Deep Learning and Natural Language Processing: A Review and Outlook

Hang Li
Bytedance AI Lab
Two Questions

• Why does Deep Learning work very well for Natural Language Processing?
• What will be next beyond Deep Learning for Natural Language Processing?

• This talk tries to answer the questions
Outline

• *Human Language Processing*
• Deep Learning
• Deep Learning for Natural Language Processing
• Attention: Soft Association Mechanism
• Future of Natural Language Processing
• Summary
Damasio’s Hypothesis

Having a mind means that an organism forms neural representations which can become images, be manipulated in a process called thought.

Antonio Damasio
Embodied Simulation Hypothesis

• Question: “Does a gorilla have a nose?”
• To answer this question, one evokes the image of gorilla in consciousness
• Concepts are stored in memory as associated visual, auditory, motor images
• Language understanding is simulation on the basis of images of related concepts
Thinking is neural processing in sub-consciousness that generates images in consciousness.
Symbols for Humans

- Words and symbols are based on topographically organized representations and are images as well.
- Mathematicians and physicists describe their thinking as dominated by images.
Outline

• Human Language Processing
• *Deep Learning*
• Deep Learning for Natural Language Processing
• Attention: Soft Association Mechanism
• Future of Natural Language Processing
• Summary
Advantages

• Function Approximation
• Sample Efficiency
• Generalization

Disadvantages

• Robustness
• “Appropriateness”
Function Approximation

- Universal function approximation theorem:
  - For continuous function $F: [0, 1]^n \rightarrow \mathcal{R}$ and $\varepsilon > 0$, there exists

  $$f(x) = \alpha^T \sigma(wx + b)$$

  $$= \sum_i \alpha_i \sigma \left[ \sum_j w_{ij} x_i + b_i \right]$$

  such that for all $x$ in $[0,1]^n$, $|F(x) - f(x)| < \varepsilon$ holds

  Cybenko 1989
Sample Efficiency

• Theorem:
  – There exist Boolean functions computable with a polynomial size logic gates circuit of depth $k$ that require exponential size when restricted to depth $k - 1$

• Deep networks have better sample efficiency than shallow networks

Hastad 1986
Generalization

• Deep neural networks exhibit remarkably generalization ability

• Findings
  – Deep neural networks easily fit random labels
  – Explicit regularizers like dropout and weight-decay may not be essential for generalization
  – SGD may act as an implicit regularizer

Zhang et al. 2017
Generalization

• Theorem:
  – Two-layer overparameterized ReLU neural network for multi-class classification
  – Stochastic gradient descent (SGD) from random initialization
  – If data is from mixtures of well-separated distributions
  – Then SGD learns a network with small generalization error

Li & Liang 2018
No Free Lunch Theorem

• Theorem
  – \( \mathcal{F} = \text{set of all possible functions, } y = f(x) \)
  – Given any distribution \( \mathcal{D} \) on \((x, y)\) and training data set \( \mathcal{S} \)
  – For any learner \( L \), \( \frac{1}{|\mathcal{F}|} \sum_{\mathcal{F}} Acc(L) = \frac{1}{2} \) holds, where \( Acc \) is generalization accuracy

• Corollary
  – For any two learners \( L_1, L_2 \)
  – If \exists \text{ function, s.t. } Acc(L_1) > Acc(L_2)
  – Then \exists \text{ function, s.t. } Acc(L_2) > Acc(L_1)

Wolpert & Macready 1997
Robustness (强健性)

- Adversarial robustness

\[ \min_{\theta} E_x \left[ \max_{\|x - x'\|_\infty \leq \epsilon} L(\theta, x') \right] \]

- Theorem:
  - If data distribution of binary classification is two Gaussians
  - Then sample complexity of robust generalization is significantly larger than that of standard generalization

- More training data is needed for robust classification

Schmidt et al. 2018
Bengio’s Comment

What can we conclude about the failure of our current systems? I would say the strongest thing I see is that they are learning in a way that exploits superficial clues that help to do the task they are asked to do. But often these are not the clues humans would consider to be the most important.

Yoshua Bengio
Appropriateness

• Learned representation might not be appropriate, due to
  – Data bias
  – Model bias
  – Training bias

• Deep networks may exhibit pathological behavior
Interpretability

• Should neural networks have interpretability?
• It depends on applications, e.g., health care, finance
• We are not consciously aware how our minds process information
Outline

• Human Language Processing
• Deep Learning
• *Deep Learning for Natural Language Processing*
• Attention: Soft Association Mechanism
• Future of Natural Language Processing
• Summary
Natural Language Processing Problems

- Classification: $x \rightarrow c$
- Matching: $x, y \rightarrow \mathcal{R}$
- Sequence-to-Sequence: $x \rightarrow y$
- Structured Prediction: $x \rightarrow [x]$
- Sequential Decision Process: $\pi: s \rightarrow a$

Li 2017
Natural Language Problems

- Classification
  - Text classification
  - Sentiment analysis
- Matching
  - Search
  - Question answering
  - Single-turn dialogue (retrieval)
- Sequence to Sequence
  - Machine translation
  - Summarization
  - Single-turn dialogue (generation)
- Structured Prediction
  - Sequential labeling
  - Semantic parsing
- Sequential Decision Process
  - Multi turn dialogue
Deep Learning for Natural Language Processing

A game of mimicking human behaviors using neural processing tools

\[ y = f(x) \]

\[ \max_f P_f (y|x) \]
Neural Processing Techniques

• Models
  – Feedforward Neural Network
  – Convolutional Neural Network
  – Recurrent Neural Network
  – Sequence-to-Sequence Model
  – Attention
  – …..

• Input: word embedding
• Output: softmax function
• Loss function: cross entropy
• Learning algorithm: stochastic gradient descent
• Regularization, e.g., dropout
McAllester’s Argument

Progress toward AI is coming from advances in general purpose (differentiable) programming language features, including residual connections, gating, attention, GAN, VAE.

David McAllester
Outline

• Human Language Processing
• Deep Learning
• Deep Learning for Natural Language Processing
• Attention: Soft Association Mechanism
• Future of Natural Language Processing
• Summary
Attention Is All You Need?

- Attention (including self-attention) is a powerful mechanism, used in many models including Transformer.
- Attention = ‘soft’ association mechanism, cf., associative memory.

\[
\text{query}
\]

\[
\text{value} = \sum_{i=1}^{n} \pi(\text{query}, \text{key}_i) \cdot \text{value}_i
\]

\[
\begin{array}{|c|c|}
\hline
\text{key}_1 & \text{value}_1 \\
\hline
\text{key}_2 & \text{value}_2 \\
\hline
\text{...} & \text{...} \\
\hline
\text{key}_n & \text{value}_n \\
\hline
\end{array}
\]
Sequence-to-Sequence Model: Transformer

- Encoder + decoder
- Multi-head attention
- Multi-layer encoder and decoder
- Three types of attention
- Parallel processing
- Position embedding
- BERT: Transformer encoder
Transformer Builds Hierarchical Sentence Representation with Attention.
Outline

• Human Language Processing
• Deep Learning
• Deep Learning for Natural Language Processing
• Attention: Soft Association Mechanism
• *Future of Natural Language Processing*
• Summary
New Opportunities in Future

- Language Generation
- Multimodal Processing
- Prior Representation Learning
- Neural symbolic processing?
Prior Representation Learning

- BERT: prior language representation learning
- Enhanced state-of-the-art models in many tasks

\[ y = f(x, h(x')) \]
Neural Symbolic Processing

- Incorporate knowledge (structured symbols) into neural networks
- Still very challenging

\[ y = f(x, g(s)) \]
Hinton’s Comment

• Combining neural processing and symbolic processing is just like combining electric cars and gasoline cars (paraphrase)
• Personal communication, NeurIPS 2018

Geoffrey Hinton
Interpretability

- Extract knowledge (structured symbols) from neural networks
- Another neural symbolic processing problem

\[ y = f(x) \]
Outline

• Human Language Processing
• Deep Learning
• Deep Learning for Natural Language Processing
• Attention: Soft Association Mechanism
• Future of Natural Language Processing
• Summary
Summary

• Human language processing is neural processing
• Advantages of DL: function approximation, sample efficiency, generalization
• Disadvantages of DL: robustness and appropriateness
• DL for NLP is game of mimicking human behaviors using neural processing tools
• Attention: soft association mechanism
• Future directions include generation, multimodality, and prior representation learning
• Neural symbolic processing is a challenging yet important problem
References

• 李航, 智能与计算, 计算机学会通讯, 2019年第一期。
• Yuanzhi Li, Yingyu Liang, Learning Overparameterized Neural Networks via Stochastic Gradient Descent on Structured Data, 2018.
• Ludwig Schmidt, Shibani Santurkar, Dimitris Tsipras, Kunal Talwar, Aleksander Madry, Adversarially Robust Generalization Requires More Data, 2018.
• David McAllester, Universality in Deep Learning and Models of Computation, Second Workshop on Symbolic-Neural Learning, Nagoya, 2018.
Thank you!

lihang.lh@bytedance.com