Paper Presentations

- If you are presenting next week, please come and see us in office hours this week. Ideally, we would like to walk through your slides with you.

- Each paper -- 25 mins of Paper Presentation + 5 mins Q/A

Guidelines for Paper Presentation

- Motivation
- Problem Statement
- Summary of Contributions
- Related Work
- Preliminaries + Background (Intuition First!)
- Approach (Intuition First!)
- Key Experimental Results
- Conclusions
  - Your Perspective on the Weaknesses of Paper
  - What would you do differently?

Generalized Additive Models

Intelligible Models for HealthCare

Caruana et. al.

Contributions

- Two case studies where Generalized Additive Models (intelligible) yield state-of-the-art accuracy.
  - Pneumonia risk prediction
  - 30-day hospital readmission

- Claim: GAMs is a class of models that can handle interpretability/accuracy trade-off quite well

Roadmap

- Motivation
- Intelligible Models
- Case study: Pneumonia risk
- Case study: 30 day readmission

Roadmap

- **Motivation**
- Intelligible Models
- Case study: Pneumonia risk
- Case study: 30 day readmission

Motivation

- A large project to evaluate application of ML to healthcare problems
  - Predicting probability of death (POD) for pneumonia patients
  - Most accurate models: neural nets (0.86 AUC)
  - Logistic regression: 0.77

- Logistic regression was used instead. Why?

Motivation

- Rule based learning method was also used
  - Insight: HasAsthma(x) $\Rightarrow$ LowerRisk(x)
  - Counterintuitive?

- Rule based system was intelligible making it easy to recognize and remove dangerous rules

- Lack of intelligibility made it harder to deploy neural nets because it was difficult to know other problems with the model

Motivation

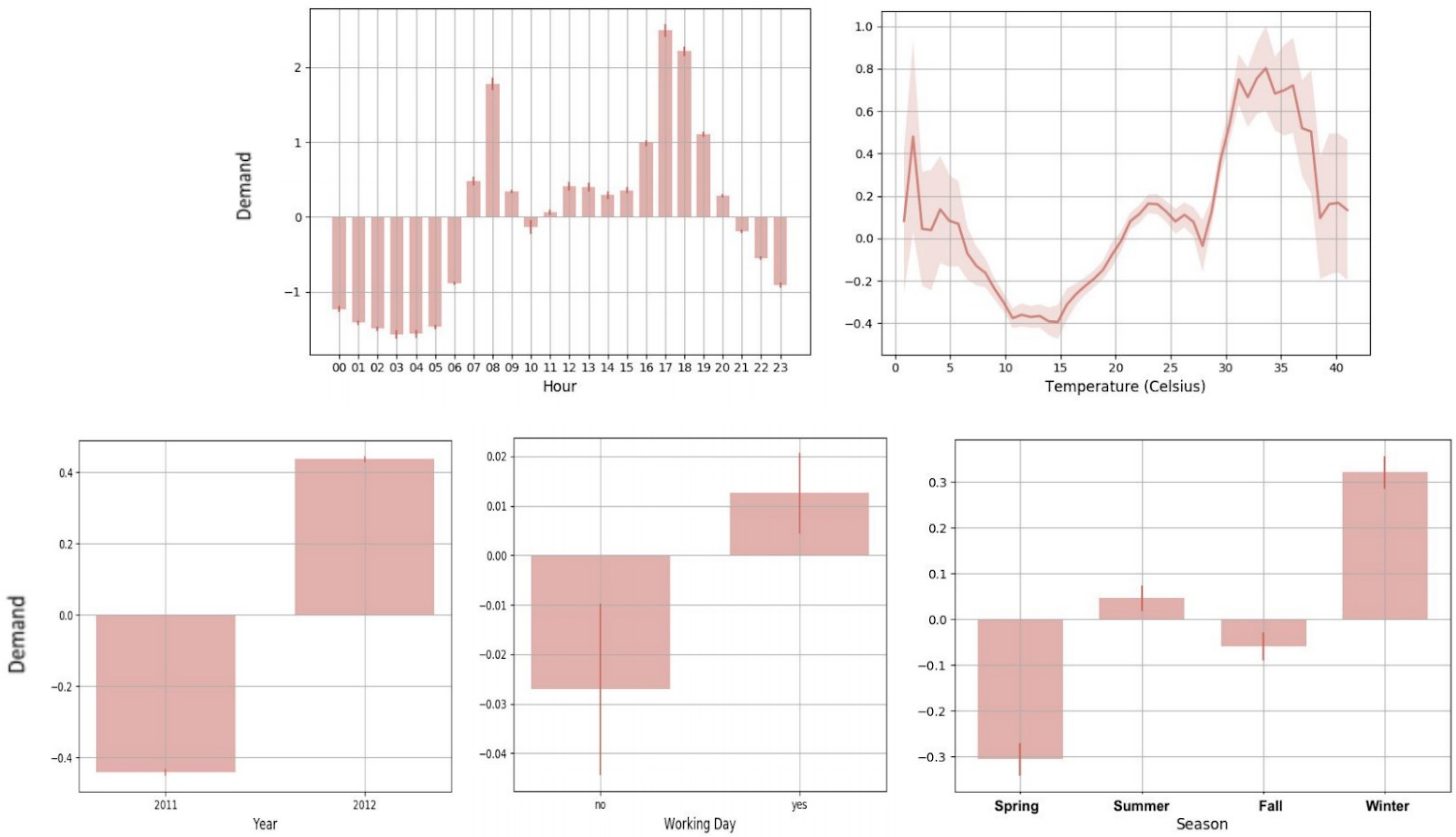
- Many more models but equally unintelligible today
  - SVMs, random forests, boosted trees

- GAMs are both intelligible and accurate!
  - Editable by experts

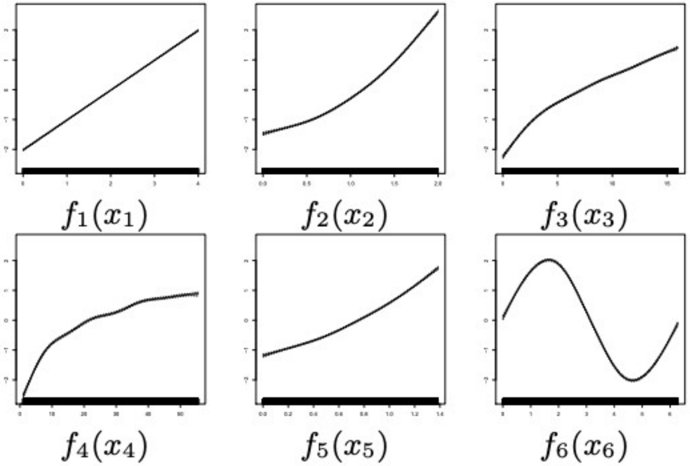
Roadmap

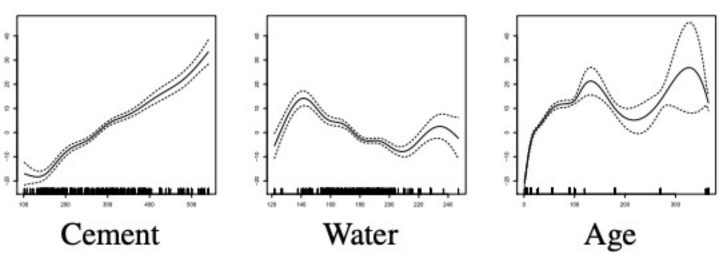
- ~~Motivation~~
- **Intelligible Models: GAMs and GA²MS**
- Case study: Pneumonia risk
- Case study: 30 day readmission

GAMs


Demand
Hour
00 01 02 03 04 05 06 07 08 09 10 11 12 13 14 15 16 17 18 19 20 21 22 23
Temperature (Celsius)
Demand
2011
2012
Year
no
yes
Working Day
Spring
Summer
Fall
Winter
Season

GAMs





GAMs and GA²Ms

$$g(E[y]) = \beta_0 + \sum f_j(x_j),$$

$$g(E[y]) = \beta_0 + \sum_j f_j(x_j) + \sum_{i \neq j} f_{ij}(x_i, x_j).$$

g is a link function: identity (additive model e.g., regression); log (E[y] / 1 – E[y]) (generalized additive model e.g., classification)

f_j is a shape function

Intelligibility and Accuracy

Model	Form	Intelligibility	Accuracy
Linear Model	$y = \beta_0 + \beta_1 x_1 + ... + \beta_n x_n$	+++	+
Generalized Linear Model	$g(y) = \beta_0 + \beta_1 x_1 + ... + \beta_n x_n$	+++	+
Additive Model	$y = f_1(x_1) + ... + f_n(x_n)$	++	++
Generalized Additive Model	$g(y) = f_1(x_1) + ... + f_n(x_n)$	++	++
Full Complexity Model	$y = f(x_1, ..., x_n)$	+	+++

Lou et. al., 2012

Shape Functions

- Regression Splines
- Trees
- Ensembles of Trees

GAMs and GA^2Ms

- Learning:
  - Represent each component as a spline
    - Least squares formulation; Optimization problem to balance smoothness and empirical error
  - Regression trees on a single/pair of features
  - Gradient boosting with bagging of shallow trees

- GA^2Ms: Build GAM first and and then detect and rank all possible pairs of interactions in the residual
  - Choose top k pairs
  - k determined by CV

Roadmap

- ~~Motivation~~
- ~~Intelligible Models~~
- **Case study: Pneumonia risk**
- Case study: 30 day readmission

Pneumonia Risk

- 14,199 pneumonia patients
- train set: 9847
- test set: 4352
- 46 features
- Predict POD
- 10.86% patients died from pneumonia (1542)

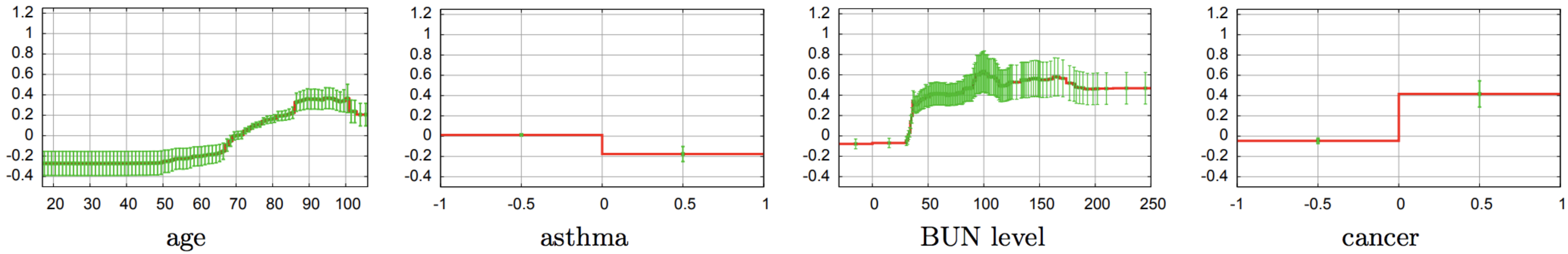
Pneumonia Risk: Features

Patient-history findings			
chronic lung disease	-	age	C
re-admission to hospital	-	gender	-
admitted through ER	-	diabetes mellitus	-
admitted from nursing home	-	asthma	-
congestive heart failure	-	cancer	-
ischemic heart disease	-	number of diseases	C
cerebrovascular disease	-	history of seizures	-
chronic liver disease	-	renal failure	-
history of chest pain	-		
Physical examination findings			
diastolic blood pressure	C	wheezing	-
gastrointestinal bleeding	-	stridor	-
respiration rate	C	heart murmur	-
altered mental status	-	temperature	C
heart rate	C		
Laboratory findings			
liver function tests	-	BUN level	C
glucose level	C	creatinine level	C
potassium level	C	albumin level	C
hematocrit	C	WBC count	C
percentage bands	C	pH	C
pO2	C	pCO2	C
sodium level	C		
Chest X-ray findings			
positive chest x-ray	-	lung infiltrate	-
pleural effusion	-	pneumothorax	-
cavitation/empyema	-	chest mass	-
lobe or lung collapse	-		

Pneumonia Risk: AUC

Model	Pneumonia
Logistic Regression	0.8432
GAM	0.8542
GA^2M	0.8576
Random Forests	0.8460
LogitBoost	0.8493

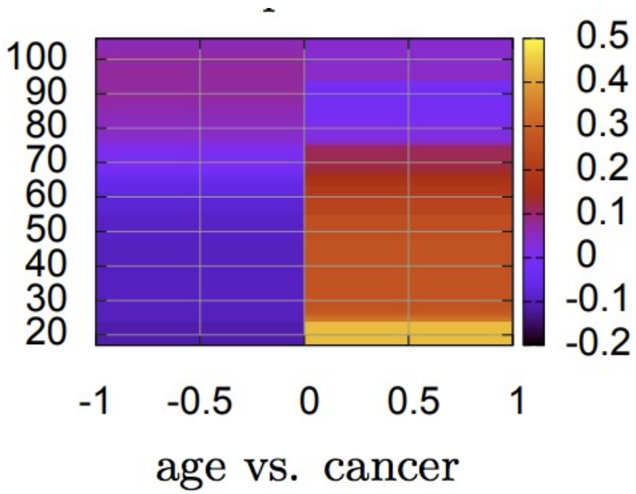
Understanding Outputs of GA²M



Blood Urea Nitrogen

Normal value: 10 to 20
0 means not ordered

Understanding Outputs of GA²M



Childhood cancers are associated with high risk of death

Roadmap

- ~~Motivation~~
- ~~Intelligible Models~~
- ~~Case study: Pneumonia risk~~
- **Case study: 30 day readmission**

30 day readmission

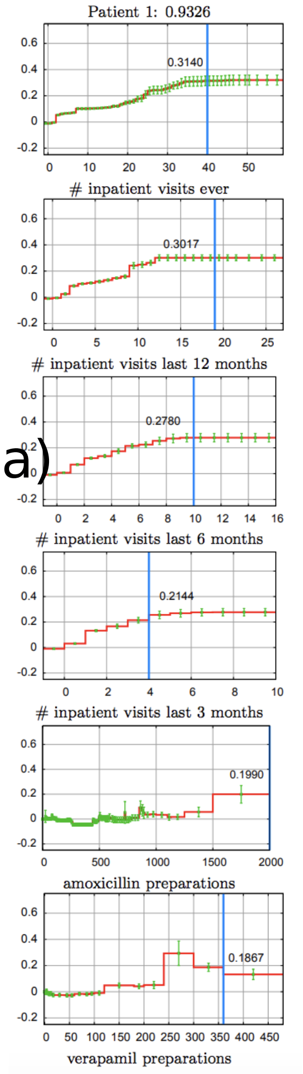
- 195K patients in train, 100K patients in test
- 3956 features
- Predict which patients are likely to be readmitted within 30 days of being released
- Hospitals with high readmission rates are penalized financially
  - Did not provide adequate care earlier
- 8.91% of patients readmitted within 30 days

AUC

Model	Readmission
Logistic Regression	0.7523
GAM	0.7795
GA^2M	0.7833
Random Forests	0.7671
LogitBoost	0.7835

Patient level insights

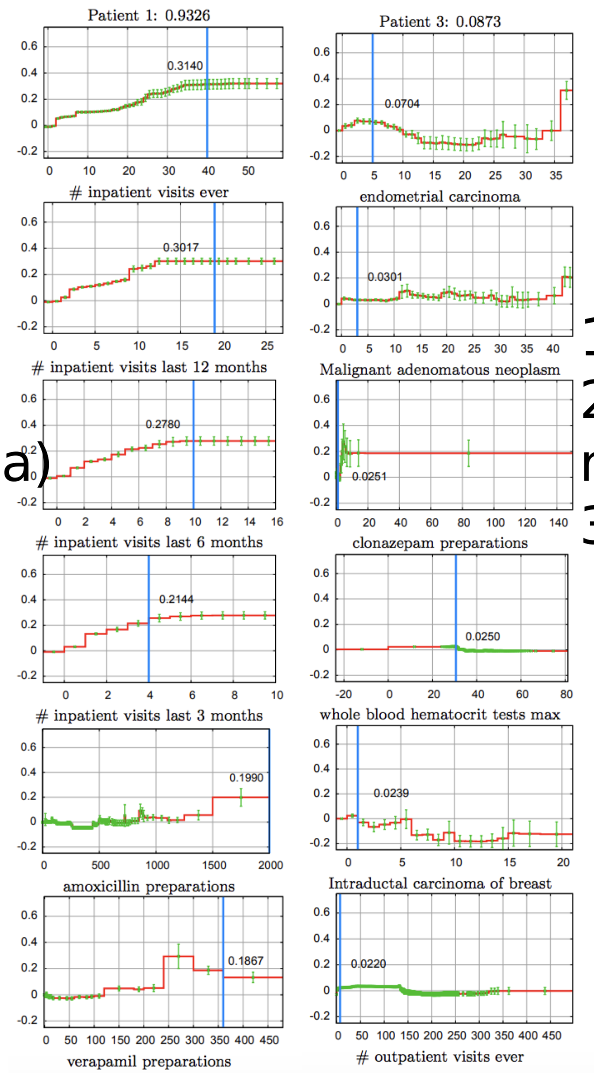
. Lots of admissions
. Received lot of
Amoxycilin (strep/pneumonia)
. Verapamil (hypertension)



p(risk) = 0.9326

Patient level insights

. Lots of admissions
. Received lot of
moxycilin (strep/pneumonia)
. Verapamil (hypertension)



1. Post menopausal
2. Cancers that
respond well to treatme
3. Not hospitalized much

p(risk) = 0.9326

p(risk) = 0.0873

Roadmap

- ~~Motivation~~
- ~~Intelligible Models~~
- ~~Case study: Pneumonia risk~~
- ~~Case study: 30 day readmission~~

Modularity

- Bias term + contributions of individual features + contributions of pairwise interactions

- This structure helps us clearly understand the model

Sorting Terms By Importance

- For each patient, we can compute which term is resulting in what risk score
- Rank terms based on the values of risk score
- This ranking tells us which features are contributing to the risk of each patient

Feature Shaping vs. Expert Discretization

- Instead of GAMs learning function shapes, experts could also provide inputs by discretizing features
- Expert discretized features were used for logistic regression model
- However, GAMs outperformed LR indicating that feature shaping is valuable

Correlation != Causation

- GAMs and GA^2Ms are intelligible
- But, they are not causal
- What we see in plots are associations captures from the data but are not causal implications
- It is often easy to confuse intelligibility of predictive models with causality
  - Please don’t make that mistake!

Prototype Based Approaches

Deep Learning for Case-Based Reasoning through Prototypes

Oscar Li, Hao Liu, Chaofan Chen, Cynthia Rudin

Contributions

- Proposed and developed a novel network architecture for deep learning
- Explains its own reasoning for each prediction
- Not post-hoc explanations
- Prototypes learned during training
  - Explanations are faithful to what the network computes

Motivation

- ML models are increasingly deployed to answer societal questions interpretability/transparency
- Radiology: Lack of transparency poses challenges to FDA approval for deep learning models
- Neural nets are particularly difficult to understand because of the high degree of non-linearity

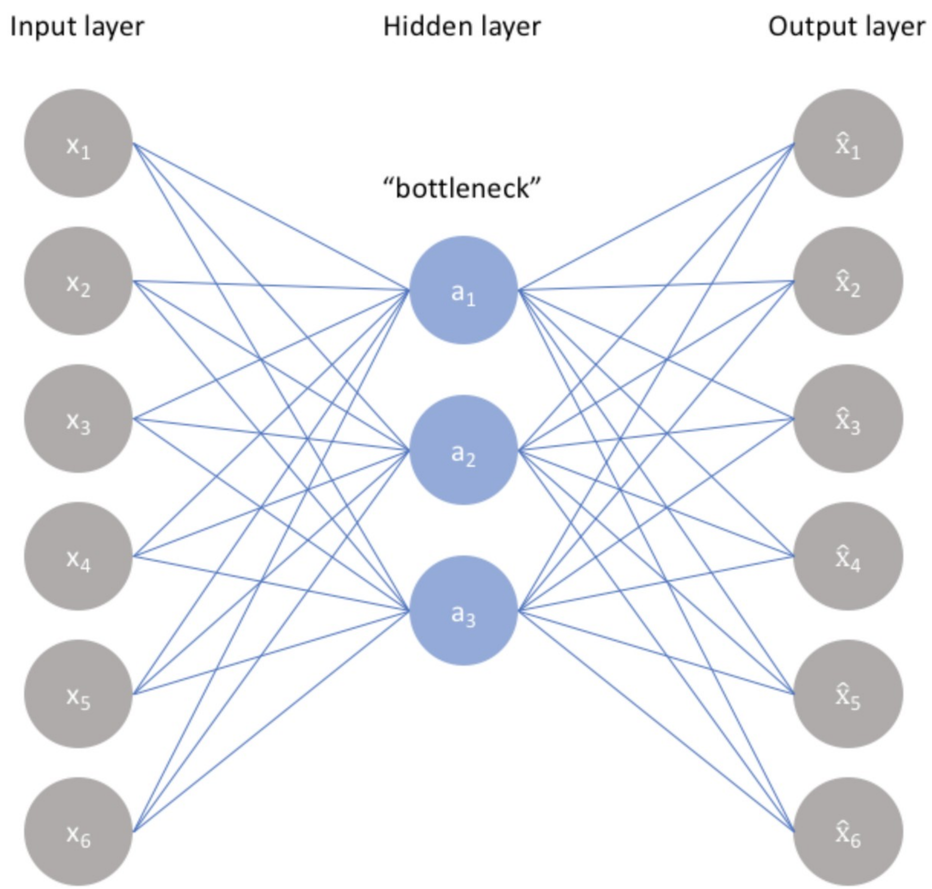
Related Work: Post-hoc explanations

- Past: neural nets designed mainly for accuracy with post-hoc explanations
  - Build neural net first, then interpret!

- Problem: post-hoc explanations may not be faithful to the model

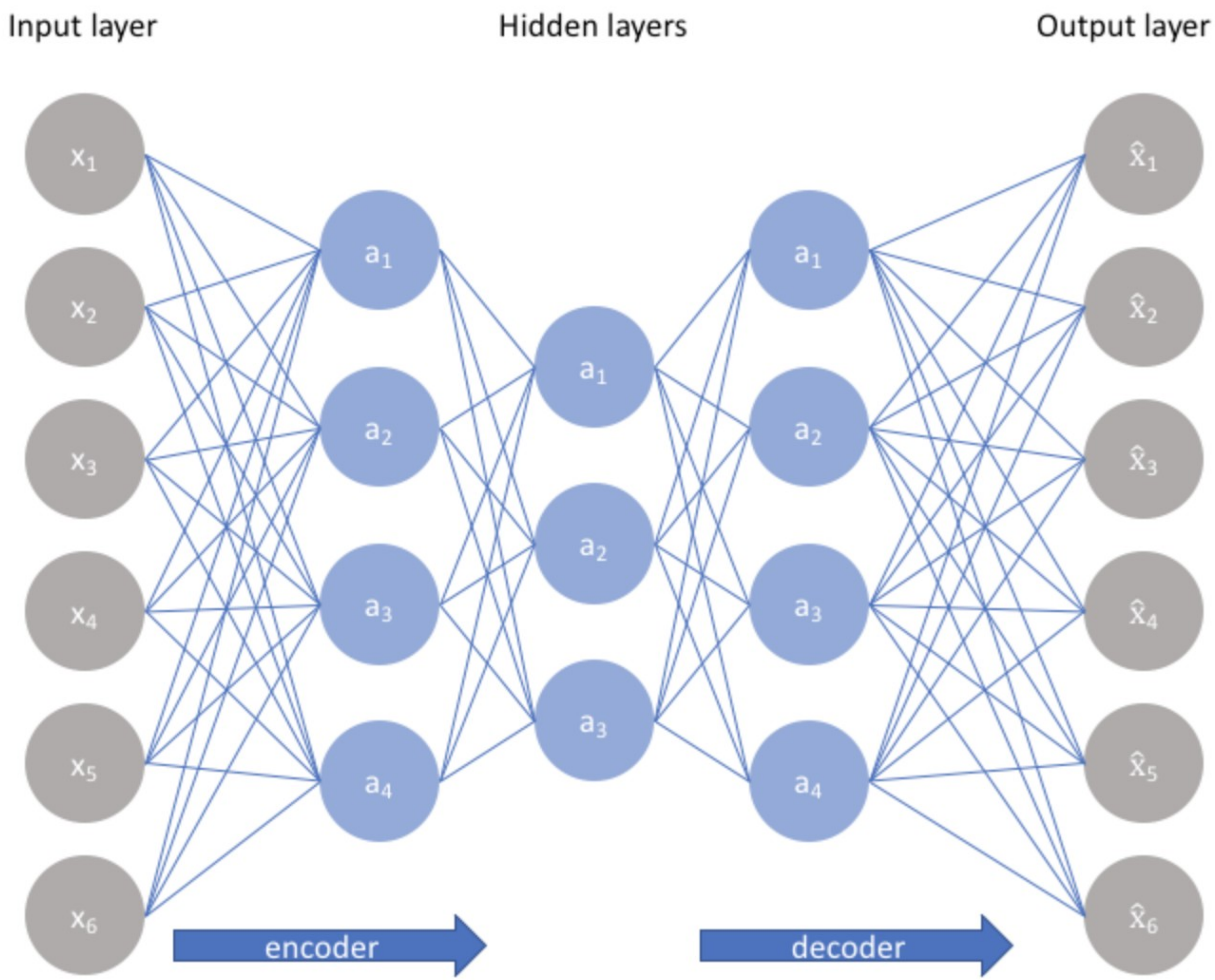
- Easy to create multiple conflicting yet convincing explanations, none of which is correct

Background: Autoencoder



Non-linear Dimensionality Reduction and Reconstruction

Background: Autoencoder



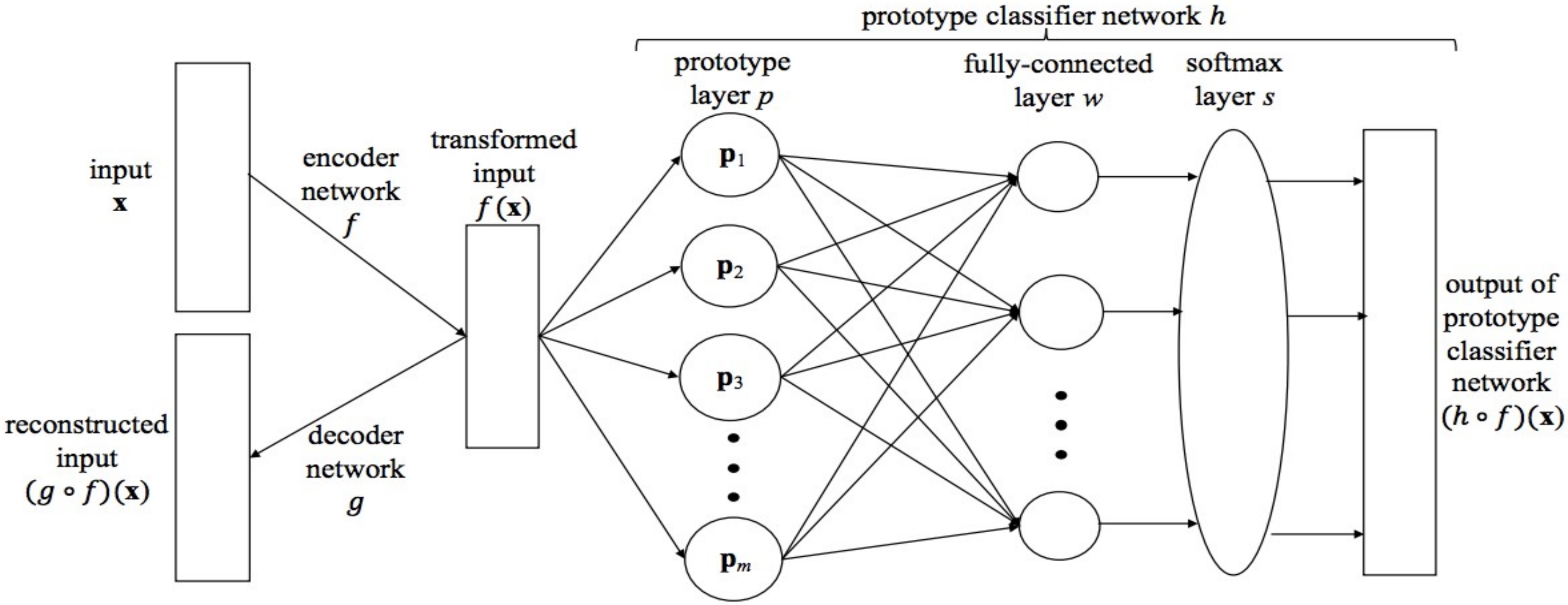
Background: Constructing an Autoencoder

- Constrain the number of nodes present in the hidden layer(s) of the network,
  - limiting the amount of information that can

The encoding will learn and describe latent attributes of the input data.

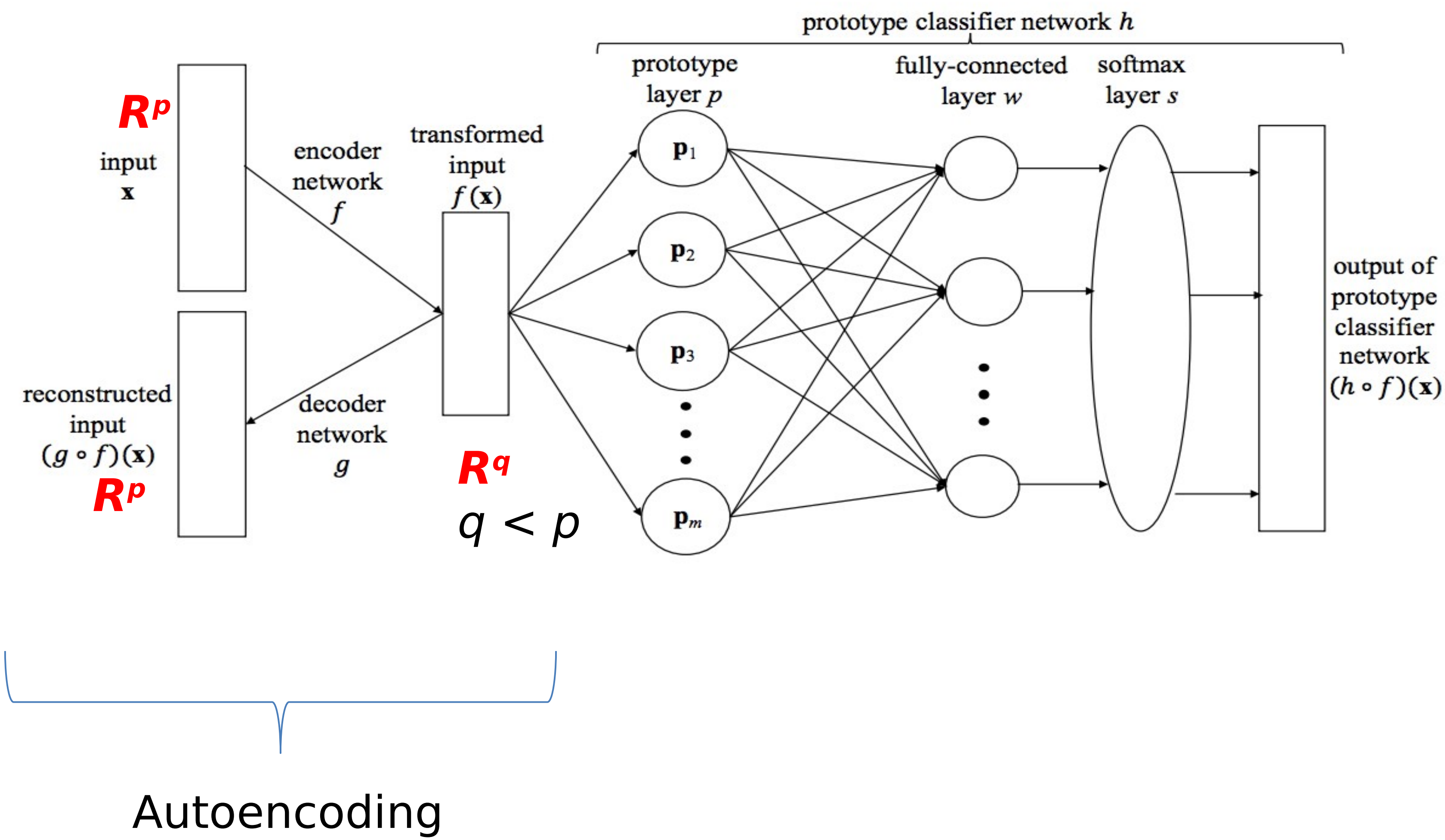
the reconstruction error, our model can learn the most important attributes of the input data and how to best reconstruct the original input from an "encoded" state.

Proposed Network Architecture

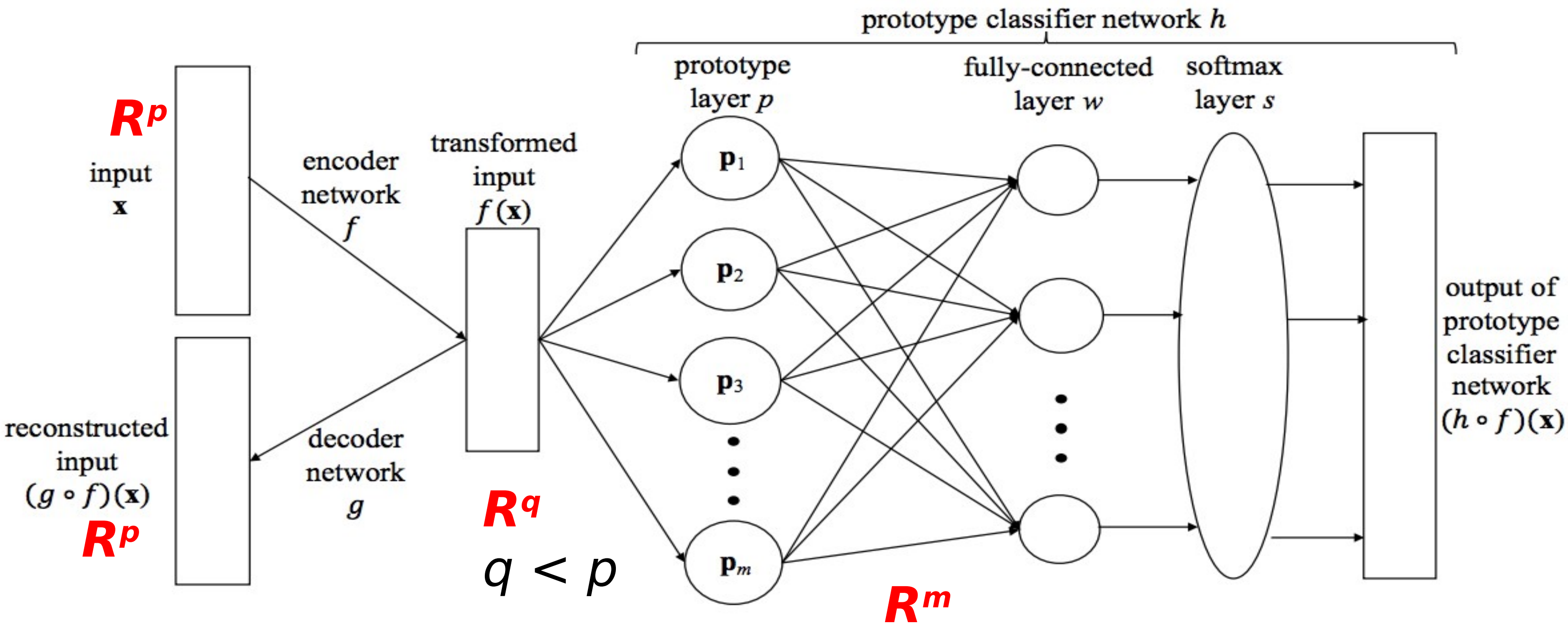


Autoencoding

Autoencoder



Prototype Layer

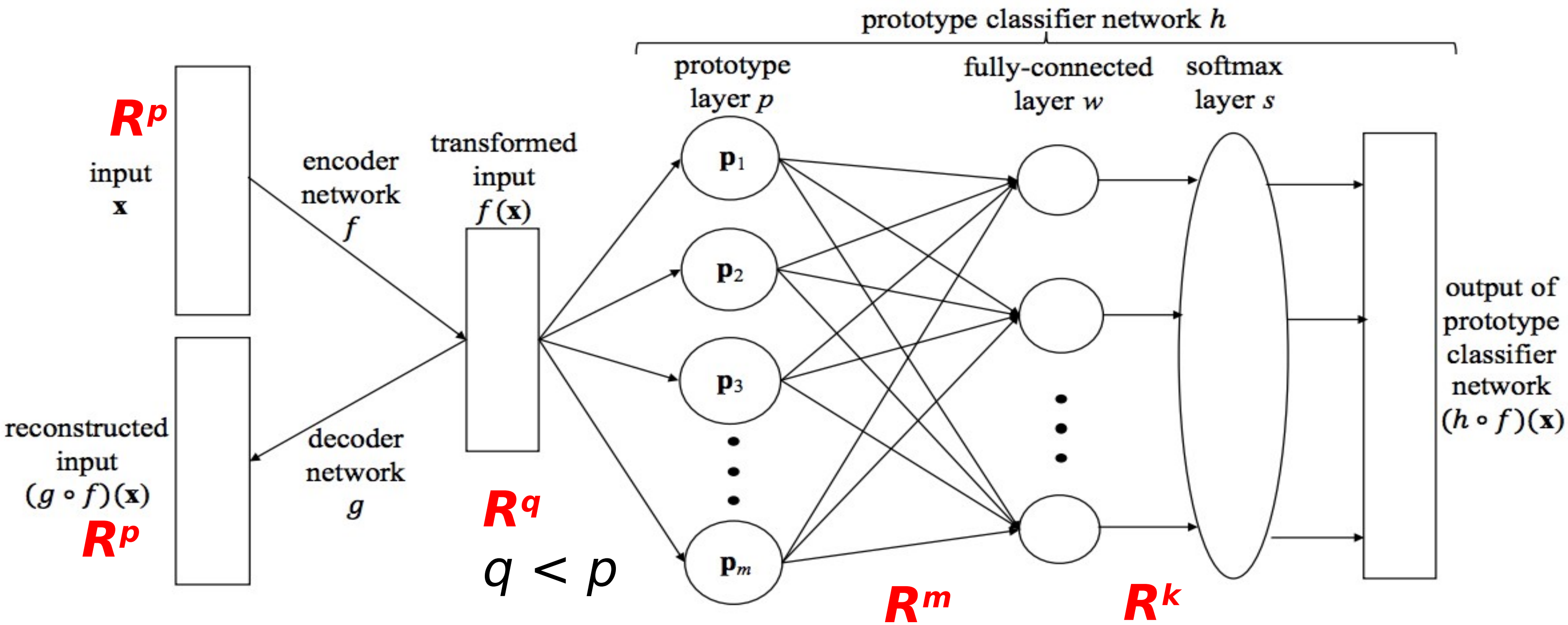


Prototype layer is responsible for computing the following:

$$\mathbf{z} = f(\mathbf{x}) \quad p(\mathbf{z}) = \left[\|\mathbf{z}-\mathbf{p}_1\|_2^2, \quad \|\mathbf{z}-\mathbf{p}_2\|_2^2, \quad \ldots \quad \|\mathbf{z}-\mathbf{p}_m\|_2^2\right]^\top$$

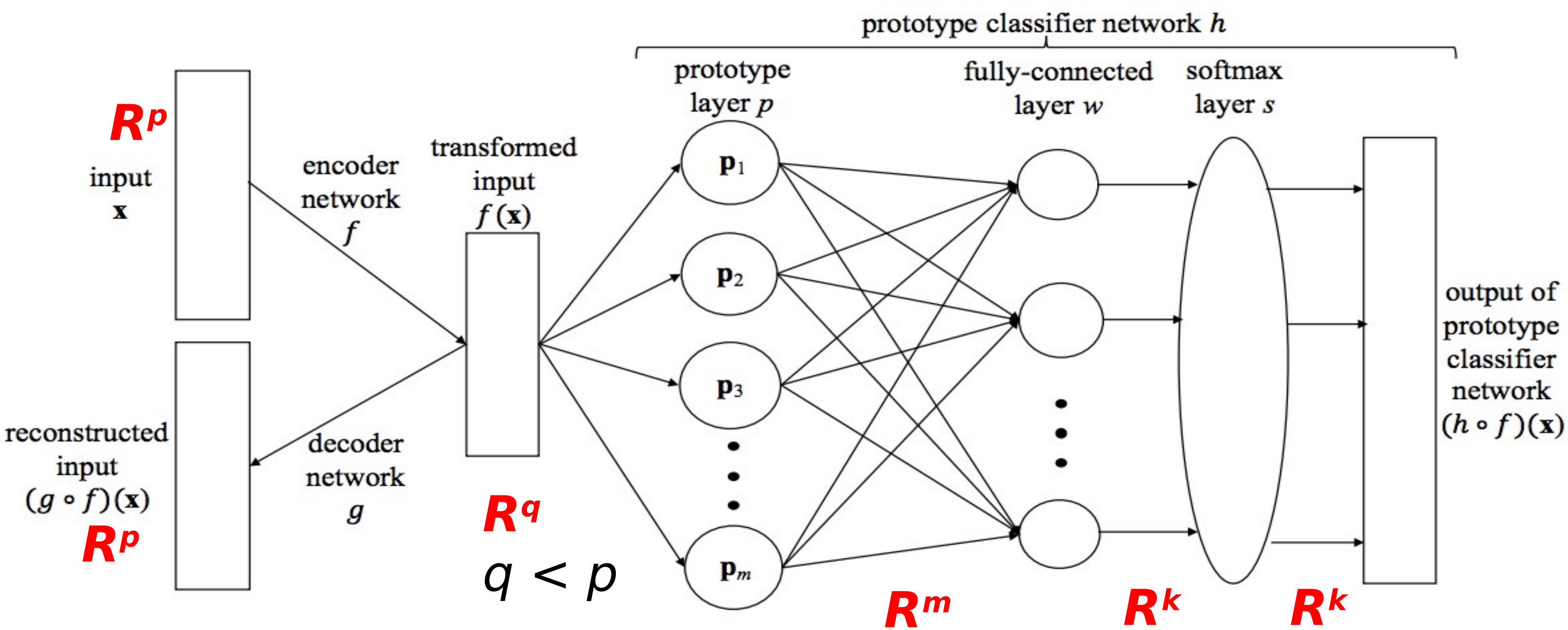
Each node in layer p computes one of the above elements

Fully Connected Layer



- The fully connected layer computes weighted sums of the data $\|\mathbf{z}-\mathbf{p}_j\|_2^2$ $Wp(\mathbf{z})$:
- W is a k x m matrix

Softmax Layer



The weighted sums $Wp(\mathbf{z})$ are normalized by the softmax layer to output a probability distribution over K classes

Advantages of Proposed Architecture

- Automatically learns useful features
  - Non-linear dimensional reduction
  - Suitable for high-dimensional data such as images
- Prototypes vectors can be decoded and visualized
  - Same latent space as encoded inputs
- Ability to interpret without post-hoc analysis

Cost Function

- Cross entropy loss

$$E(h \circ f, D) = \frac{1}{n} \sum_{i=1}^{n} \sum_{k=1}^{K} -\mathbb{1}[y_i = k] \log((h \circ f)_k(\mathbf{x}_i))$$

- Reconstruction error

$$R(g \circ f, D) = \frac{1}{n} \sum_{i=1}^{n} \|(g \circ f)(\mathbf{x}_i) - \mathbf{x}_i\|_2^2.$$

Cost Function: Interpretability Regularizers

$$R_1(\mathbf{p}_1, ..., \mathbf{p}_m, D) = \frac{1}{m}\sum_{j=1}^{m} \min_{i\in[1,n]} \|\mathbf{p}_j - f(\mathbf{x}_i)\|_2^2,$$

$$R_2(\mathbf{p}_1, ..., \mathbf{p}_m, D) = \frac{1}{n}\sum_{i=1}^{n} \min_{j\in[1,m]} \|f(\mathbf{x}_i) - \mathbf{p}_j\|_2^2.$$

Each prototype vector should be as close as possible to at least one training example

Each training example should be as close as possible to one prototype

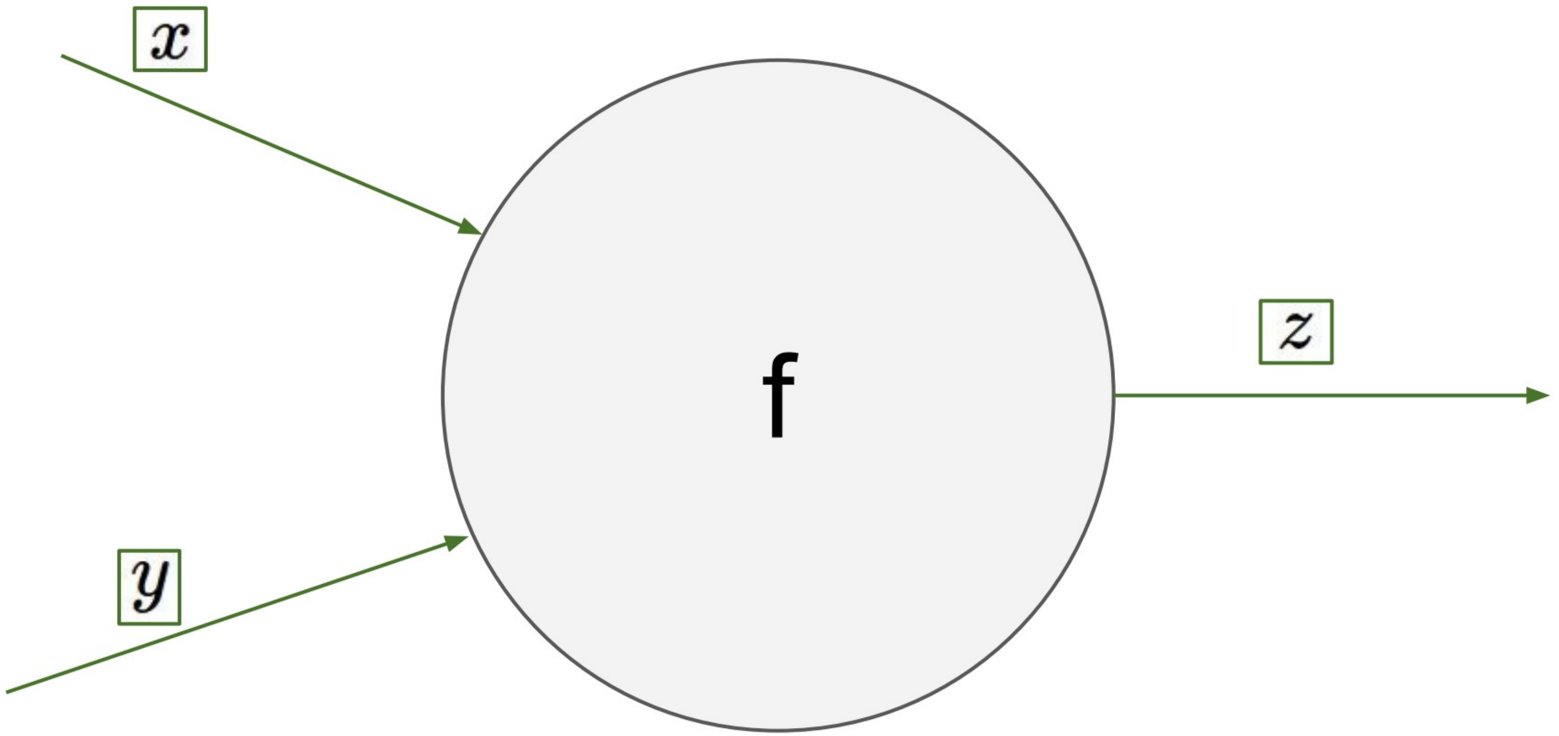
Cost Function

$$
\begin{aligned}
L((f, g, h), D) = E(h \circ f, D) &+ \lambda R(g \circ f, D) \\
&+ \lambda_1 R_1(\mathbf{p}_1, \ldots, \mathbf{p}_m, D) \\
&+ \lambda_2 R_2(\mathbf{p}_1, \ldots, \mathbf{p}_m, D),
\end{aligned}
$$

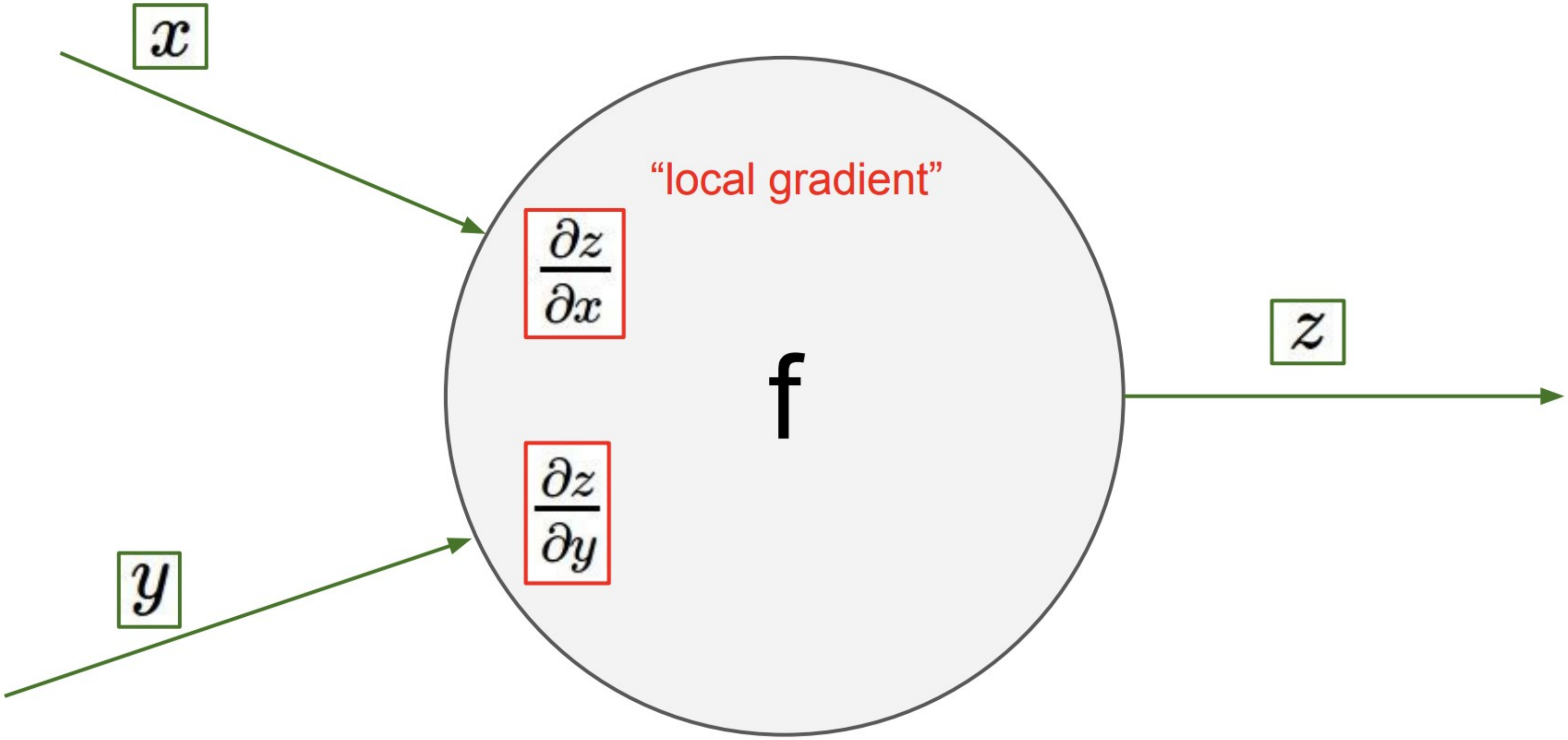
Neural Networks

- Step 1: Define architecture
- Step 2: Outline cost function
- Step 3: Forward pass, compute derivatives, back propagate, update parameters – repeat!
  - Min functions are not technically differentiable
  - But, in practice, packages allow it
  - This is essentially gradient descent

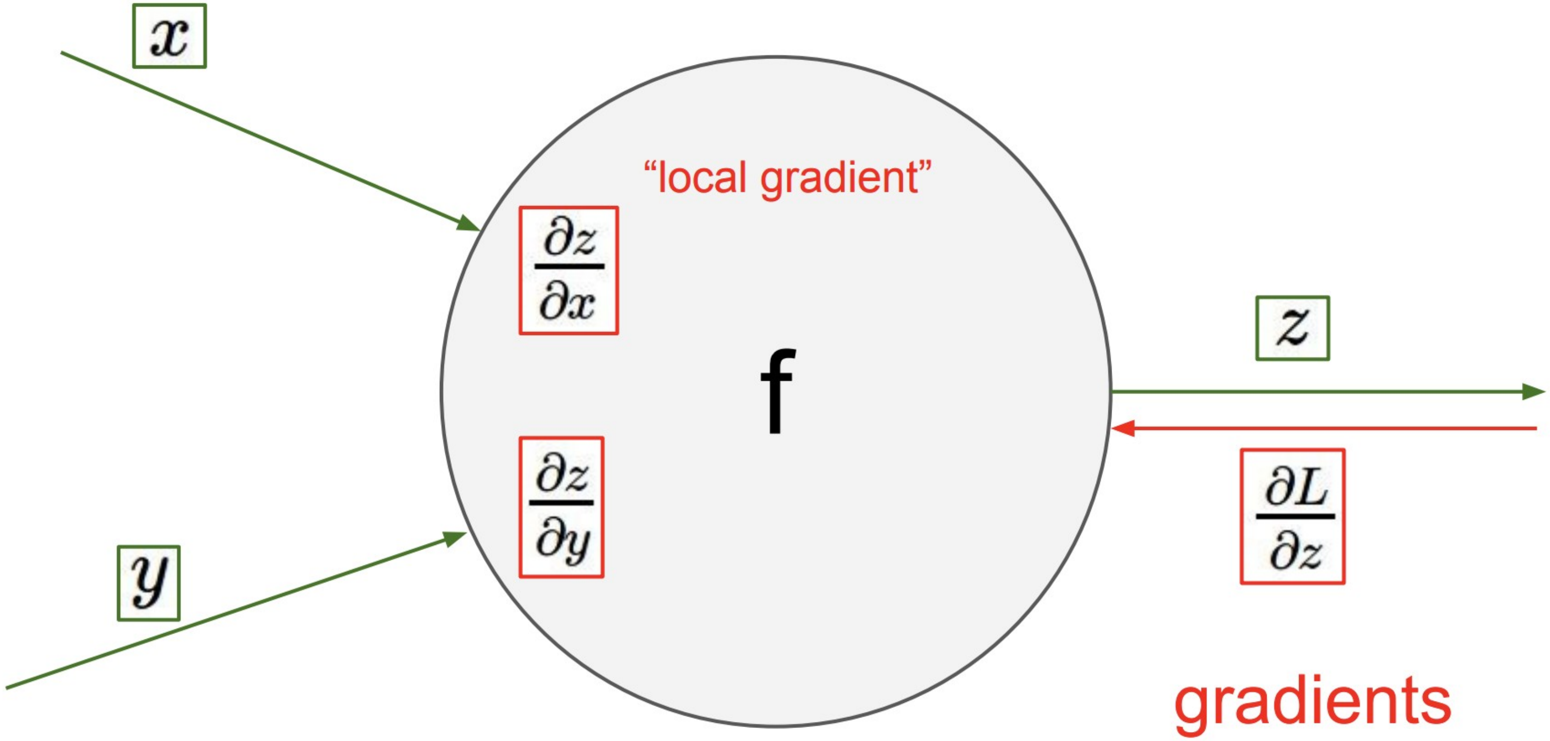
Backpropagation: Intuition



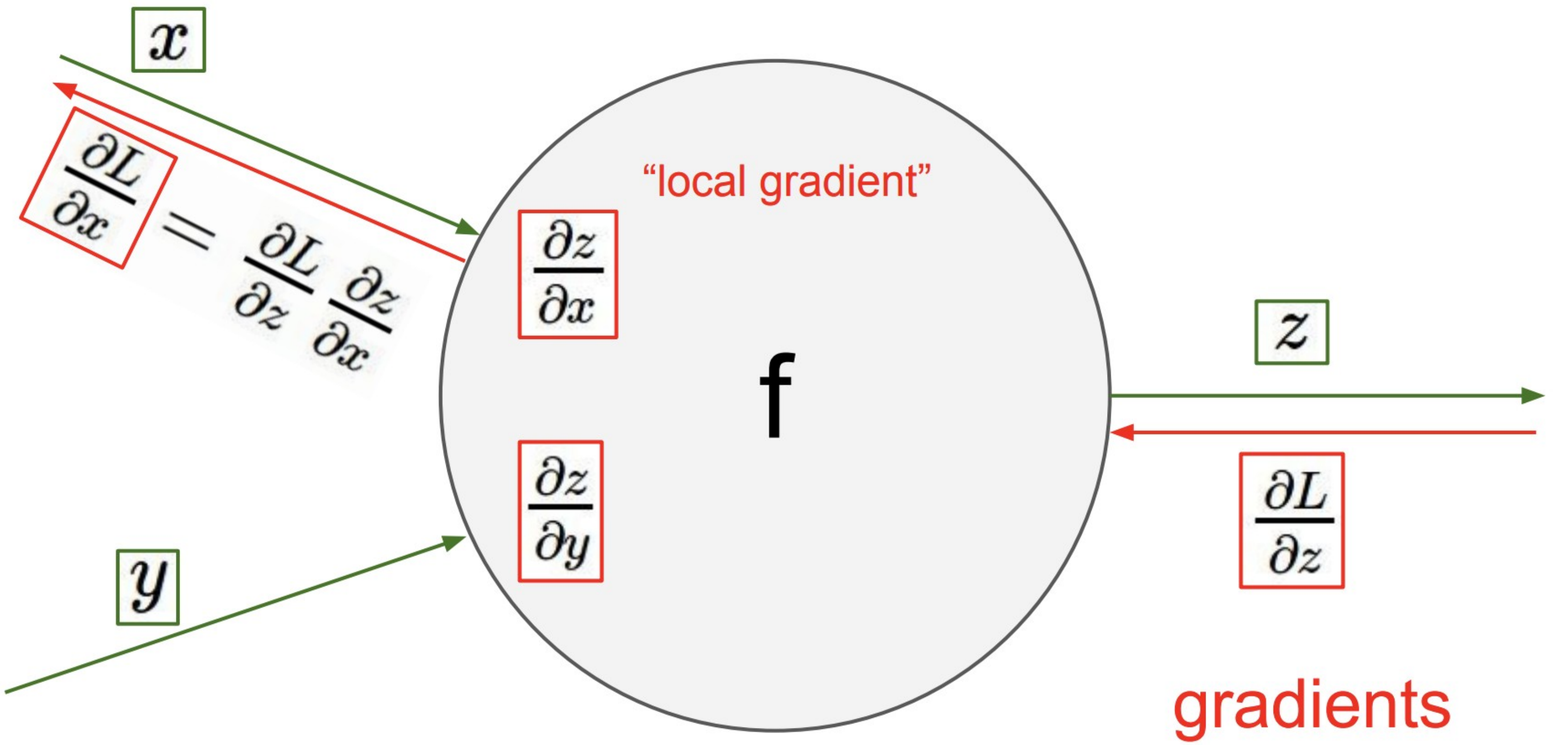
Backpropagation: Intuition



Backpropagation: Intuition

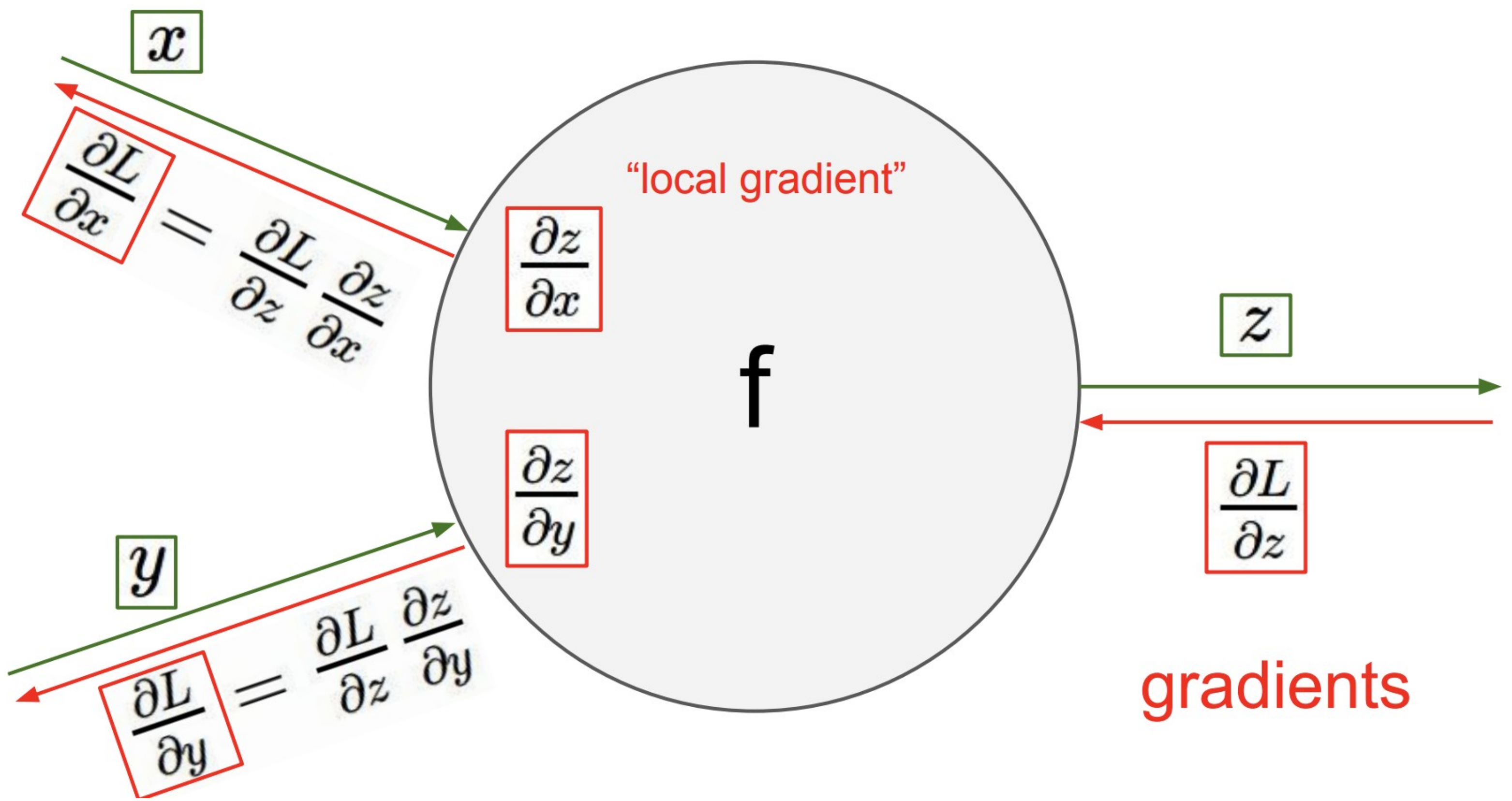


Backpropagation: Intuition



Local gradient x upstream gradient

Backpropagation: Intuition



Local gradient x upstream gradient

MNIST Data

- Test accuracy above 99% and on par with SOTA

- Re


- Lea


Learned Weight Matrix

	0	1	2	3	4	5	6	7	8	9
8	-0.07	7.77	1.81	0.66	4.01	2.08	3.11	4.10	-20.45	-2.34
9	2.84	3.29	1.16	1.80	-1.05	4.36	4.40	-0.71	0.97	-18.10
0	-25.66	4.32	-0.23	6.16	1.60	0.94	1.82	1.56	3.98	-1.77
7	-1.22	1.64	3.64	4.04	0.82	0.16	2.44	-22.36	4.04	1.78
3	2.72	-0.27	-0.49	-12.00	2.25	-3.14	2.49	3.96	5.72	-1.62
6	-5.52	1.42	2.36	1.48	0.16	0.43	-11.12	2.41	1.43	1.25
3	4.77	2.02	2.21	-13.64	3.52	-1.32	3.01	0.18	-0.56	-1.49
1	0.52	-24.16	2.15	2.63	-0.09	2.25	0.71	0.59	3.06	2.00
6	0.56	-1.28	1.83	-0.53	-0.98	-0.97	-10.56	4.27	1.35	4.04
6	-0.18	1.68	0.88	2.60	-0.11	-3.29	-11.20	2.76	0.52	0.75
5	5.98	0.64	4.77	-1.43	3.13	-17.53	1.17	1.08	-2.27	0.78
2	1.53	-5.63	-8.78	0.10	1.56	3.08	0.43	-0.36	1.69	3.49
2	1.71	1.49	-13.31	-0.69	-0.38	4.55	1.72	1.59	3.18	2.19
4	5.06	-0.03	0.96	4.35	-21.75	4.25	1.42	-1.27	1.64	0.78
2	-1.31	-0.62	-2.69	0.96	2.36	2.83	2.76	-4.82	-4.14	4.95

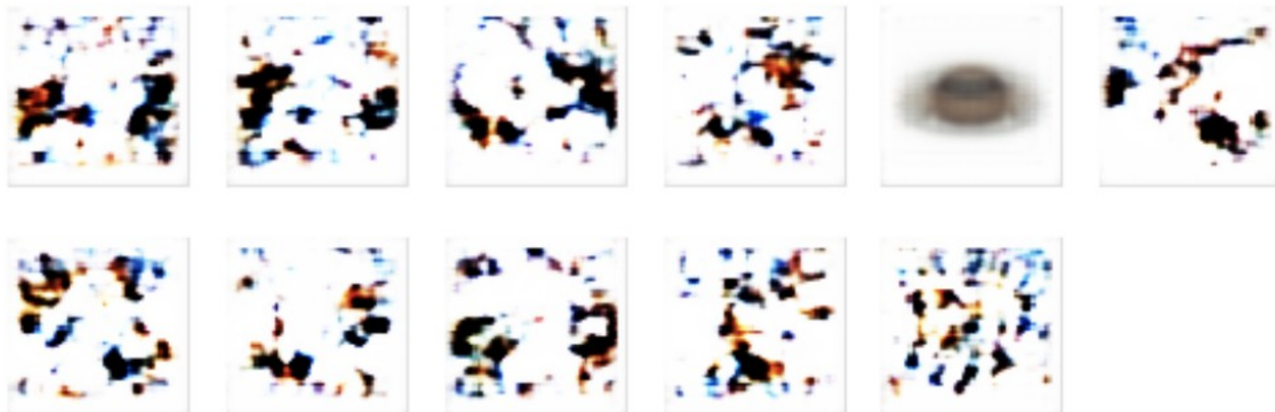
Ablation Study on Cars Data

Learned Prototypes

Without R_1 and R_2

Ablation Study on Cars Data

Without R_1



Without R_2