# <<Text Understanding from Scratch>>

Tianyi Luo
2015-03-18

# Author & Publication


- Xiang Zhang, Yann LeCun

  Computer Science Department,

  Courant Institute of Mathematical Sciences,

  New York University



- arXiv:1502.01710v1 [cs.LG] 5 Feb 2015

# Abstract

This article demontrates that we can apply deep learning to text understanding from character-level inputs all the way up to abstract text concepts, using temporal convolutional networks(LeCun et al., 1998) (ConvNets). We apply ConvNets to various large-scale datasets, including ontology classification, sentiment analysis, and text categorization. We show that temporal ConvNets can achieve astonishing performance without the knowledge of words, phrases, sentences and any other syntactic or semantic structures with regards to a human language. Evidence shows that our models can work for both English and Chinese.

# Text understanding

- What does Text understanding consist?
  - Consist in reading texts formed in natural languages.
  - Consist in **determining the explicit or implicit meaning** of each elements such as words, phrases, sentences and paragraphs.
  - Consist in **making inferences about the implicit or explicit properties** of these texts.

# Text understanding

- Disadvantages of Traditional methods of Text understanding
  - **Prior knowledge** is required and not cheap.

    ( They need to pre-define a dictionary of interested words, etc.)
  - Work well enough when applied to a **narrowly defined domain**.
  - Specialized to **a particular language**.

    ( They need structural parser for specific language.)
  - **After applying word2vector, there are still some engineered layers to represent structures such as words, phrases and sentences**

# Text Understanding From Scratch

- Contributions of this paper
  - ConvNets **do not require knowledge of words** – working with characters is fine.
  - ConvNets **do not require knowledge of syntax or semantic structures** – inference directly to high-level targets is fine.

# Text Understanding From Scratch

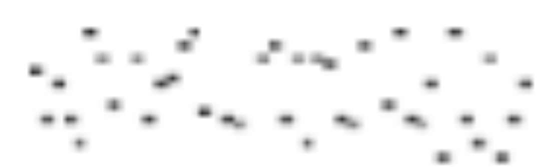
- Motivation of this paper
  - Our approach is **partly inspired by ConvNet's success in computer vision**. It has outstanding performance in various image recognition tasks.
  - These successful results usually involve some **end-to-end ConvNet model that learns hierarchical representation from raw pixels**.
  - ConvNets Similarly, we hypothesize that **when trained from raw characters, temporal ConvNet is able to learn the hierarchical representations of words, phrases and sentences in order to understand text**.

# Text Understanding From Scratch

- Character quantization
  - Encode 69 characters: **abcdefghijklmnopqrstuvwxyz0123456789-,;.!?:’’’/\|_@#$%ˆ & * ˜‘+-=<>()[]{}**
  - **“a”: {1,0,0,…,0}  “b”:{0,1,…,0}  “)”: {0,…,1,0,0,0,0}**

    **The dimension of these vectorc is 69.**

    **Other characters including blank characters are quantized as all-zero vectors.**



Inspired by (RSTM)work, we quantize characters in backward order.

The binary expression of
“International Conference on Machine Learning”

# Text Understanding From Scratch

- Model Design

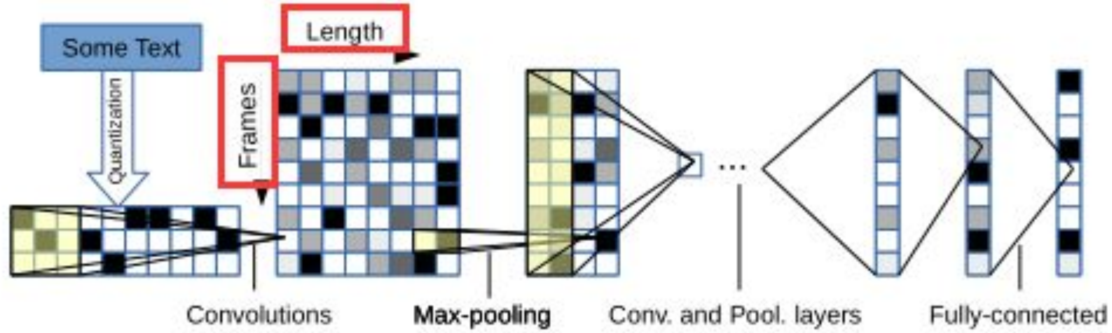



*Figure 2.* Illustration of our model

**The input have number of frames equal to 69** due to our character quantization method, and **the length of each frame is dependent on the problem**.

# Text Understanding From Scratch

- Model Design

*Table 1.* Convolutional layers used in our experiments. The convolutional layers do not use stride and pooling layers are all non-overlapping ones, so we omit the description of their strides.

| Layer | Large Frame | Small Frame | Kernel | Pool |
|---|---|---|---|---|
| 1 | 1024 | 256 | 7 | 3 |
| 2 | 1024 | 256 | 7 | 3 |
| 3 | 1024 | 256 | 3 | N/A |
| 4 | 1024 | 256 | 3 | N/A |
| 5 | 1024 | 256 | 3 | N/A |
| 6 | 1024 | 256 | 3 | 3 |

# Text Understanding From Scratch

- Model Design

*Table 2.* Fully-connected layers used in our experiments. The number of output units for the last layer is determined by the problem. For example, for a 10-class classification problem it will be 10.

| Layer | Output Units Large | Output Units Small |
|---|---|---|
| 7 | 2048 | 1024 |
| 8 | 2048 | 1024 |
| 9 | Depends on the problem | |

# Text Understanding From Scratch

- Data Augmentation using Thesaurus
  - **Image recognition** a model should **have some controlled invariance towards changes in translating, scaling, rotating and flipping** of the input image.
  - Similarly, in **speech recognition** we usually **augment data** by **adding artificial noise background and changing the tone or speed of speech signal**.
  - In terms of texts, the most natural choice in data augmentation for us is **to replace words or phrases with their synonyms** because the exact order of characters may form rigorous syntactic and semantic meaning.

# Text Understanding From Scratch

- Comparison Model
  - **Bag of Words**

The bag-of-words model is pretty straightforward. For each dataset, we count how many times each word appears in the training dataset, and choose 5000 most frequent ones as the bag. Then, we use multinomial logistic regression as the classifier for this bag of features.

  - **wore2vec**

As for the word2vec model, we first ran k-means on the word vectors learnt from Google News corpus with $k = 5000$, and then use a bag of these centroids for multinomial logistic regression. This model is quite similar to the bag-of-words model in that the number of features is also 5000.

# Text Understanding From Scratch

- DBpedia Ontology Classification

*Table 3.* DBpedia ontology classes. The numbers contain only samples with both a title and a short abstract.

| Class | Total | Train | Test |
|---|---|---|---|
| Company | 63,058 | 40,000 | 5,000 |
| Educational Institution | 50,450 | 40,000 | 5,000 |
| Artist | 95,505 | 40,000 | 5,000 |
| Athlete | 268,104 | 40,000 | 5,000 |
| Office Holder | 47,417 | 40,000 | 5,000 |
| Mean Of Transportation | 47,473 | 40,000 | 5,000 |
| Building | 67,788 | 40,000 | 5,000 |
| Natural Place | 60,091 | 40,000 | 5,000 |
| Village | 159,977 | 40,000 | 5,000 |
| Animal | 187,587 | 40,000 | 5,000 |
| Plant | 50,585 | 40,000 | 5,000 |
| Album | 117,683 | 40,000 | 5,000 |
| Film | 86,486 | 40,000 | 5,000 |
| Written Work | 55,174 | 40,000 | 5,000 |

The length of input used was $l_0 = 1014$.

# Text Understanding From Scratch

- DBpedia Ontology Classification

*Table 4.* DBpedia results. The numbers are accuracy.

| Model | Thesaurus | Train | Test |
|---|---|---|---|
| Large ConvNet | No | **99.95%** | 98.26% |
| Large ConvNet | Yes | 99.81% | **98.40%** |
| Small ConvNet | No | 99.70% | 97.99% |
| Small ConvNet | Yes | 99.64% | 98.15% |
| Bag of Words | No | 96.62% | 96.43% |
| word2vec | No | 89.64% | 89.41% |

Experiment Result

# Text Understanding From Scratch

- Amazon Review Sentiment Analysis

*Table 5.* Amazon review datasets. Column "total" is the total number of samples for each score. Column "chosen" is the number of samples whose length is between 100 and 1000. Column "full" and "polarity" are number of samples chosen for full score dataset and polarity dataset, respectively.

| | Total | Chosen | Full | Polarity |
|---|---|---|---|---|
| 1 | 2,746,559 | 2,206,886 | 1,250,000 | 2,200,000 |
| 2 | 1,791,219 | 1,290,278 | 1,250,000 | 1,250,000 |
| 3 | 2,892,566 | 1,975,014 | 1,250,000 | 0 |
| 4 | 6,551,166 | 4,576,293 | 1,250,000 | 1,250,000 |
| 5 | 20,705,260 | 16,307,871 | 1,250,000 | 2,200,000 |

The length of input used was $l_0 = 1014$.

# Text Understanding From Scratch

- Amazon Review Sentiment Analysis

*Table 6.* Result on Amazon review full score dataset. The numbers are accuracy.

| Model | Thesaurus | Train | Test |
|---|---|---|---|
| Large ConvNet | No | **93.73%** | **73.28%** |
| Large ConvNet | Yes | 83.67% | 71.37% |
| Small ConvNet | No | 82.10% | 70.12% |
| Small ConvNet | Yes | 84.42% | 68.18% |
| Bag of Words | No | 52.13% | 51.93% |
| word2vec | No | 38.22% | 38.25% |

*Table 7.* Result on Amazon review polarity dataset. The numbers are accuracy.

| Model | Thesaurus | Train | Test |
|---|---|---|---|
| Large ConvNet | No | **99.71%** | **96.34%** |
| Large ConvNet | Yes | 99.51% | 96.08% |
| Small ConvNet | No | 98.24% | 95.84% |
| Small ConvNet | Yes | 98.57% | 96.01% |
| Bag of Words | No | 88.46% | 85.54% |
| word2vec | No | 75.15% | 73.07% |

Experiment Result

# Text Understanding From Scratch

- Yahoo! Answers Topic Classification

*Table 8.* Yahoo! Answers topic classification dataset

| Category | Total | Train | Test |
|---|---|---|---|
| Society & Culture | 295,340 | 140,000 | 5,000 |
| Science & Mathematics | 169,586 | 140,000 | 5,000 |
| Health | 278,942 | 140,000 | 5,000 |
| Education & Reference | 206,440 | 140,000 | 5,000 |
| Computers & Internet | 281,696 | 140,000 | 5,000 |
| Sports | 146,396 | 140,000 | 5,000 |
| Business & Finance | 265,182 | 140,000 | 5,000 |
| Entertainment & Music | 440,548 | 140,000 | 5,000 |
| Family & Relationships | 517,849 | 140,000 | 5,000 |
| Politics & Government | 152,564 | 140,000 | 5,000 |

The length of input used was $l_0 = 1014$.

# Text Understanding From Scratch

- Yahoo! Answers Topic Classification

*Table 9.* Results on Yahoo! Answers dataset. The numbers are accuracy.

| Model | Thesaurus | Train | Test |
|---|---|---|---|
| Large ConvNet | No | 71.76% | 69.84% |
| Large ConvNet | Yes | **72.23%** | **69.92%** |
| Small ConvNet | No | 70.10% | 69.92% |
| Small ConvNet | Yes | 70.73% | 69.81% |
| Bag of Words | No | 66.75% | 66.44% |
| word2vec | No | 58.84% | 59.01% |

Experiment Result

# Text Understanding From Scratch

- News Categorization in English

*Table 10.* AG's news corpus. Only categories used are listed.

| Category | Total | Train | Test |
|---|---|---|---|
| World | 81,456 | 40,000 | 1,100 |
| Sports | 62,163 | 40,000 | 1,100 |
| Business | 56,656 | 40,000 | 1,100 |
| Sci/Tech | 41,194 | 40,000 | 1,100 |

The length of input used was $l_0 = 1014$.

# Text Understanding From Scratch

- News Categorization in English

*Table 11.* Result on AG's news corpus. The numbers are accuracy

| Model | Thesaurus | Train | Test |
|---|---|---|---|
| Large ConvNet | No | 99.00% | 91.12% |
| Large ConvNet | Yes | **99.00%** | **91.64%** |
| Small ConvNet | No | 98.94% | 89.32% |
| Small ConvNet | Yes | 98.97% | 90.39% |
| Bag of Words | No | 88.35% | 88.29% |
| word2vec | No | 85.30% | 85.28% |

Experiment Result

# Text Understanding From Scratch

- News Categorization in Chinese

*Table 12.* Sogou News dataset

| Category | Total | Train | Test |
|---|---|---|---|
| Sports | 645,931 | 150,000 | 10,000 |
| Finance | 315,551 | 150,000 | 10,000 |
| Entertainment | 160,409 | 150,000 | 10,000 |
| Automobile | 167,647 | 150,000 | 10,000 |
| Technology | 188,111 | 150,000 | 10,000 |

The length of input used was $l_0 = 1014$.

# Text Understanding From Scratch

- News Categorization in Chinese
  - The **romanization** or **latinization** form we have used is **Pinyin**, which is a **phonetic system** for transcribing the **Mandarin pronunciations**.
  - During this procedure, we used the **pypinyin** package combined with **jieba Chinese segmentation system**. The resulting Pinyin text had **each tone appended their finals as numbers between 1 and 4.**

# Text Understanding From Scratch

- News Categorization in Chinese

*Table 13.* Result on Sogou News corpus. The numbers are accuracy

| Model | Thesaurus | Train | Test |
|---|---|---|---|
| Large ConvNet | No | **97.64%** | **97.05%** |
| Small ConvNet | No | 97.45% | 97.03% |
| Bag of Words | No | 95.69% | 95.46% |

Experiment Result

# Text Understanding From Scratch

- Some ideas
  - 1. This model is successfully applied in Text Understanding, can we apply this model **in language model**? **Letter embedding**?
  - 2. Recent research shows that **it is possible to generate text description of images from the features learnt in a deep image recognition model**. The models in this article show very good ability for understanding natural languages, and we are interested in **using the features from our model to generate a response sentence** in similar ways?
  - 3. **Natural language in its essence is time-series** in disguise. Can we **extend application for our approach towards time-series data**?

# Text Understanding From Scratch

- Some ideas
  - 4. In this article we **only apply ConvNets to text understanding for its semantic or sentiment meaning**.

    Can we extend this approach towards **NER** or **POS**?
  - 5. Can we learn from **symbolic systems** such as **mathematical equations, logic expressions or programming languages**?
  - 6. Can we extend this approach towards **other tasks**, not just classification task?

# Text Understanding From Scratch

- Some ideas
  - 7. The same idea came in the paper called <<Deep Speech: Scaling up end-to-end speech recognition>>.

We present a state-of-the-art speech recognition system developed using end-to-end deep learning. Our architecture is significantly simpler than traditional speech systems, which rely on laboriously engineered processing pipelines; these traditional systems also tend to perform poorly when used in noisy environments. In contrast, our system does not need hand-designed components to model background noise, reverberation, or speaker variation, but instead directly learns a function that is robust to such effects. We do not need a phoneme dictionary, nor even the concept of a "phoneme." Key to our approach is a well-optimized RNN training system that uses multiple GPUs, as well as a set of novel data synthesis techniques that allow us to efficiently obtain a large amount of varied data for training. Our system, called Deep Speech, outperforms previously published results on the widely studied Switchboard Hub5'00, achieving 16.0% error on the full test set. Deep Speech also handles challenging noisy environments better than widely used, state-of-the-art commercial speech systems.

## Thank You !