Modeling Difficulty to Understand Deep Learning Performance

John Lalor (1), Hao Wu (2), Hong Yu (1)

(1) University of Massachusetts, (2) Boston College

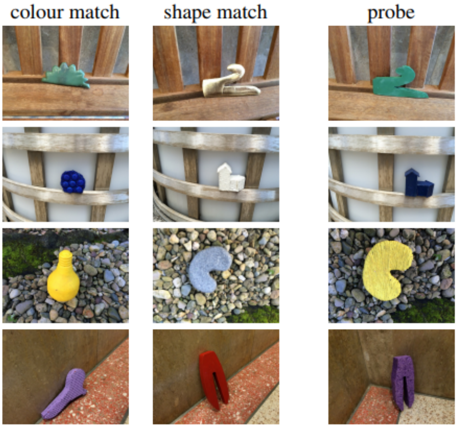
10 January, 2018

Motivation

- What do we know about how Deep Neural Networks (DNNs) learn?
- One way is by looking at specific outputs for which we have detailed information
  - E.g. probing networks with methods from cognitive psychology[1]

[1] Ritter et al. Cognitive psychology for deep neural networks: A shape bias case study, *ICML* 2017.

Motivation



Motivation

- In this work we look at *difficulty* to determine if it has an effect on learning
  - Item Response Theory from Psychometrics[2]
- Do DNNs learn easy examples differently than harder examples?

[2] Baker and Kim. *Item Response Theory: Parameter Estimation Techniques, 2nd Ed.* CRC Press, 2014

IRT Introduction

- Psychometric methodology for scale construction and evaluation
- Jointly model individual ability and item characteristics
- Widely used in educational testing: construction, evaluation, and scoring of standardized tests (e.g. TOEFL, GRE)

IRT Terminology

- “item”: single test question
- “evaluation scale”: set of items to measure some ability
- “response pattern”: an individual’s graded (0/1) answers to an evaluation scale
- “ability score” (θ): Score assigned to an individual based on her response pattern on the evaluation scale

IRT Assumptions

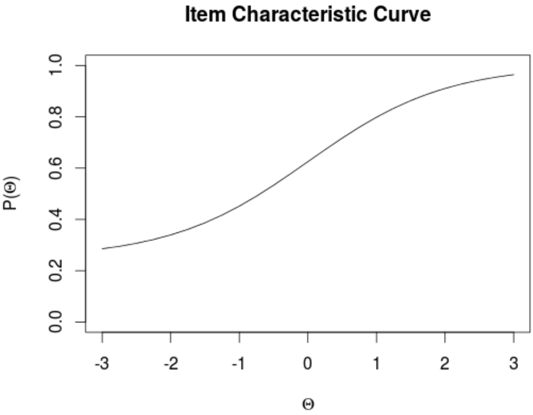
- Individuals differ from each other on an unobserved latent trait dimension (“ability”)
- The probability of correctly answering an item is a function of the person’s ability.
- Responses to different items are independent of each other for a given ability level of the person (“local independence assumption”)
- Responses from different individuals are independent of each other

3 Parameter Logistic Model (3PL)

$$p_{ij}(\theta_j) = c_i + \frac{1 - c_i}{1 + e^{-a_i(\theta_j - b_i)}}$$

- p_{ij}: Probability of individual j answering item i correctly
- θ_j: Ability estimate for individual j
- a_i: Discrimination parameter for item i
- b_i: Difficulty parameter for item i
- c_i: Guessing parameter for item i

Item Characteristic Curve (ICC)



ICC: Varying Parameters

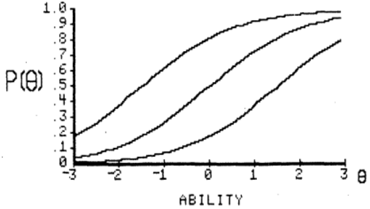


FIGURE 1-2. Three item characteristic curves with the same discrimination but different levels of difficulty

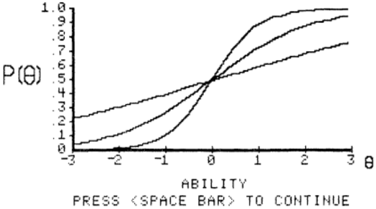


FIGURE 1-3. Three item characteristic curves with the same difficulty but with different levels of discrimination

Estimating Item Paramters

$$p_{ij}(\theta_j) \quad = \quad c_i + \frac{1 - c_i}{1 + e^{-a_i(\theta_j - b_i)}}$$

Estimating Item Paramters

$$\begin{aligned} p_{ij}(\theta_j) &= c_i + \frac{1 - c_i}{1 + e^{-a_i(\theta_j - b_i)}} \\ q_{ij}(\theta_j) &= 1 - p_{ij}(\theta_j) \end{aligned}$$

Estimating Item Paramters

$$
\begin{aligned}
p_{ij}(\theta_j) &= c_i + \frac{1 - c_i}{1 + e^{-a_i(\theta_j - b_i)}} \\
q_{ij}(\theta_j) &= 1 - p_{ij}(\theta_j) \\
L &= \prod_{j=1}^{J} \prod_{i=1}^{I} p_{ij}(\theta_j)^{y_{ij}} q_{ij}(\theta_j)^{1-y_{ij}}
\end{aligned}
$$

Estimating Item Paramters

$$
\begin{aligned}
p_{ij}(\theta_j) &= c_i + \frac{1 - c_i}{1 + e^{-a_i(\theta_j - b_i)}} \\
q_{ij}(\theta_j) &= 1 - p_{ij}(\theta_j) \\
L &= \prod_{j=1}^{J} \prod_{i=1}^{I} p_{ij}(\theta_j)^{y_{ij}} q_{ij}(\theta_j)^{1-y_{ij}} \\
\log L &= \sum_{j=1}^{J} \sum_{i=1}^{I} [y_{ij} p_{ij}(\theta_j) + (1 - y_{ij}) q_{ij}(\theta_j)]
\end{aligned}
$$

Solve with Expectation-Maximization[3]

[3] Bock and Aitkin. Marginal Maximum Likelihood Estimation of Item Parameters: Application of an EM Algorithm. *Psychometrica*, 1981.

Using Difficulty as a Predictor

- Can we use what we know about certain examples (via IRT) to understand performance?
- Train models with different training set sizes to see how performance changes with regards to training size and item difficulty

Data

- Stanford SNLI Corpus[4]
  - 550K English premise-hypothesis sentence pairs
    - All human generated
  - IRT Data: 180 randomly sampled sentence pairs from *quality-control* data[5]
    - originally annotated by 5 Amazon Mechanical Turk workers (Turkers)
    - We obtained an additional 1000 labels per sentence pair

[4] Bowman et al. A large annotated corpus for learning natural language inference. *EMNLP*, 2015.

[5] Lalor et al. Building an evaluation scale using item response theory. *EMNLP* 2016.

DNN Models

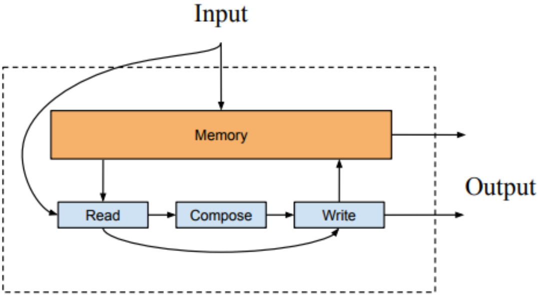
- Long-Short Term Memory (LSTM) Network [6]
- Convolutional Neural Network [7]
- Memory augmented LSTM: Neural Semantic Encoder[8]

[6] Hochreiter and Schmidhuber. Long short-term memory. *Neural computation*, 1997.

[7] Kim. Convolutional neural networks for sentence classification. arXiv:1408.5882, 2014.

[8] Munkhdalai and Yu. Neural semantic encoders. *EACL*, 2017

Neural Semantic Encoder

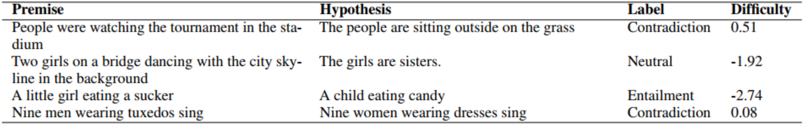


Parameter Estimation

- Collect AMT data
- Fit item parameters
  - *mirt* R package[9]
- Use IRT Data as test set for pre-trained DNN models
  - Models trained on random sample of SNLI training data
- Predict likelihood of model answering correctly with difficulty and training set size as dependent variables

[9] Chalmers. mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software*, 2012.

Examples



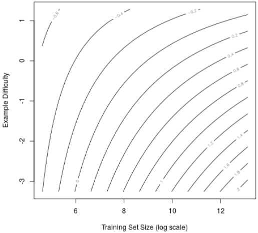
Premise	Hypothesis	Label	Difficulty
People were watching the tournament in the stadium	The people are sitting outside on the grass	Contradiction	0.51
Two girls on a bridge dancing with the city skyline in the background	The girls are sisters.	Neutral	-1.92
A little girl eating a sucker	A child eating candy	Entailment	-2.74
Nine men wearing tuxedos sing	Nine women wearing dresses sing	Contradiction	0.08

Overview

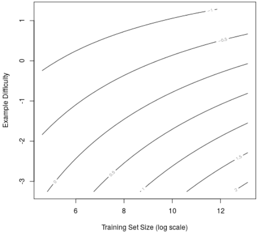
- Goal was to determine whether *difficulty* has an impact on performance
- Logistic regression with training size and difficulty as predictors
- Plot log-odds of answering a question correctly

Contour Plots

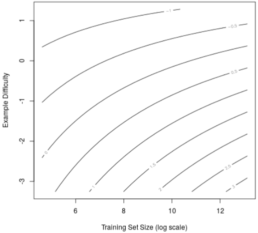
LSTM



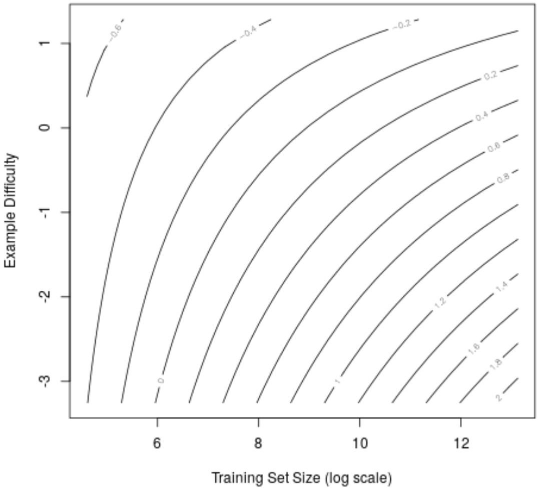
CNN



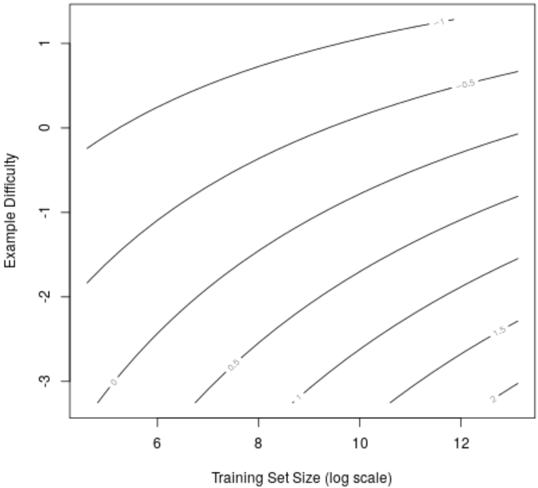
NSE



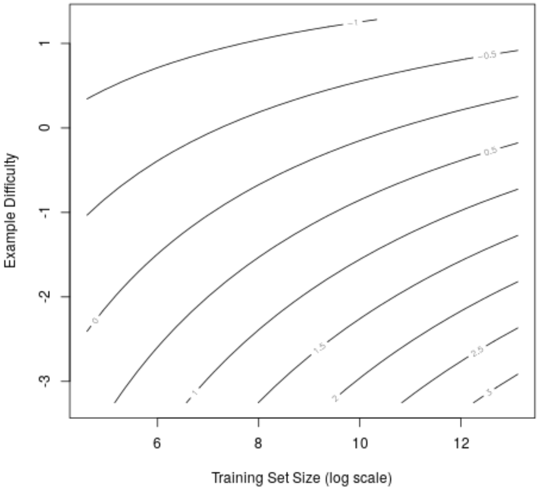
LSTM


-0.6
-0.4
-0.2
0.2
0.4
0.6
0.8
0
1
1.2
1.4
1.6
1.8
2
1
0
-1
-2
-3
Example Difficulty
6
8
10
12
Training Set Size (log scale)

CNN


-1
-0.5
0
0.5
1
1.5
2
1
0
-1
-2
-3
Example Difficulty
6
8
10
12
Training Set Size (log scale)

NSE


-1
-0.5
0.5
0
1
1.5
2
2.5
3
1
0
-1
-2
-3
Example Difficulty
6
8
10
12
Training Set Size (log scale)

Conclusion

- New use case for IRT in DNNs
  - Difficulty as measured by human responses impacts performance
- Models distinguish between difficulty levels as training size increases
- As they learn more, the get better at easier questions faster
  - This is similar to expected *performance* of humans

Thank you!

email: lalor@cs.umass.edu