Big Data Information Extraction:

Generating Seeded Named Entity Dictionary

Mid-program Presentation

Supervisor: Annie En-Shuin Lee, VerticalScope Inc.
Team Members: Gian Alix, Yufeng Li, Mingxuan Wang

Outline

1. Pre-midterm Summary
2. Post-midterm
3. Challenges
4. Conclusions and Final Outcomes

Forum is Unstructured and Noisy Conversations!!

i had to replace the belt, alternator, and the tensioner a few times because of squeaks on my old xj. i don't know why but after a few swaps it got a set it liked and stopped making noise

Yeah, i don't get it. Could this cause any damage? The belt looks fine and the noise developed over the past 2 weeks. It sounds like its coming out of the center pulleys.

I have the same issue. I've replaced the idler pulley, power steering pump and two belts. Still have a slight squeak. Sounds like it's coming from the water pump. Could it be the water pump squeaking? Is this common?

I replaced that with the reman engine. Very odd. I'm just worried i'm doing additional damage. i didn't have a chance to fully check it out this weekend but i'll take a look next weekend.

Just a word of caution. Last august I was on my way home from town 25 mile trip one way. I had the power steering pulley explode and destroy the electric fan blade. Apparently some of the newer ones are made of plastic. Atleast it looks like plastic. I fought mine with the squeaks as well even with a new belt. Dayco makes a tension gauge. I just kept putting tension on till it wouldn't squeal with the a.c. on and revving the engine. Just a final note. I noticed a few days before the pulley broke that there were cracks in the center of the pulley where it meets the hub. Broke before I could replace it.

So not only is this the second belt in a week to squeak but now this is the second one the shred while driving.

OK,. a bit of an update. i think the sound is coming from the water pump pulley. odd way of diagnosing but this morning i spit on the pulley and the sound went away for about 2 minutes then came back, does this mean it's a tension problem and not a pulley problem? I'll try to adjust tomorrow AM. Do I need to keep the e-fan in there if it's winter time in NY?

Motivation

Find Named Entities

1. **Human Annotation**
2. **Seed Catalogue**

Pre-midterm Summary

01

Methodology

- Extract posts
- Pattern
- Extract annotation
- Update seed catalogue

02

Algorithm

- Use of dictionaries
- Brute force matching

03

Evaluation

- Scoring through Test Statistics
- Performance via F1-score

04

Results

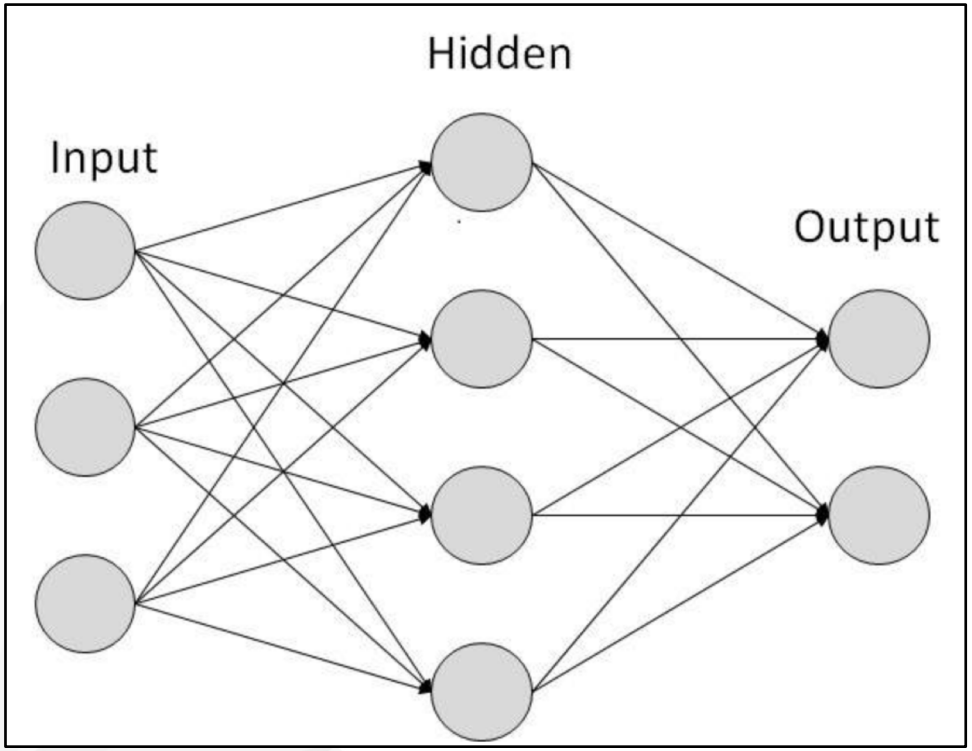
- Approach/Retrieval Techniques
- Population-Selection Methods
- Set Retrieval on a biased selection gives the best F1-score: 0.7222

Post-midterm

1. Deep Learning
2. Mapping Functions
3. Word Embedding
4. Word2vec

Deep Learning

Feedforward Networks



Feedforward

Information flows through the function being evaluated

Network

Composed together of many different functions

Goal

To approximate some function $y = f(x)$
Example: $y = f(x, \theta)$
Best θ that gives the best approximation

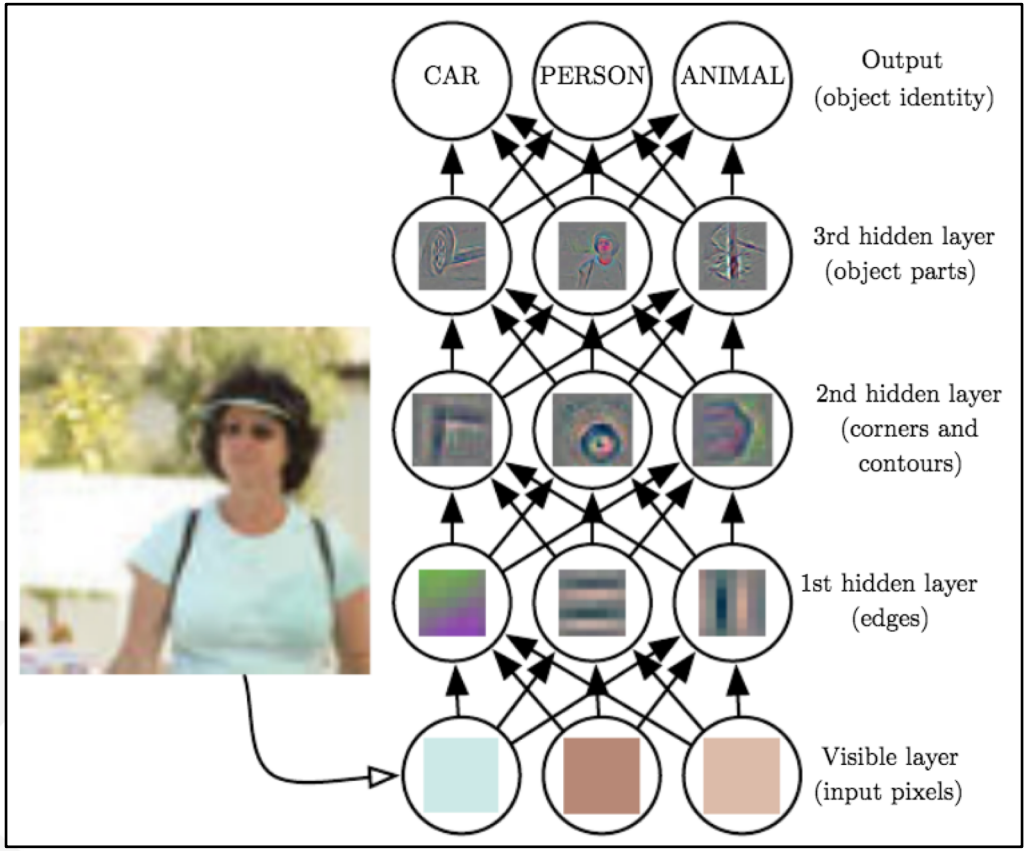
Layers

Three main layers:

- input
- hidden
- output

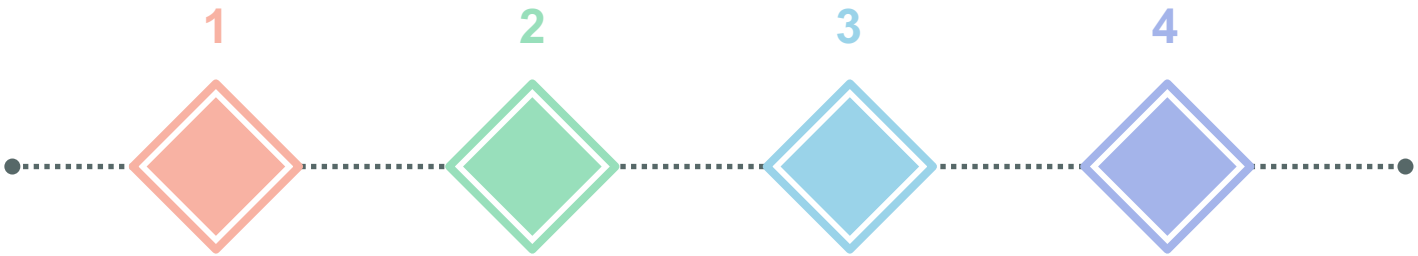
Deep Learning

Visual Example



Deep Learning

Practical Design Process



1. Determine the Goal (error metrics)
2. Establish a working end-to-end pipeline
3. Instrument the system well to determine bottlenecks in performance.
4. Repeatedly make incremental changes
   - Gathering new data
   - Adjusting hyperparameter
   - Debugging

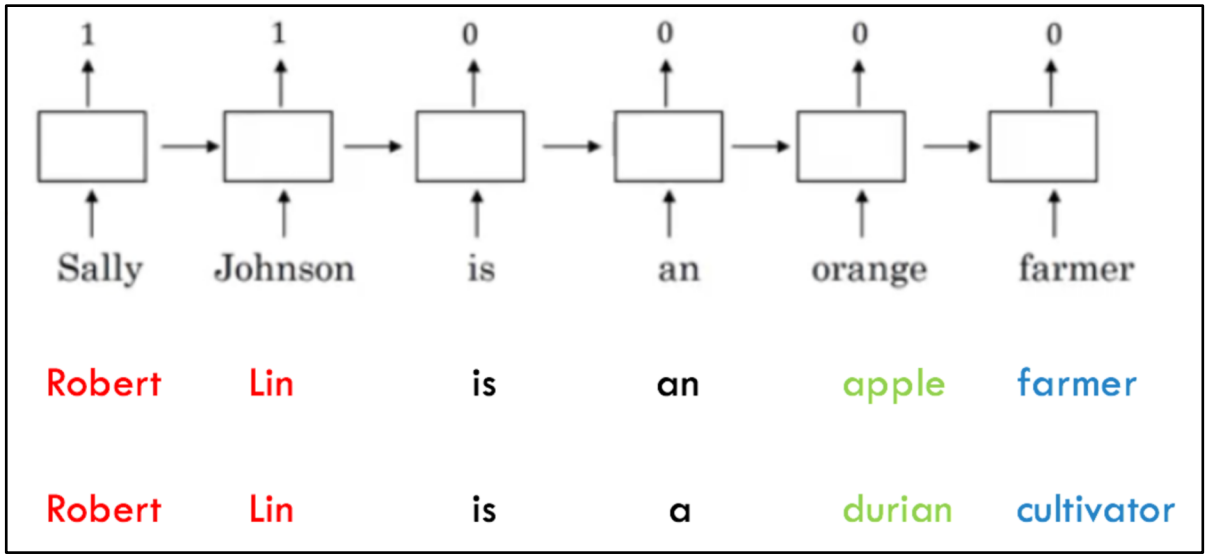
Deep Learning

Word Embedding

1. **One-hot Vector:** Evaluate Similarities
2. **Featurized Representation**
3. **Optimize Embedding Matrix**

Deep Learning

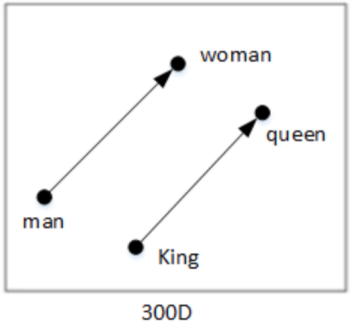
Word Embedding



Deep Learning

Word Embedding

	Man (5391)	Woman (9853)	King (4914)	Queen (7157)	Apple (456)	Orange (6257)
Gender	-1	1	-0.95	0.97	0.00	0.01
Royal	0.01	0.02	0.93	0.95	-0.01	0.00
Age	0.03	0.02	0.7	0.69	0.03	-0.02
Food	0.09	0.01	0.02	0.01	0.95	0.97
⋮	⋮	⋮	⋮	⋮	⋮	⋮



$$e_{man} - e_{woman} \approx e_{king} - e_?$$

$$e_{man} - e_{woman} = \begin{bmatrix} -1 \\ 0.01 \\ 0.03 \\ 0.09 \end{bmatrix} - \begin{bmatrix} 1 \\ 0.02 \\ 0.02 \\ 0.01 \end{bmatrix} \approx \begin{bmatrix} -2 \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

$$e_{king} - e_{queen} = \begin{bmatrix} -0.95 \\ 0.93 \\ 0.70 \\ 0.02 \end{bmatrix} - \begin{bmatrix} 0.97 \\ 0.95 \\ 0.69 \\ 0.01 \end{bmatrix} \approx \begin{bmatrix} -2 \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

Challenges

Human Annotation

- Consistency issues
- Ambiguous NE

Main Textbook

- Too advanced for an undergraduate level

Resources

- Needed a balance of easy and difficult resources

Conclusions

Final Outcomes and Key Accomplishments

1. **Human Annotations**
   Annotated over 1000+ sentences on 10 verticals
2. **Algorithms**
   Wrote scripts to recognize named entities
3. **Experiments and Results**
   Measured our algorithm's performance via the F1-score
4. **Deep Learning**
   Gained more knowledge on Neural Nets and Word Embedding

Thank you