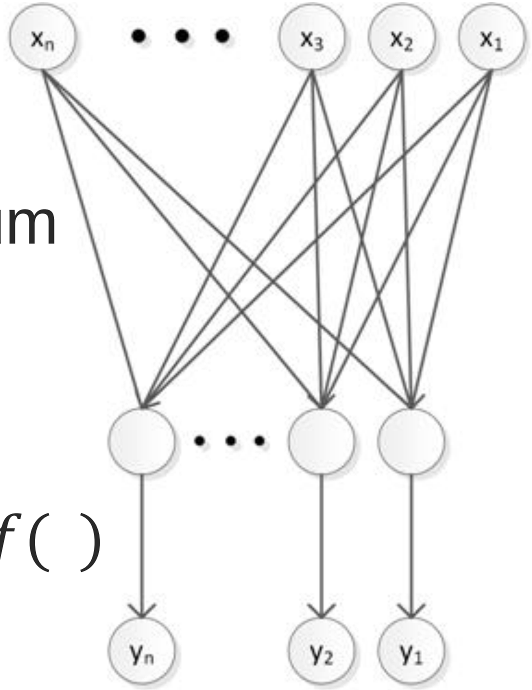
Basic Neural Network building block

Input

Weighted sum
$Wx + b$

Activation function $f(\ )$

Output



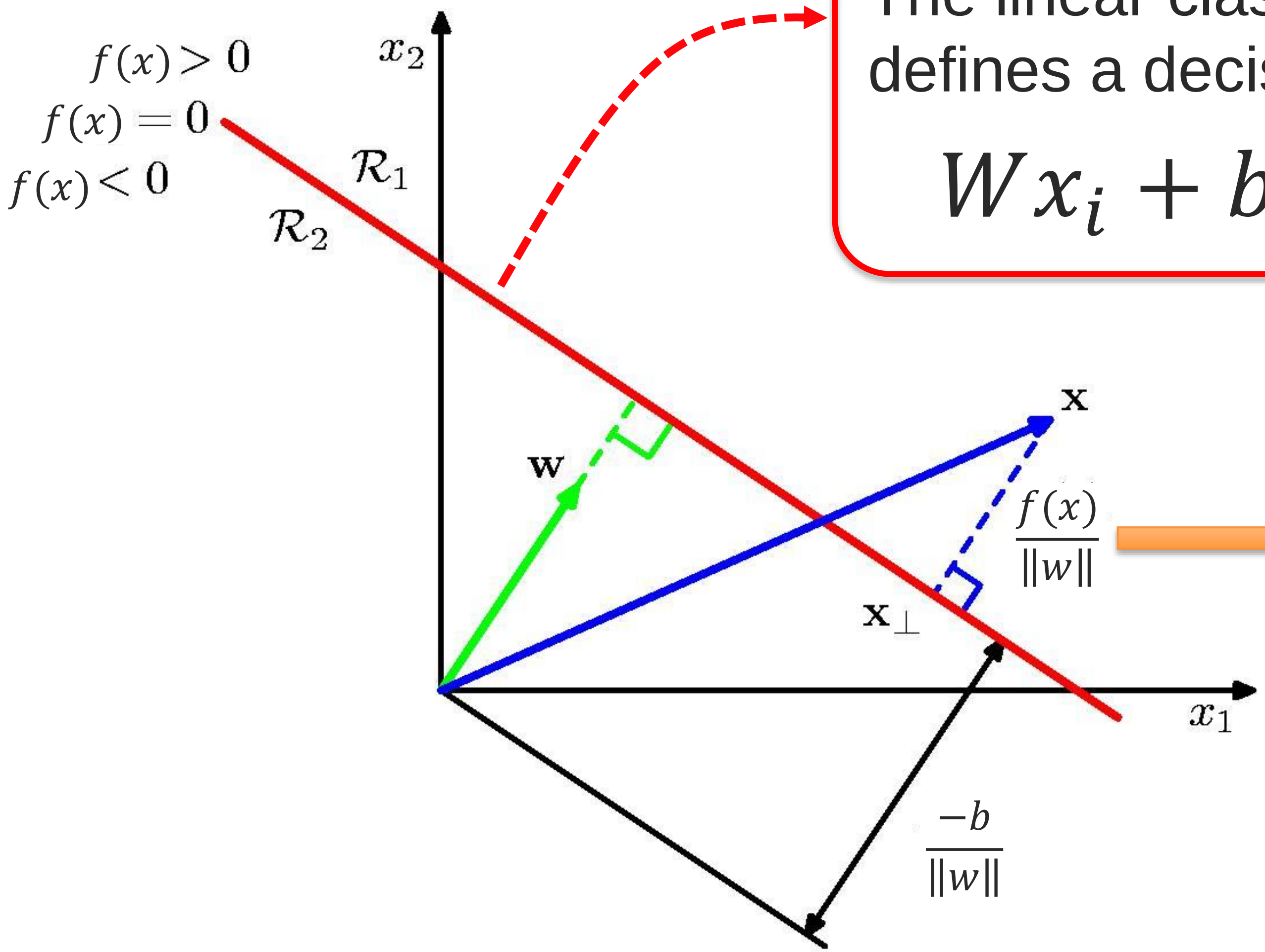
Linear classifier

Nonlinear activation function

$$\hat{y} = f(Wx + b)$$

Interpreting a Linear Classifier

$$f(x_i; W, b) = Wx_i + b$$



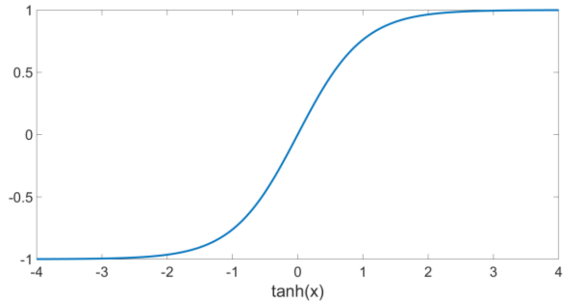
The linear classifier defines a decision plane:

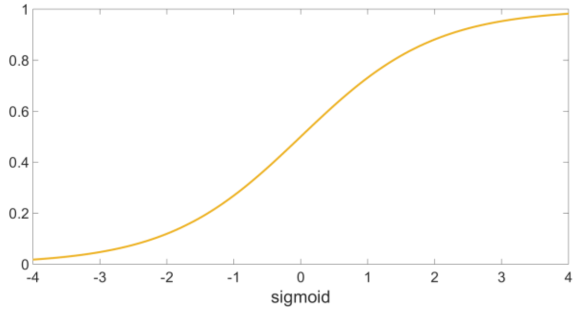
$$Wx_i + b > 0$$

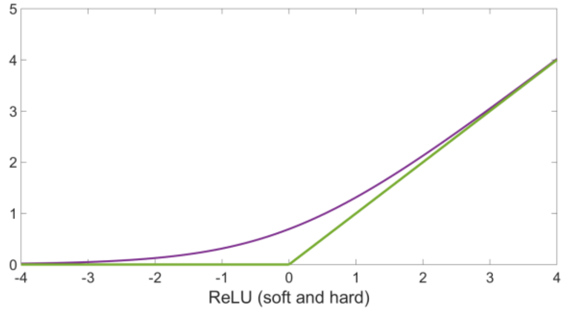
Output can be seen as distance to decision plane

Neural Networks – activation function

- $f(x) = \tanh(x)$
- Sigmoid - $f(x) = (1 + e^{-x})^{-1}$
- Linear – $f(x) = ax + b$
- ReLU $f(x) = \max(0, x) \sim \log(1 + \exp(x))$
  - Rectifier Linear Units
  - Faster training - no gradient vanishing
  - Induces sparsity







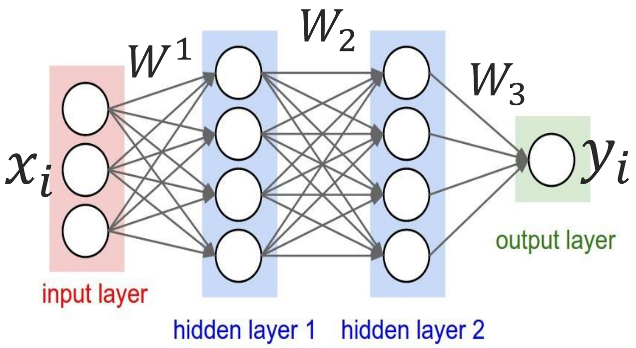
Multi-Layer Feedforward Network

Activation functions (individual layers)

$$f_{1;W_1}(x) = \sigma(W_1 x + b_1)$$

$$f_{2;W_2}(x) = \sigma(W_2 x + b_2)$$

$$f_{3;W_3}(x) = \sigma(W_3 x + b_3)$$



Score function

$$y_i = f(x_i) = f_{3;W_3}(f_{2;W_2}(f_{1;W_1}(x_i)))$$

How to integrate all the output scores?

Neural Network – Loss Function

(or cost function or objective)

Scores $f(x_i; W)$ | **Label** $y_i = 2\ (dog)$ → **Loss** $L_i = ?$

Image x_i

(Size: 32*32*3)

0 (duck) ?	-12.3
1 (cat) ?	45.6
2 (dog) ?	98.7
3 (pig) ?	12.2
4 (bird) ?	-45.3

Multi-class problem

How to assign only one number representing how “unhappy” we are about these scores?

The loss function quantifies the amount by which the prediction scores deviate from the actual values.

How to normalize the scores?

First Loss Function: Cross-Entropy Loss

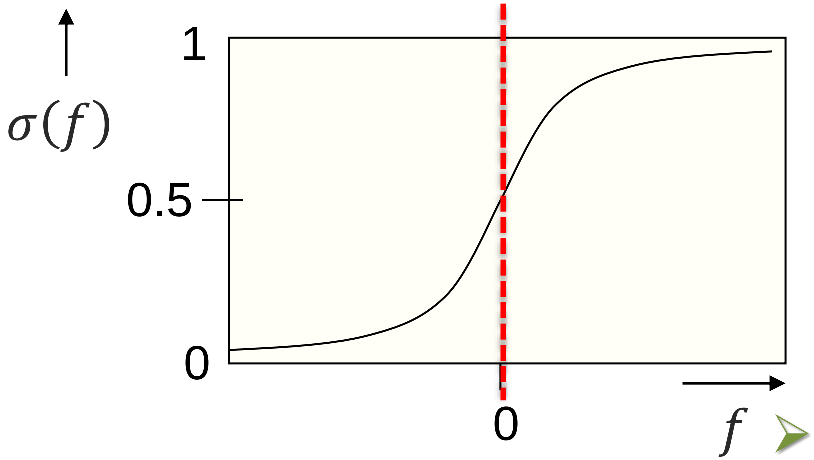
(or logistic loss)

Logistic function: $$\sigma(f) = \frac{1}{1+e^{-f}}$$

Logistic regression:
(two classes)

$$p(y_i = \text{"}dog\text{"} | x_i; w) \;\; = \sigma(w^T x_i)$$

= true
for two-class problem



➢ **Score function**

First Loss Function: Cross-Entropy Loss

(or logistic loss)

Logistic function: $$\sigma(f) = \frac{1}{1+e^{-f}}$$

Logistic regression:
(two classes)

$$p(y_i = \text{"}dog\text{"} | x_i; w) \quad = \sigma(w^T x_i)$$

= true
for two-class problem

Softmax function:
(multiple classes)

$$p(y_i | x_i; W) = \frac{e^{f_{y_i}}}{\sum_j e^{f_j}}$$

Cross-entropy loss:

$$L_i = \boxed{-\log\left(\boxed{\frac{e^{f_{y_i}}}{\sum_j e^{f_j}}}\right)}$$

Softmax function

Minimizing the negative log likelihood.

Second Loss Function: Hinge Loss

(or max-margin loss or Multi-class SVM loss)

$$\boxed{L_i} = \boxed{\sum_{j \neq y_i}} \max(0, \boxed{f(x_i, W)_j - f(x_i, W)_{y_i}} + \Delta)$$

loss due to example i

sum over all incorrect labels

difference between the correct class score and incorrect class score

Optimization – Learning model parameters

Learning model parameters

We have our training data

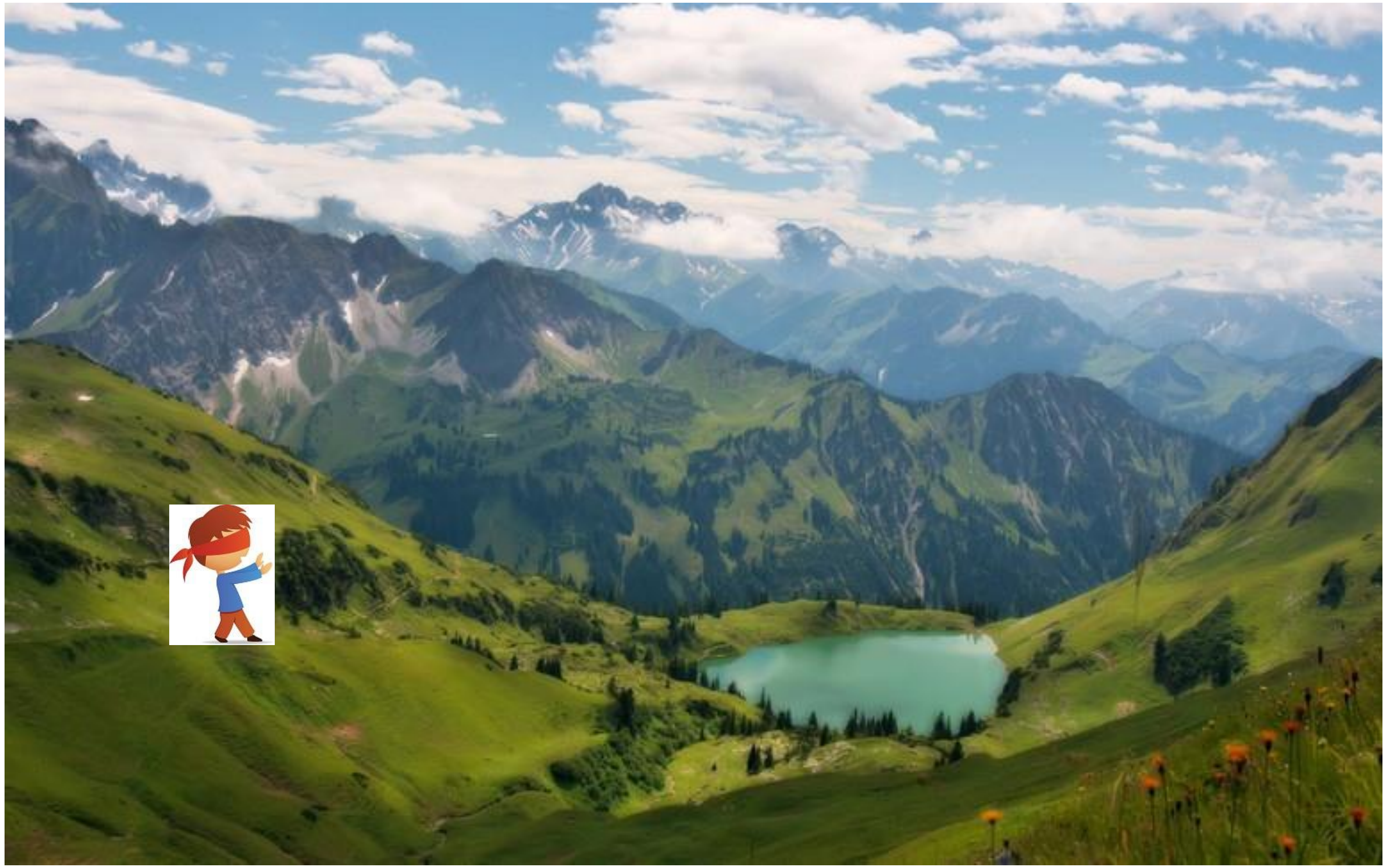
- $\mathrm{X} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \dots, \boldsymbol{x}_n\}$ (e.g. images, videos, text etc.)
- $\mathrm{Y} = \{y_1, y_2, \dots, y_n\}$ (labels)

We want to learn the W (weights and biases) that leads to best loss

$$\underset{W}{\operatorname{argmin}}[L(\mathrm{X}, \mathrm{Y}, W)]$$

The notation means find W for which $L(\mathrm{X}, \mathrm{Y}, W)$ has the lowest value

Optimization



Analytical gradient

If we know the function and it is **differentiable**

- Derivative/gradient is defined at every point in *f*
- Sometimes use differentiable approximations
- Some are locally differentiable

Examples:

$$f(x) = \frac{1}{1+e^{-x}}; \frac{df}{dx} = (1 - f(x))f(x)$$

$$f(x) = (x - y)^2; \frac{df}{dx} = 2(x - y)$$

How to follow the gradient

Many methods for optimization

- **Gradient Descent (actually the “simplest” one)**
- Newton methods (use Hessian – second derivative)
- Quasi-Newton (use approximate Hessian)
  - BFGS
  - LBFGS
  - Don’t require learning rates (fewer hyperparameters)
  - But, do not work with stochastic and batch methods so rarely used to train modern Neural Networks

All of them look at the gradient

- Very few non gradient based optimization methods

Parameter Update Strategies

Gradient descent:

$$\theta^{(t+1)} = \theta^{t} - \epsilon_k \boxed{\nabla_\theta L}$$

Gradient of our loss function

New model parameters

Previous parameters

Learning rate at iteration *k*

$$\epsilon_k = (1-\alpha)\epsilon_0 + \alpha \boxed{\epsilon_\tau}$$

Decay learning rate linearly until iteration τ

Learning rate at iteration *k*

Decay

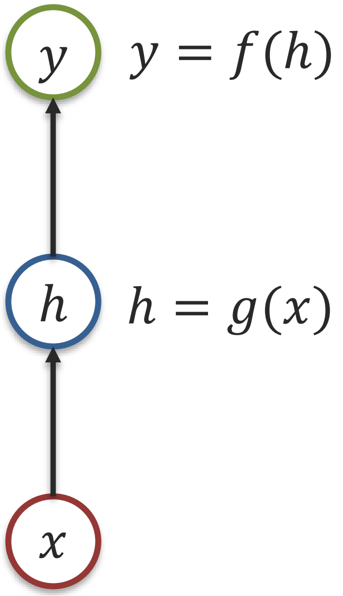
Initial learning rate

Neural Network Gradient

Gradient Computation

Chain rule:

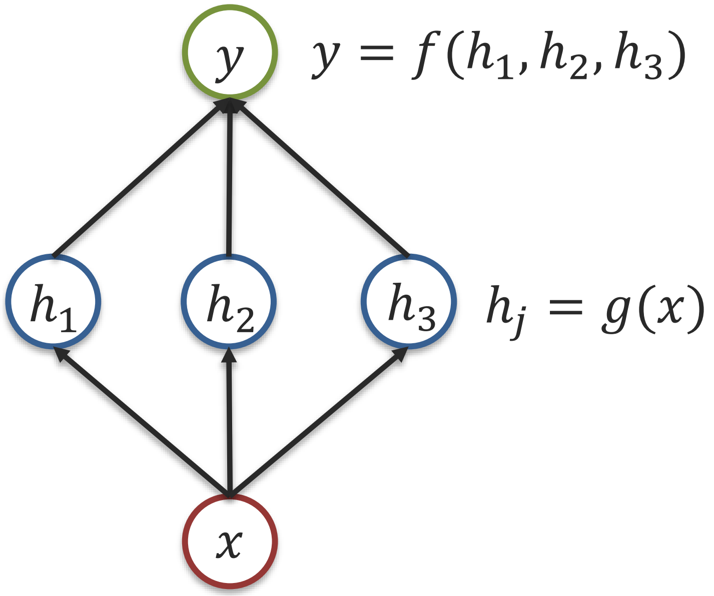
$$\frac{\partial y}{\partial x} = \frac{\partial y}{\partial h}\frac{\partial h}{\partial x}$$



Optimization: Gradient Computation

Multiple-path chain rule:

$$\frac{\partial y}{\partial x} = \sum_j \frac{\partial y}{\partial h_j} \frac{\partial h_j}{\partial x}$$



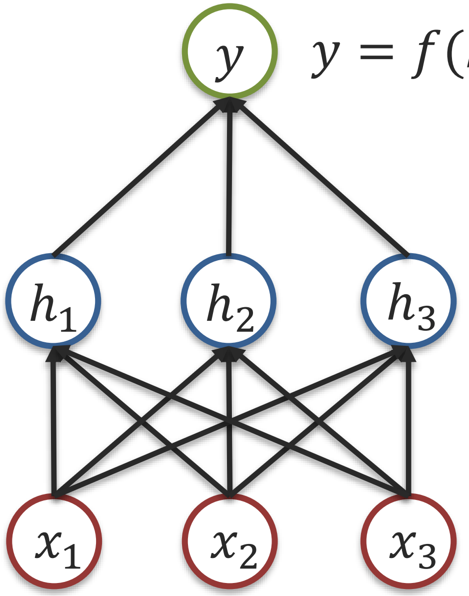
Optimization: Gradient Computation

Multiple-path chain rule:

$$\frac{\partial y}{\partial x_1} = \sum_j \frac{\partial y}{\partial h_j}\frac{\partial h_j}{\partial x_1}$$

$$\frac{\partial y}{\partial x_2} = \sum_j \frac{\partial y}{\partial h_j}\frac{\partial h_j}{\partial x_2}$$

$$\frac{\partial y}{\partial x_3} = \sum_j \frac{\partial y}{\partial h_j}\frac{\partial h_j}{\partial x_3}$$



$y = f(h_1, h_2, h_3)$

$h_j = g(\boldsymbol{x})$

Optimization: Gradient Computation

Vector representation:

$$\nabla_{\boldsymbol{x}}\, y = \left[\frac{\partial y}{\partial x_1}, \frac{\partial y}{\partial x_2}, \frac{\partial y}{\partial x_3}\right]$$

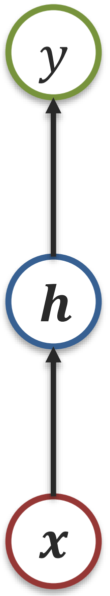
Gradient

$$\nabla_{\boldsymbol{x}}\, y = \left(\boxed{\frac{\partial \boldsymbol{h}}{\partial \boldsymbol{x}}}\right)^T \nabla_{\boldsymbol{h}}\, y$$

"backprop" Gradient

"local" Jacobian

(matrix of size $|h| \times |x|$ computed using partial derivatives)



$y = f(\boldsymbol{h})$

$\boldsymbol{h} = g(\boldsymbol{x})$

Backpropagation Algorithm (efficient gradient)

Forward pass

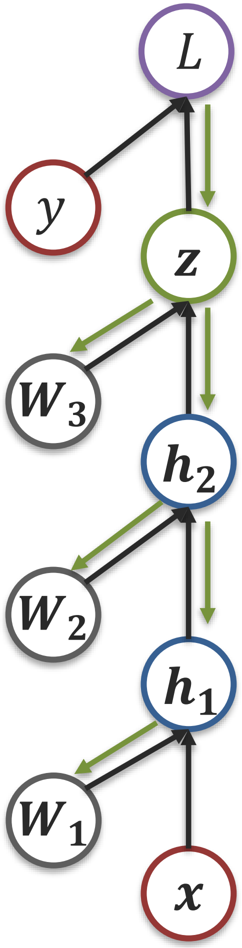
- Following the graph topology, compute value of each unit

Backpropagation pass

- Initialize output gradient = 1
- Compute "local" Jacobian matrix using values from forward pass
- Use the chain rule:

Gradient = "local" Jacobian x "backprop" gradient

- Why is this rule important?



$L = -logP(Y = y|\boldsymbol{z})$

(cross-entropy)

$\boldsymbol{z} = matmult(\boldsymbol{h}_2, \boldsymbol{W}_3)$

$\boldsymbol{h}_2 = f(\boldsymbol{h}_1; \boldsymbol{W}_2)$

$\boldsymbol{h}_1 = f(\boldsymbol{x}; \boldsymbol{W}_1)$

Computational Graph: Multi-layer Feedforward Network

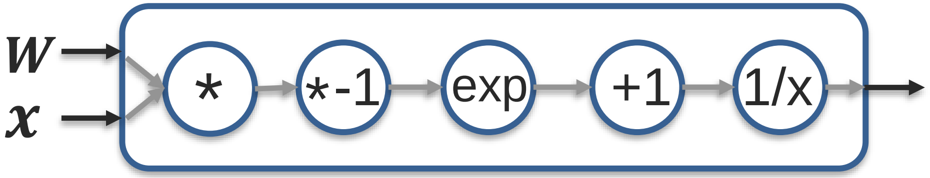
Computational unit:

$\boldsymbol{h} = f(\boldsymbol{x}; \boldsymbol{W})$

- Multiple input
- One output
- Vector/tensor

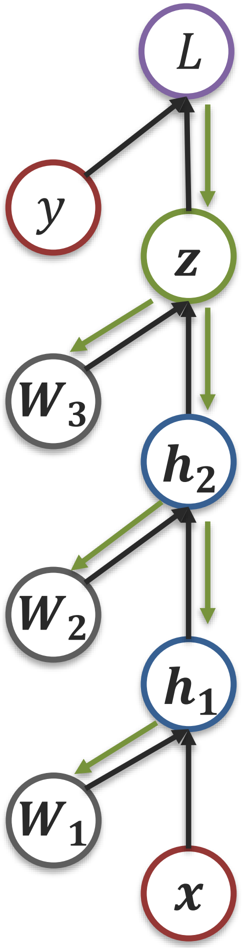
- Sigmoid unit:

$$h_j = (1 + e^{-W_j x})^{-1}$$



Differentiable "unit" function!

(or close approximation to compute "local Jacobian)



$L = -logP(Y = y|\boldsymbol{z})$

(cross-entropy)

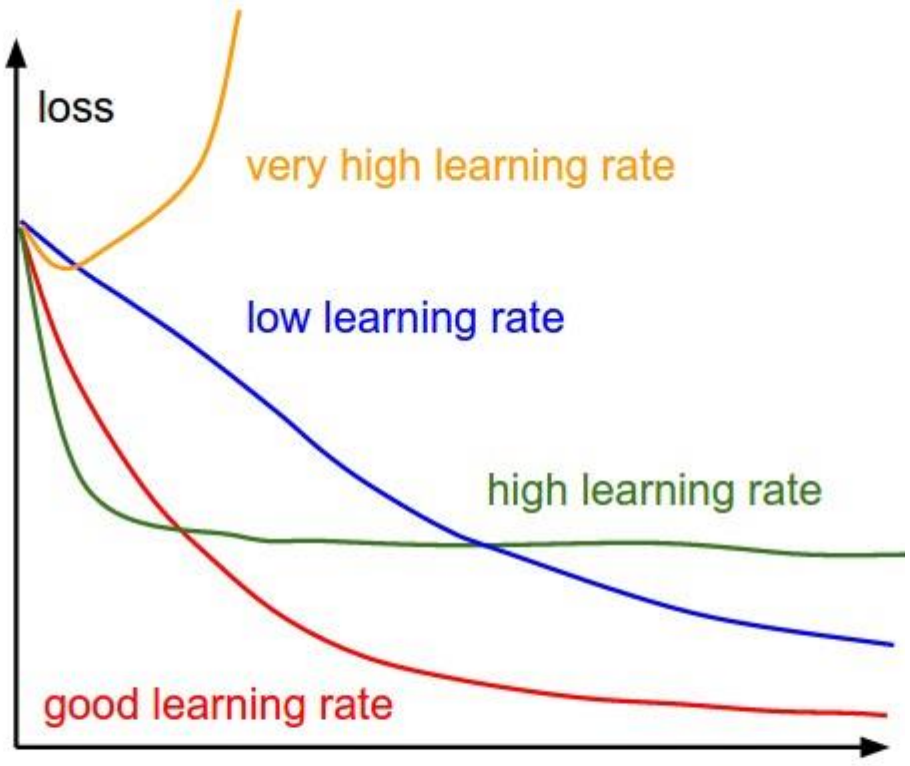
$\boldsymbol{z} = matmult(\boldsymbol{h}_2, \boldsymbol{W}_3)$

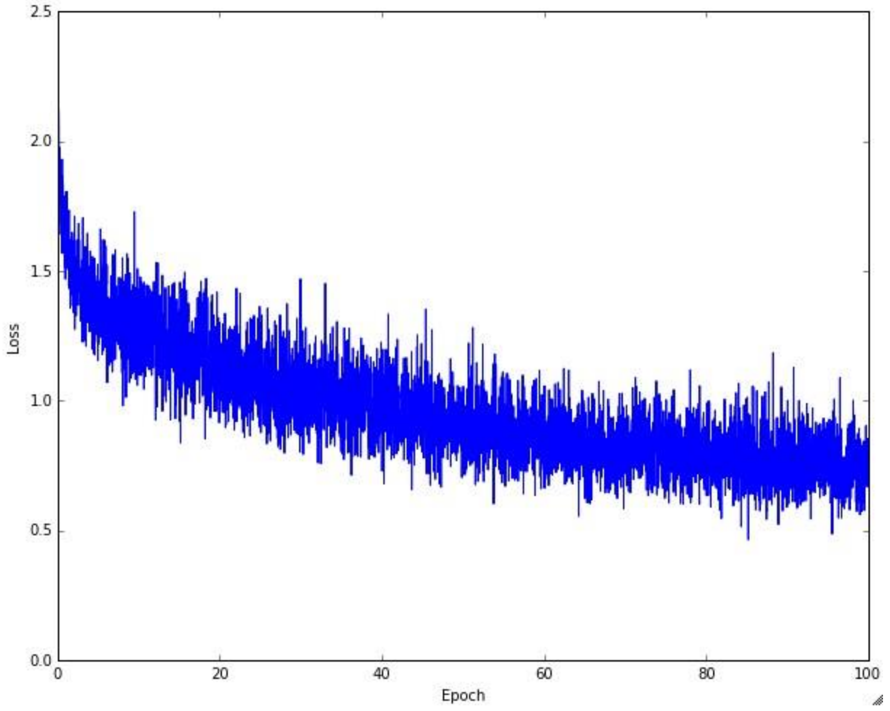
$\boldsymbol{h}_2 = f(\boldsymbol{h}_1; \boldsymbol{W}_2)$

$\boldsymbol{h}_1 = f(\boldsymbol{x}; \boldsymbol{W}_1)$

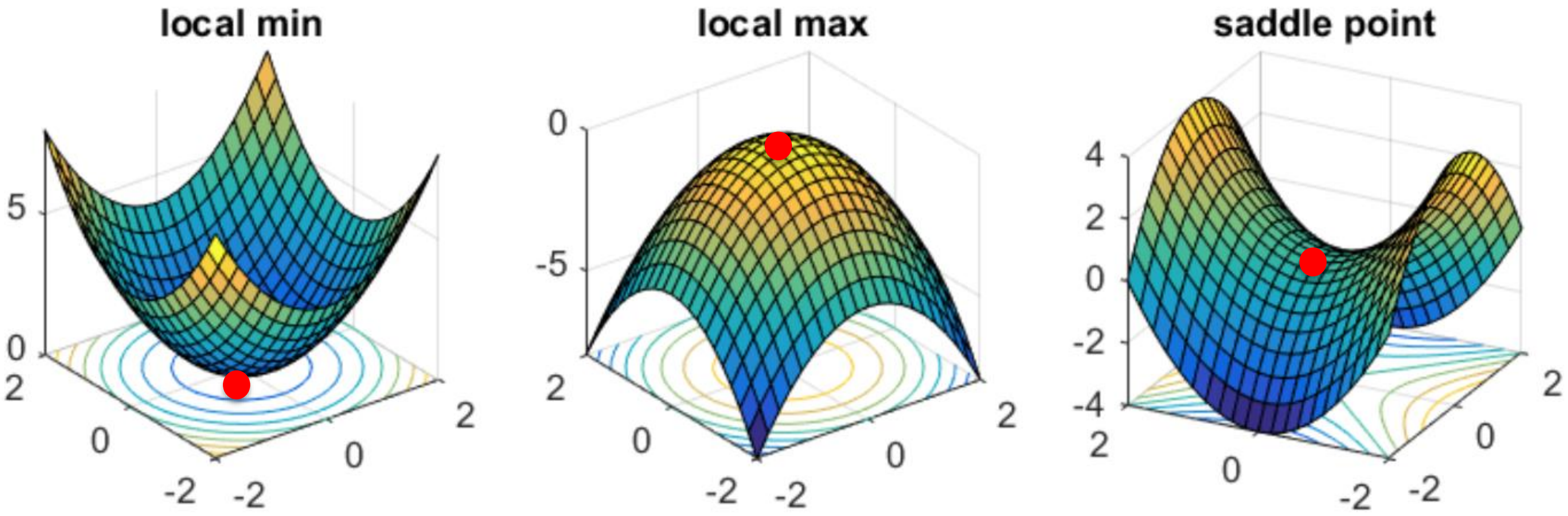
Optimization:
Some Practical Guidelines

Interpreting learning rates





Critical Points



Detecting Saddles

One way to detect saddles:

- Calculate Hessian at point x
- If Hessian is indefinite you have a saddle for sure.
- If Hessian is not indefinite you really can’t tell.

“My loss isn’t changing”

- You are definitely close to a critical point
  - You may be in a saddle point
  - You may be in the local minima/maxima
- One trick: quickly check the surrounding
  - Best practical trick if Hessian is not indefinite.

Adaptive Learning Rate

Key Idea: Let neurons who just started learning have huge learning rate.

Adaptive Learning Rate is an active area of research:

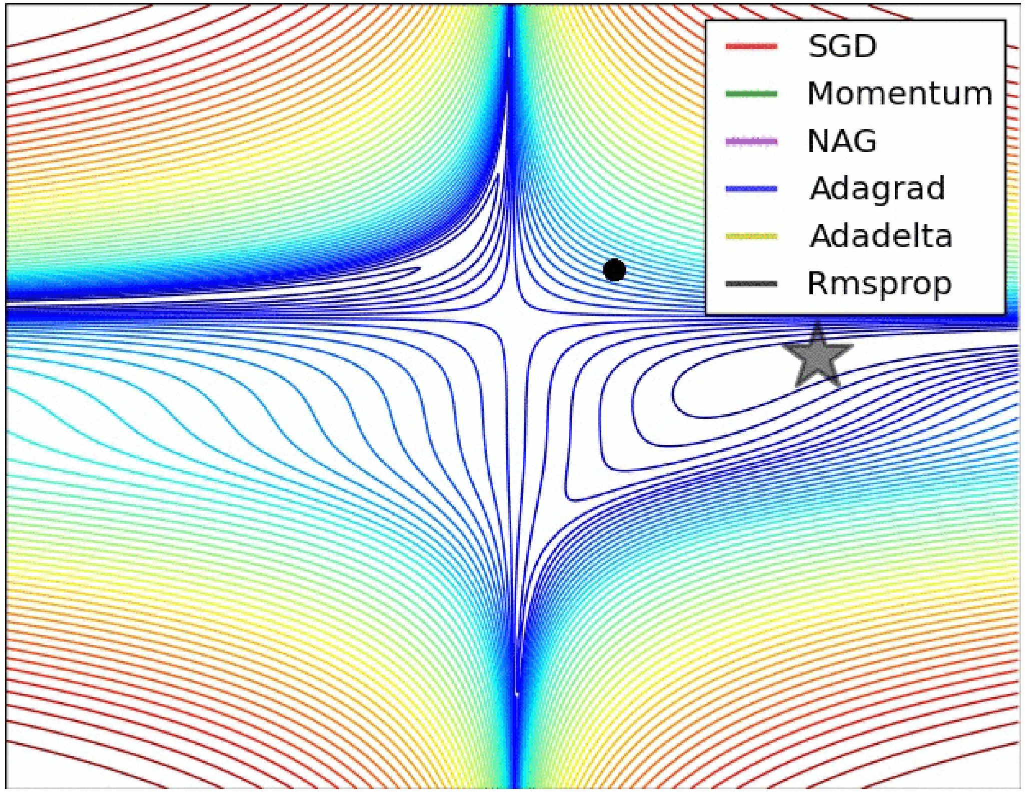
- Adadelta
- RMSProp

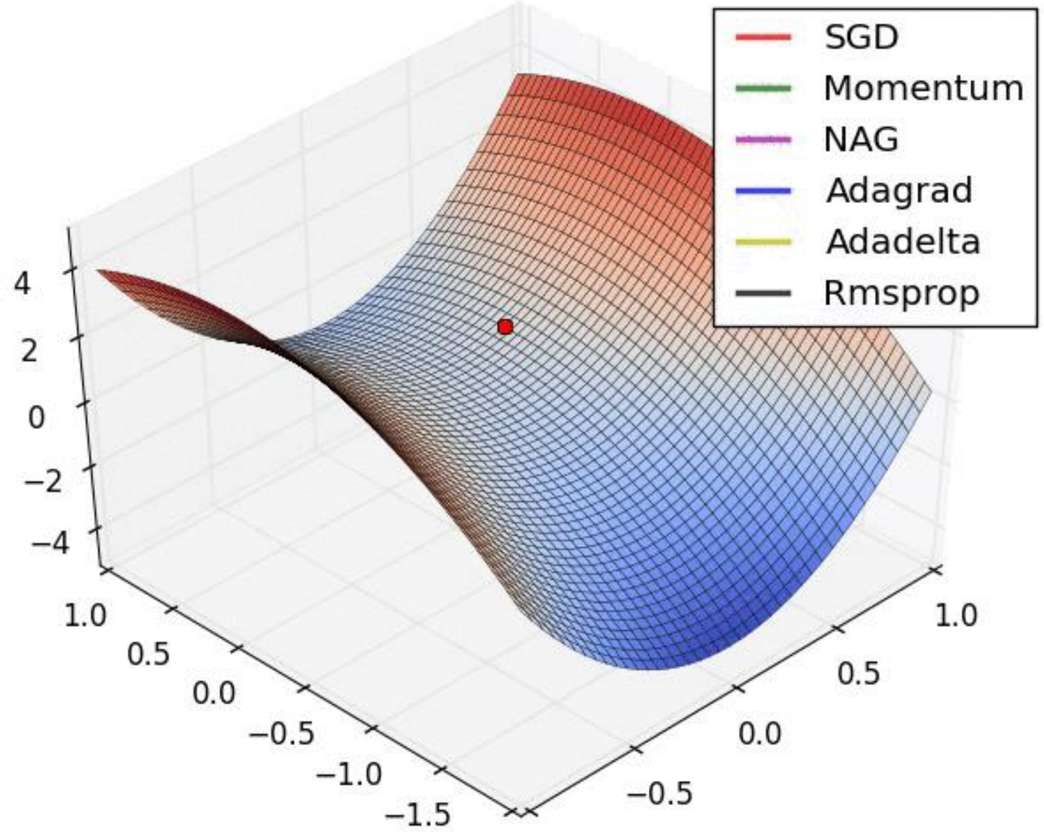
```
cache = decay_rate * cache + (1 - decay_rate) * dx**2
x += - learning_rate * dx / (np.sqrt(cache) + eps)
```

- Adam

```
m = beta1*m + (1-beta1)*dx
v = beta2*v + (1-beta2)*(dx**2)
x += - learning_rate * m / (np.sqrt(v) + eps)
```

Adaptive Learning Rate



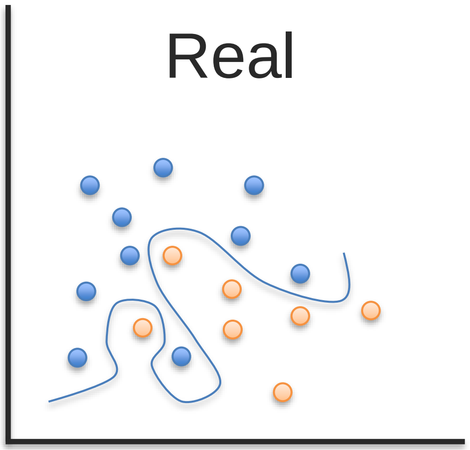


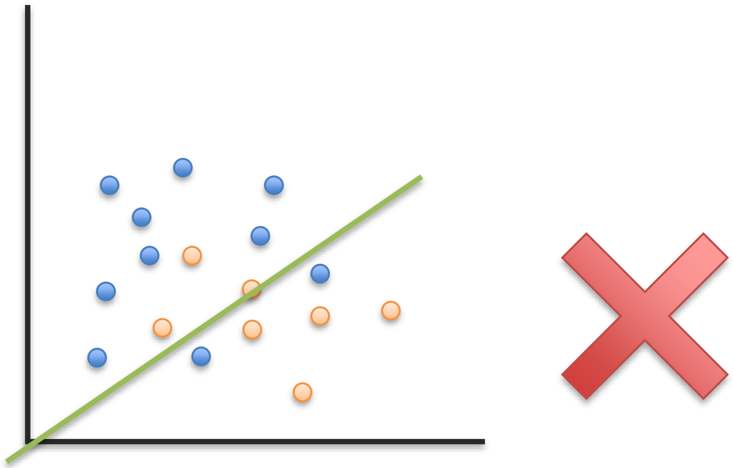
Bias-Variance

Problem of bias and variance

- Simple models are unlikely to find the solution to a hard problem, thus probability of finding the right model is low.

Not an issue these days!

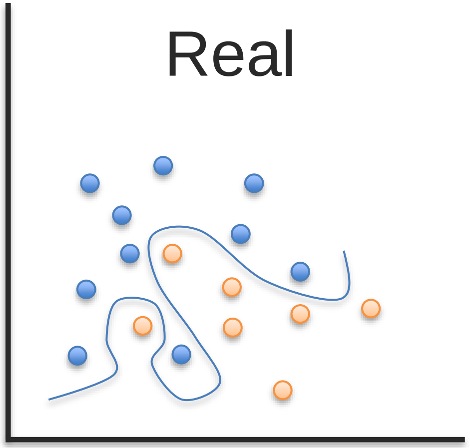


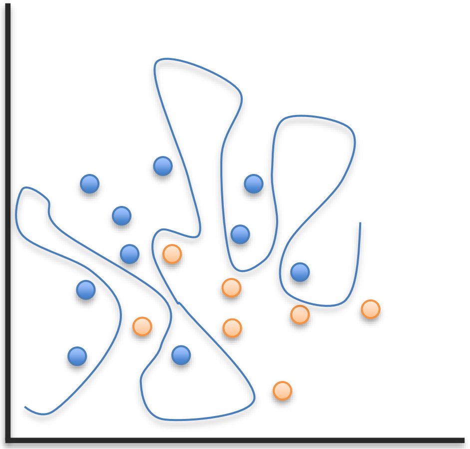


Bias-Variance

Problem of bias and variance

- Simple models are unlikely to find the solution to a hard problem, thus probability of finding the right model is low.
- Complex models find many solutions to a problem, thus probability of finding the right model is again low.



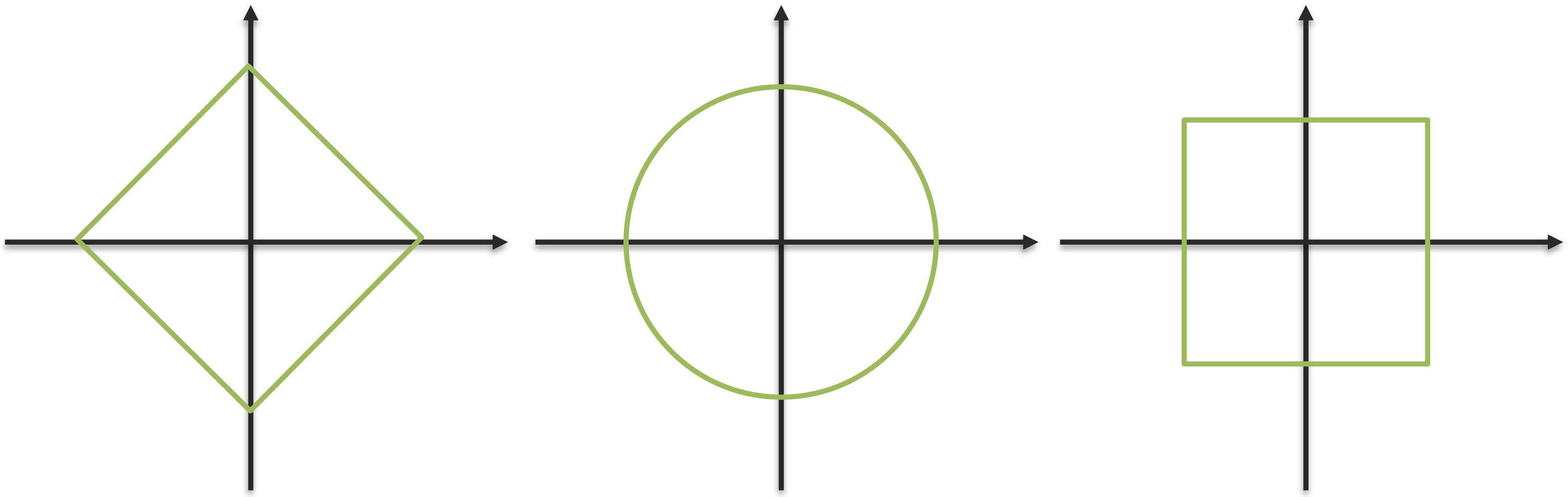


A big issue with deep learning!

Parameter Regularization

Adding prior to the network parameters

- L^p Norms



L^1 $\quad$ L^2 $\quad$ L^∞

Minimize: $Loss(x;\theta) + \propto ||\theta||$

Structural Regularization

Lots of models can learn everything.

- Go for simpler ones.

Occam's razor

Take advantage of the structure and "invariances" present in each modality:

- CNNs: translation invariance
- LSTMs: sequential structure
- GRUs: sequential structure