Research on Deep Learning for Natural Language Processing at Huawei Noah’s Ark Lab

Hang Li
Noah’s Ark Lab
Huawei Technologies
Talk Outline

• Research at Huawei Noah’s Ark Lab
• New Breakthrough in Natural Language Processing with Deep Learning
• Research on DL for NLP at Noah’s Ark Lab
• Summary
Noah’s Ark Lab

- Our future challenge is not about better chips, larger bandwidth, and more complex signal interference models,
  - *It is about data*

- Have you seen the movie ‘2012’?
  - The flood is coming, but this time it is the ‘data flood’
  - Once here, it will never recede

- Noah’s Ark Lab will lead us to tackle the challenges in
  - Data Mining
  - Artificial Intelligence

Founder and President: Mr. Ren, Zhengfei
Noah’s Ark Lab

- Founded in 2012
- Located in Hong Kong and Shenzhen
- Collaborations with Universities in Hong Kong
  - A large number of researchers earned PhD at the universities
  - Many interns
  - Strong connections with professors and students

- Research Areas
  - Machine Learning
  - Data Mining
  - Speech and Language Processing
  - Information and Knowledge Management
  - Intelligent Systems
  - Human Computer Interaction
We want to build

Intelligent Mobile Devices

Data Mining & Artificial Intelligence

Intelligent Telecommunication Networks

Intelligent Enterprise
Intelligent Telecommunication Networks

- Software-defined Network
- Network Maintenance
- Network Planning and Optimization
Intelligent Enterprise

- Supply Chain Management
- Customer Relationship Management
- Human Resources Management
- Information and Knowledge Management
- Communication
Intelligent Mobile Devices

- Information Recommendation
- Personal Information Management
- Natural Language Dialogue (Question Answering)
- Information Extraction
- Machine Translation
Research Topics We Are Working on

• Machine Translation
• Natural Language Dialogue
• Automatic Speech Recognition
• Image Retrieval
• Deep Learning Platform
Talk Outline

• Research at Huawei Noah’s Ark Lab
• New Breakthrough in Natural Language Processing with Deep Learning
• Research on DL for NLP at Noah’s Ark Lab
• Summary
Distributional Linguistics

• Distributional semantics
• *You shall know a word by the company it keeps* - John Firth

• Sylvan?
• Companying words: woods, forest, tree, living, woodland, beautiful, grove
Representation of Word Meaning

Using high-dimensional real-valued vectors to represent the meaning of words
Representation of Sentence Meaning

*New finding: This is possible*

Using high-dimensional real-valued vectors to represent the meaning of sentences.
Recent Breakthrough in Distributional Linguistics

• From words to sentences
• Representing syntax, semantics, even pragmatics
How Is Learning of Sentence Meaning Possible?

- Deep neural networks (complicated non-linear models)
- Big Data
- Task-oriented
- Error-driven and gradient-based
- Compositional
Natural Language Tasks

• Classification: assigning a label to a string
  \[ S \rightarrow C \]

• Generation: creating a string
  \[ \rightarrow S \]

• Matching: matching two strings
  \[ s, t \rightarrow R^+ \]

• Translation: transforming one string to another
  \[ s \rightarrow t \]
Natural Language Applications Can Be Formalized as Tasks

- **Classification**
  - Sentiment analysis
- **Generation**
  - Language modeling
- **Matching**
  - Search
  - Question answering
- **Translation**
  - Machine translation
  - Natural language dialogue (single turn)
  - Text summarization
  - Paraphrasing
Learning of Representations in Tasks

- Classification
  \[ \text{S} \rightarrow \text{r} \rightarrow \text{C} \]

- Generation
  \[ \rightarrow s(r) \]

- Matching
  \[ s, t \rightarrow \text{r} \rightarrow \mathbb{R}^+ \]

- Translation
  \[ \text{S} \rightarrow \text{r} \rightarrow \text{t} \]
Deep Learning Tools for Learning Sentence Representations

• Neural Word Embedding
• Recurrent Neural Networks
• Recursive Neural Networks
• Convolutional Neural Networks
**Word Representation: Neural Word Embedding**

\[
\log \frac{P(w, c)}{P(w)P(c)} = W C^T
\]

<table>
<thead>
<tr>
<th></th>
<th>(C_1)</th>
<th>(C_2)</th>
<th>(C_3)</th>
<th>(C_4)</th>
<th>(C_5)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(W_1)</td>
<td>10</td>
<td></td>
<td>1</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>(W_2)</td>
<td></td>
<td>2</td>
<td></td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>(W_3)</td>
<td>2.5</td>
<td></td>
<td></td>
<td>1</td>
<td></td>
</tr>
</tbody>
</table>

\[
M = WC^T
\]

**Matrix Factorization**

- **Word Embedding** or **word2vec**

**Diagram:**
- Matrix factorization: \(M = WC^T\)
- Word embedding or word2vec
Recurrent Neural Network (RNN) (Mikolov et al. 2010)

- On sequence
- Variable length
- Long dependency: long short term memory (LSTM)

\[ h_t = f(h_{t-1}, w_t) \]
Recursive Neural Network (RNN)  
(Socher et al. 2011)

- Based on parse tree
- Auto-encoding and decoding

the cat sat on the mat
Convolutional Neural Network (CNN) (Hu et al. 2014)

- Robust parsing
- Shared parameter
- Fixed length, zero padding

the cat sat on the mat

max pooling

convolution
Talk Outline

• Research at Huawei Noah’s Ark Lab
• New Breakthrough in Natural Language Processing with Deep Learning
• Research on DL for NLP at Noah’s Ark Lab
• Summary
Researchers Working on Deep Learning in Noah’s Ark Lab

Zhengdong Lu  Lifeng Shang  Lin Ma
Natural Language Dialogue

- Vast amount of conversation data is available
- Powerful technologies like deep learning developed
- Single turn vs multi-turn dialogue
- Two approaches
  - Retrieval based
  - Generation based

NTCIR: Short Text Conversation Contest
5.5 million message response pairs
Natural Language Dialogue System - Retrieval based Approach

- Retrieval based Approach
- Index of messages and responses matching
- Retrieved messages and responses
- Matching
- Matched responses
- Ranking
- Ranked responses
- Online
- Offline
- Best response
- Index of messages and responses
- Matching models
- Ranking model
Deep Match CNN
- Architecture I

• First represent two sentences, and then match
Deep Match CNN - Architecture II

• Represent and match two sentences simultaneously

• Two dimensional convolution and pooling
Retrieval based Approach:
Accuracy = 70%+

上海今天好熱，堪比新加坡。

上海今天热的不一般。

想去武当山 有想同游的么？

我想跟帅哥同游~哈哈

It is very hot in Shanghai today, just like Singapore.
It is unusually hot.

I want to go to Mountain Wudang, it there anybody going together with me?

Haha, I want to go with you, handsome boy
Natural Language Dialogue System - Generation based Approach

- Encoding messages to intermediate representations
- Decoding intermediate representations to responses
- Recurrent Neural Network (RNN)
Generation based Approach
Accuracy = 70%+

占中终于结束了。  
Occupy Central is finally over.

下一个是陆家嘴吧？  
Will Lujiazui (finance district in Shanghai) be the next?

我想买三星手机。  
I want to buy a Samsung phone

还是支持一下国产的吧。  
Why not buy our national brands?
Image Retrieval System

• Task
  – Image search on smartphone
  – Key: matching text to images

• Technology
  – Deep model for matching text and image

Find the picture that I had dinner with my friends at an Italian restaurant in Hong Kong
Deep Match Model for Image and Text
-CNN Based

- Represent text and image and then match the two
Deep Match Model for Image and Text - Top 10 Recall > 60%

Outperforms all existing methods in image retrieval
Machine Translation System

- Bilingual Corpus
- Phrase Alignment
- Rule Extraction
- Translation Rule Table
- Monolingual Corpus
- Language Model Training
- Language Model
- Translation Model Training
- Translation Model
- Training Data
- Translation Model
- Target Language Sentence
- Decoder
- Translation Template
- Translation Dictionary
- Source Language Sentence
- offline
- online
CNN based Translation Model

- Improved BLEU Score
  1 point vs BBN model (ACL’14 best paper) on benchmark data
CNN based Language Model

Chile holds parliament and presidential election

Improved BLEU Score
1 point on benchmark dataset
Talk Outline

• Research at Huawei Noah’s Ark Lab
• New Breakthrough in Natural Language Processing with Deep Learning
• Research on DL for NLP at Noah’s Ark Lab
• Summary
Research Finding and Advantages

• Possible to learn sentence representations in different tasks
• Completely data-driven, no human intervention
• With big data and powerful computers, we can really build human intelligence
Limitation and Challenges

• Black box: difficult to understand mechanism

• Naturalness vs factualness
  – E.g. “Clinton was born in DDDDD, and was educated in the university of edinburgh”

• Negation
  – E.g. “love” vs “does not love”

• Non-compositional expressions
  – E.g. “kick the bucket”
Take-away Messages

• Noah’s Ark Lab is working on
  – Intelligent telecommunication networks
  – Intelligent mobile devices
  – Intelligent enterprise

• Recent breakthrough in representation learning of sentence meaning
  – Neural networks, i.e., deep learning
  – Big Data
  – Task-oriented
  – Error-driven and gradient-based
  – Compositional

• Noah’s Ark Lab is working on DL for NLP
  – Natural language dialogue
  – Machine translation
  – Image retrieval
References


Thank you!


Our Email Address: noahlab@huawei.com