CS224d: Deep NLP

Lecture 15: Applications of DL to NLP

Richard Socher
richard@metamind.io
Overview

• Model overview: How to properly compare to other models and choose your own model
  • Word representation
  • Phrase composition
  • Objective function
  • Optimization

• Character RNNs on text and code
• Morphology
• Logic
• Question Answering
• Image – Sentence mapping
Model overview: Word Vectors

- Random
- Word2Vec
- Glove

- Dimension – often defines the number of model parameters

- Or work directly on characters or morphemes
Model overview: Phrase Vector Composition

• Composition Function governs how exactly word and phrase vectors interact to compose meaning

• Averaging: \( p = a + b \)
  • Lots of simple alternatives

• Recursive neural networks

• Convolutional neural networks

• Recurrent neural network
Composition: Bigram and Recursive functions

• Many related models are special cases of MV-RNN

\[ p = f \left( W \begin{bmatrix} Ba \\ Ab \end{bmatrix} \right) \]

• Mitchell and Lapata, 2010; Zanzotto et al., 2010:

\[ p = Ba + Ab = id \left( \begin{bmatrix} I_{n \times n} & I_{n \times n} \\ I_{n \times n} & I_{n \times n} \end{bmatrix} \begin{bmatrix} Ba \\ Ab \end{bmatrix} \right) \]

• Baroni and Zamparelli (2010): A is an adjective matrix and b is a noun vector

\[ p = Ab = id \left( \begin{bmatrix} 0_{n \times n} & I_{n \times n} \\ I_{n \times n} & I_{n \times n} \end{bmatrix} \begin{bmatrix} Ba \\ Ab \end{bmatrix} \right) \]

• RNNs of Socher et al. 2011 (ICML, EMNLP, NIPS) are also special cases

\[ p = f \left( W \begin{bmatrix} I_{n \times n}a \\ I_{n \times n}b \end{bmatrix} \right) \]

• Recursive neural tensor networks bring quadratic and multiplicative interactions between vectors
Additional choice for recursive neural nets

- Dependency trees focus more on semantic structure

1. Constituency Tree

2. Dependency Tree

3. Balanced Tree
Composition: CNNs

• Several variants also:
  • No pooling layers
  • Pooling layers: simple max-pooling or dynamic pooling
  • Pooling across different dimensions

• Somewhat less explored in NLP than RNNs²

• Not linguistically nor cognitively plausible
Composition: Recurrent Neural Nets

- Vanilla
- GRU
- LSTM
- Many variants of LSTMs
  “LSTM: A Search Space Odyssey” by Greff et al. 2015
Model overview: Objective function

- **Max-margin**

- **Cross-entropy**
  - Supervised to predict a class
  - Unsupervised: predict surrounding words

- **Auto-encoder**
  - My opinion: Unclear benefits for NLP
  - Unless encoding another modality
Optimization

- Initialization (word vector and composition parameters)!!

- Optimization algorithm
  - SGD
  - SGD + momentum
  - L-BFGS
  - AdaGrad
  - Adelta

- Optimization tricks
  - Regularization (some define as part of model)
  - Dropout
Character RNNs on text and code

http://karpathy.github.io/2015/05/21/rnn-effectiveness/
Character RNNs on text and code

• Haven’t yet produced useful results on real datasets

• Shows that RNNs can memorize sequences and keep memory (mostly LSTMs)

• Most interesting results simply train on dataset and sample from it afterwards (first shown by Sutskever et al. 2011: Generating Text with Recurrent Neural Networks)

• Results from an LSTM (karpathy.github.io)
PANDARUS:
Alas, I think he shall be come approached and the day
When little srain would be attain'd into being never fed,
And who is but a chain and subjects of his death,
I should not sleep.

Second Senator:
They are away this miseries, produced upon my soul,
Breaking and strongly should be buried, when I perish
The earth and thoughts of many states.

DUKE VINCENTIO:
Well, your wit is in the care of side and that.
Naturalism and decision for the majority of Arab countries' capitalide was grounded by the Irish language by [[John Clair]], [[An Imperial Japanese Revolt]], associated with Guangzham's sovereignty. His generals were the powerful ruler of the Portugal in the [[Protestant Immineners]], which could be said to be directly in Cantonese Communication, which followed a ceremony and set inspired prison, training. The emperor travelled back to [[Antioch, Perth, October 25|21]] to note, the Kingdom of Costa Rica, unsuccessful fashioned the [[Thrales]], [[Cynth's Dajoard]], known in western [[Scotland]], near Italy to the conquest of India with the conflict.
Proof. Omitted.

Lemma 0.1. Let $\mathcal{C}$ be a set of the construction.

Let $\mathcal{C}$ be a gerber covering. Let $\mathcal{F}$ be a quasi-coherent sheaves of $\mathcal{O}$-modules. We have to show that

\[ \mathcal{O}_{\mathcal{O}_X} = \mathcal{O}_X(\mathcal{L}) \]

Proof. This is an algebraic space with the composition of sheaves $\mathcal{F}$ on $X_{\text{etale}}$ we have

\[ \mathcal{O}_X(\mathcal{F}) = \{ \text{morph}_1 \times \mathcal{O}_X (\mathcal{G}, \mathcal{F}) \} \]

where $\mathcal{G}$ defines an isomorphism $\mathcal{F} \to \mathcal{F}$ of $\mathcal{O}$-modules.

Lemma 0.2. This is an integer $\mathcal{Z}$ is injective.

Proof. See Spaces, Lemma ??.

Lemma 0.3. Let $S$ be a scheme. Let $X$ be a scheme and $X$ is an affine open covering. Let $U \subset X$ be a canonical and locally of finite type. Let $X$ be a scheme. Let $X$ be a scheme which is equal to the formal complex.

The following to the construction of the lemma follows.

Let $X$ be a scheme. Let $X$ be a scheme covering. Let

\[ b : X \to Y' \to Y \to Y' \times_X Y \to X \]

be a morphism of algebraic spaces over $S$ and $Y$.

Proof. Let $X$ be a nonzero scheme of $X$. Let $X$ be an algebraic space. Let $\mathcal{F}$ be a quasi-coherent sheaf of $\mathcal{O}_X$-modules. The following are equivalent

1. $\mathcal{F}$ is an algebraic space over $S$.
2. If $X$ is an affine open covering.

Consider a common structure on $X$ and $X$ the functor $\mathcal{O}_X(U)$ which is locally of finite type.

This since $\mathcal{F} \in \mathcal{F}$ and $x \in \mathcal{G}$ the diagram

\[ \begin{array}{ccc}
S & \longrightarrow & \mathcal{O}_{X'} \\
\xi & \downarrow & \downarrow \\
\mathcal{O}_X & \longrightarrow & \mathcal{O}_{X'} \\
\downarrow & \quad & \downarrow \\
\mathcal{O}_X & \longrightarrow & \mathcal{O}_X \\
\end{array} \]

is a limit. Then $\mathcal{G}$ is a finite type and assume $S$ is a flat and $\mathcal{F}$ and $\mathcal{G}$ is a finite type $f_\ast$. This is of finite type diagrams, and

- the composition of $\mathcal{G}$ is a regular sequence,
- $\mathcal{O}_X'$ is a sheaf of rings.

Proof. We have see that $X = \text{Spec}(R)$ and $\mathcal{F}$ is a finite type representable by algebraic space. The property $\mathcal{F}$ is a finite morphism of algebraic stacks. Then the cohomology of $X$ is an open neighbourhood of $U$.

Proof. This is clear that $\mathcal{G}$ is a finite presentation, see Lemmas ??.

A reduced above we conclude that $U$ is an open covering of $C$. The functor $\mathcal{F}$ is a "field"

\[ \mathcal{O}_{X, \ast} \longrightarrow \mathcal{F} \]

is an isomorphism of covering of $\mathcal{O}_{X, \ast}$. If $\mathcal{F}$ is the unique element of $\mathcal{F}$ such that $X$ is an isomorphism.

The property $\mathcal{F}$ is a disjoint union of Proposition ?? and we can filtered set of presentations of a scheme $\mathcal{O}_X$-algebra with $\mathcal{F}$ are opens of finite type over $S$.

If $\mathcal{F}$ is a scheme theoretic image points.

If $\mathcal{F}$ is a finite direct sum $\mathcal{O}_{X, \ast}$ is a closed immersion, see Lemma ??, This is a sequence of $\mathcal{F}$ is a similar morphism.
/*
  * Increment the size file of the new incorrect UI_FILTER group information
  * of the size generatively.
  */
static int indicate_policy(void)
{
    int error;
    if (fd == MARN_EPT) {
        /*
        * The kernel blank will coedl it to userspace.
        */
        if (ss->segment < mem_total)
            unblock_graph_and_set_blocked();
        else
            ret = 1;
            goto bail;
    }
    segaddr = in_SB(in.addr);
    selector = seg / 16;
    setup_works = true;
Morphology

- Better Word Representations with Recursive Neural Networks for Morphology – Luong et al. (slides from Luong)

- *Problem with word vectors*: poorly estimate rare and complex words.

<table>
<thead>
<tr>
<th></th>
<th>(Collobert &amp; Weston, 2010)</th>
<th>(Huang et. al., 2012)</th>
</tr>
</thead>
<tbody>
<tr>
<td>distinct</td>
<td>different distinctive broader narrower</td>
<td>unique broad distinctive separate</td>
</tr>
<tr>
<td>distinctness</td>
<td>morphologies pesawat clefts pathologies</td>
<td>companion roskam hitoshi enjoyed</td>
</tr>
<tr>
<td>affect</td>
<td>exacerbate impacts characterize</td>
<td>allow prevent involve enable</td>
</tr>
<tr>
<td>unaffected</td>
<td>unnoticed dwarfed mitigated</td>
<td>monti sheaths krystal</td>
</tr>
</tbody>
</table>
Limitations of existing work

- Treat related words as independent entities.
- Represent unknown words with a few vectors.

Word frequencies in Wikipedia docs (986m tokens)
Luong’s approach – network structure

- **Neural Language Model**: simple **feed-forward network** (Huang, et al., 2012) with **ranking-type cost** (Collobert et al., 2011).

- **Morphology Model**: **recursive neural network** (Socher et al., 2011).
Analysis

- Blends word structure and syntactic-semantic information.

<table>
<thead>
<tr>
<th>Words</th>
<th>(Collobert et al., 2011)</th>
<th>This work</th>
</tr>
</thead>
<tbody>
<tr>
<td>commenting</td>
<td>insisting insisted focusing</td>
<td>commented comments criticizing</td>
</tr>
<tr>
<td>unaffected</td>
<td>unnoticed dwarfed mitigated</td>
<td>undesired unhindered unrestricted</td>
</tr>
<tr>
<td>distinct</td>
<td>different distinctive broader</td>
<td>divergent diverse distinctive</td>
</tr>
<tr>
<td>distinctness</td>
<td>morphologies pesawat clefts</td>
<td>distinctiveness smallness largeness</td>
</tr>
<tr>
<td>heartlessness</td>
<td>$\emptyset$</td>
<td>corruptive inhumanity ineffectual</td>
</tr>
<tr>
<td>saudi-owned</td>
<td>avatar mohajir kripalani</td>
<td>saudi-based syrian-controlled</td>
</tr>
</tbody>
</table>
Solutions to the problem of polysemous words

- Improving Word Representations Via Global Context And Multiple Word Word Prototypes by Huang et al. 2012
Natural language inference

*Claim:* Simple task to define, but engages the full complexity of compositional semantics:

- Lexical entailment
- Quantification
- Coreference
- Lexical/scope ambiguity
- Commonsense knowledge
- Propositional attitudes
- Modality
- Factivity and implicativity
First training data

- **Training data:**
  - *dance* **entails** *move*
  - *waltz* **neutral** *tango*
  - *tango* **entails** *dance*
  - *sleep* **contradicts** *dance*
  - *waltz* **entails** *dance*

Memorization (training set): Generalization (test set):

- *dance* ??? *move*  
  - *sleep* ??? *waltz*
- *waltz* ??? *tango*  
  - *tango* ??? *move*
# Natural language inference: definitions!

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Definition</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>$x \equiv y$</td>
<td>equivalence</td>
<td>$couch \equiv sofa$</td>
</tr>
<tr>
<td>$x \sqsubseteq y$</td>
<td>forward entailment</td>
<td>$crow \sqsubseteq bird$</td>
</tr>
<tr>
<td></td>
<td>(strict)</td>
<td></td>
</tr>
<tr>
<td>$x \sqsupseteq y$</td>
<td>reverse entailment</td>
<td>$European \sqsupseteq French$</td>
</tr>
<tr>
<td></td>
<td>(strict)</td>
<td></td>
</tr>
<tr>
<td>$x \land y$</td>
<td>negation</td>
<td>$human \land nonhuman$</td>
</tr>
<tr>
<td></td>
<td>(exhaustive exclusion)</td>
<td></td>
</tr>
<tr>
<td>$x \mid y$</td>
<td>alternation</td>
<td>$cat \mid dog$</td>
</tr>
<tr>
<td></td>
<td>(non-exhaustive exclusion)</td>
<td></td>
</tr>
<tr>
<td>$x \supset y$</td>
<td>cover</td>
<td>$animal \supset nonhuman$</td>
</tr>
<tr>
<td></td>
<td>(exhaustive non-exclusion)</td>
<td></td>
</tr>
<tr>
<td>$x \neq y$</td>
<td>independence</td>
<td>$hungry \neq hippo$</td>
</tr>
</tbody>
</table>
A minimal NN for lexical relations

- Words are learned embedding vectors.
- One plain RNN or RNTN layer
- Softmax emits relation labels
- Learn everything with SGD.
Recursion in propositional logic

Experimental approach: Train on relational statements generated from some formal system, test on other such relational statements.

The model needs to:

• Learn the relations between individual words. (lexical relations)
• Learn how lexical relations impact phrasal relations.
  • This needs to be recursively applicable!

• \( a \equiv a, \ a \land (\neg a), \ a \equiv (\neg (\neg a)), \ ... \)
Natural language inference with RNNs

- Two trees + learned comparison layer, then a classifier:
Natural language inference with RNNs

Accuracy

Size of longer expression

25d TreeRNTN

45d TreeRNN

45d SumNN

# only
**Question Answering: Quiz Bowl Competition**

- **QUESTION:**
  He left unfinished a novel whose title character forges his father's signature to get out of school and avoids the draft by feigning desire to join. A more famous work by this author tells of the rise and fall of the composer Adrian Leverkühn. Another of his novels features the jesuit Naptha and his opponent Settembrini, while his most famous work depicts the aging writer Gustav von Aschenbach. Name this German author of The Magic Mountain and Death in Venice.

- Iyyer et al. 2014: A Neural Network for Factoid Question Answering over Paragraphs
• **QUESTION:**
  He left unfinished a novel whose title character forges his father's signature to get out of school and avoids the draft by feigning desire to join. A more famous work by this author tells of the rise and fall of the composer Adrian Leverkühn. Another of his novels features the jesuit Naptha and his opponent Settembrini, while his most famous work depicts the aging writer Gustav von Aschenbach. Name this German author of The Magic Mountain and Death in Venice.

• **ANSWER:** Thomas Mann
Recursive Neural Networks

• Follow dependency structure
Pushing Facts into Entity Vectors
Table 1: Accuracy for history and literature at the first two sentence positions of each question and the full question. The top half of the table compares models trained on questions only, while the IR models in the bottom half have access to Wikipedia.

Qanta outperforms all baselines that are restricted to just the question data, and it substantially improves an IR model with access to Wikipedia despite being trained on much less data.

Figure 4: Comparisons of qanta+ir-wiki to human quiz bowl players. Each bar represents an individual human, and the bar height corresponds to the difference between the model score and the human score. Bars are ordered by human skill. Red bars indicate that the human is winning, while blue bars indicate that the model is winning.

qanta+ir-wiki outperforms most humans on history questions but fails to defeat the “average” human on literature questions.

A minor character in this play can be summoned by a bell that does not always work; that character also doesn’t have eyelids. Near the end, a woman who drowned her illegitimate child attempts to stab another woman in the Second Empire-style room in which the entire play takes place. For 10 points, Estelle and Ines are characters in which existentialist play in which Garcin claims “Hell is other people”, written by Jean-Paul Sartre?

Figure 3: A question on the play “No Exit” with human buzz position marked as 3. Since the buzz occurs in the middle of the second sentence, our model is only allowed to see the first sentence.

5.1 Experimental Results

Table 1 shows that when bag of words and information retrieval methods are restricted to question data, they perform significantly worse than qanta on early sentence positions. The performance of bow-dt indicates that while the dependency tree structure helps by itself, the compositional distributed representations learned by qanta are more useful. The significant improvement when we train answers as part of our vocabulary (see Section 3.2) indicates that our model uses answer occurrences within question text to learn a more informative vector space.

The disparity between ir-qb and ir-wiki indicates that the information retrieval models need lots of external data to work well at all sentence positions. ir-wiki performs better than other models because Wikipedia contains many more sentences that partially match specific words or phrases found in early clues than the question training set. In particular, it is impossible for all other models to answer clues in the test set that have no semantically similar...
Table 1: Accuracy for history and literature at the first two sentence positions of each question and the full question. The top half of the table compares models trained on questions only, while the IR models in the bottom half have access to Wikipedia.

qanta outperforms all baselines that are restricted to just the question data, and it substantially improves an IR model with access to Wikipedia despite being trained on much less data.

The disparity between ir-qb and ir-wiki indicates that the information retrieval models need lots of external data to work well at all sentence positions. ir-wiki performs better than other models because Wikipedia contains many more sentences that partially match specific words or phrases found in early clues than the question training set. In particular, it is impossible for all other models to answer clues in the test set that have no semantically similar

5.1 Experimental Results

Table 1 shows that when bag of words and information retrieval methods are restricted to question data, they perform significantly worse than qanta on early sentence positions. The performance of bow-dt indicates that while the dependency tree structure helps by itself, the compositional distributed representations learned by qanta are more useful. The significant improvement when we train answers as part of our vocabulary (see Section 3.2) indicates that our model uses answer occurrences within question text to learn a more informative vector space.

The disparity between ir-qb and ir-wiki indicates that the information retrieval models need lots of external data to work well at all sentence positions.
Visual Grounding

- Idea: Map sentences and images into a joint space

Socher et al. 2013: Grounded Compositional Semantics for Finding and Describing Images with Sentences
Discussion: Compositional Structure

- Recursive Neural Networks so far used constituency trees which results in more syntactically influenced representations

- Instead: Use dependency trees which capture more semantic structure
Convolutional Neural Network for Images

- CNN trained on ImageNet (Le et al. 2013)
- RNN trained to give large inner products between sentence and image vectors:

\[ J(W_I, \theta) = \sum_{(i,j) \in \mathcal{P}} \sum_{c \in S \setminus S(i)} \max(0, \Delta - v_i^T y_j + v_i^T y_c) \]
Results

A gray convertible sports car is parked in front of the trees. ✓
A close-up view of the headlights of a blue old-fashioned car. ✗
Black shiny sports car parked on concrete driveway. ✓
Five cows grazing on a patch of grass between two roadways. ✗

A jockey rides a brown and white horse in a dirt corral. ✓
A young woman is riding a Bay hose in a dirt riding-ring. ✗
A white bird pushes a miniature teal shopping cart. ✗
A person rides a brown horse. ✓

A motocross bike with rider flying through the air. ✓
White propeller plane parked in middle of grassy field. ✗
The white jet with its landing gear down flies in the blue sky. ✗
An elderly woman catches a ride on the back of the bicycle. ✗
## Results

<table>
<thead>
<tr>
<th>Describing Images</th>
<th>Mean Rank</th>
<th>Image Search</th>
<th>Mean Rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random</td>
<td>92.1</td>
<td>Random</td>
<td>52.1</td>
</tr>
<tr>
<td>Bag of Words</td>
<td>21.1</td>
<td>Bag of Words</td>
<td>14.6</td>
</tr>
<tr>
<td>CT-RNN</td>
<td>23.9</td>
<td>CT-RNN</td>
<td>16.1</td>
</tr>
<tr>
<td>Recurrent Neural Network</td>
<td>27.1</td>
<td>Recurrent Neural Network</td>
<td>19.2</td>
</tr>
<tr>
<td>Kernelized Canonical Correlation</td>
<td>18.0</td>
<td>Kernelized Canonical Correlation Analysis</td>
<td>15.9</td>
</tr>
<tr>
<td>Analysis</td>
<td><strong>16.9</strong></td>
<td>DT-RNN</td>
<td><strong>12.5</strong></td>
</tr>
</tbody>
</table>
Several models came out simultaneously in 2015 that follow up

Replace recursive neural network with LSTM and instead of only finding vectors they generate the description

Mostly memorized training sequences (becomes similar again)

Donahue et al. 2015: Long-term $\rightarrow$ Long-term Recurrent Convolutional Networks for Visual Recognition and Description

Karpathy and Fei-Fei 2015: Deep Visual-Semantic Alignments for Generating Image Descriptions
Image – Sentence Generation (!)

"little girl is eating piece of cake."

"baseball player is throwing ball in game."

"woman is holding bunch of bananas."

"a young boy is holding a baseball bat."

"a cat is sitting on a couch with a remote control."

"a woman holding a teddy bear in front of a mirror."
Next Lecture

- The future (?) of deep learning for NLP

- No video