Advanced Data Science

Dr. Kira Radinsky

Slides Adapted from Richard Socher
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
• Chat Bots
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

- 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
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 RNNs\(^2\)

• 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
Auto Encoders

\[ \begin{bmatrix} C'_1 \\ C'_2 \end{bmatrix} \in \mathbb{R}^{2n} \]
\[ \begin{bmatrix} C_1 \\ C_2 \end{bmatrix} \in \mathbb{R}^n \]
\[ \begin{bmatrix} C'_1 \\ C'_2 \end{bmatrix} \in \mathbb{R}^{2n} \]

New representation

Original representation

How well can we re-create from the new representation the original one?
Optimization

• Initialization (word vector and composition parameters)!!

• Optimization algorithm
  • SGD
  • SGD + momentum (a good default to start with!)
  • L-BFGS
  • AdaGrad (a good default if you have word vectors that you are learning)
  • 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)
Shakespeare

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 $C$ be a set of the construction.
Let $C$ be a gerber covering. Let $F$ be a quasi-coherent sheaves of $O$-modules. We have to show that
\[ \mathcal{O}_{O_X} = \mathcal{O}_X(C) \]

Proof. This is an algebraic space with the composition sheaves $F$ on $X_{\text{tale}}$ we have
\[ \mathcal{O}_X(F) = \{ \text{morph}_1 \times_{\mathcal{O}_X} (G, F) \} \]
where $G$ defines an isomorphism $F \to F$ of $O$-modules.

Lemma 0.2. This is an integer $Z$ is injective.

Proof. See Spaces, Lemma 262.

Lemma 0.3. Let $S$ be a scheme. Let $X$ be a scheme and $X$ is an affine open covering. Let $U \subseteq X$ 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
\[ h : X \to X' \to Y \to Y' \times X Y \to X. \]
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 $F$ be a quasi-coherent sheaf of $O_X$-modules. The following are equivalent
1. $F$ is an algebraic space over $S$.
2. If $X$ is an affine open covering.

Consider a common structure on $X$ and $X$ the functor $O_X(U)$ which is locally of finite type.

This since $F \in F$ and $x \in G$ the diagram

\[ \text{Spec}(K) \]

is a limit. Then $G$ is a finite type and assume $S$ is a flat and $F$ and $G$ is a finite type $f$. This is of finite type diagrams, and
- the composition of $G$ is a regular sequence,
- $O_X$ is a sheaf of rings.

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

Proof. This is clear that $G$ is a finite presentation, see Lemmas 262.

A reduced above we conclude that $U$ is an open covering of $C$. The functor $F$ is a "field"
\[ O_{X_{\text{t}}} \to F_{\text{t}} : (O_{X_{\text{t}}}(x)) \to O_{X_{\text{t}}}(O_{X_{\text{t}}}(x)) \]
is an isomorphism of covering of $O_X$. If $F$ is the unique element of $F$ such that $X$ is an isomorphism.
The property $F$ is a disjoint union of Proposition 262 and we can filtered set of presentations of a scheme $O_X$-algebra with $F$ are open of finite type over $S$.

If $F$ is scheme theoretic image points.

If $F$ is a finite direct sum $O_X$ is a closed immersion, see Lemma 262. This is a sequence of $F$ is a similar morphism.
Code! (Linux source code)

```c
/*
 * 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 coeld 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;

    return ret;
}
```
So why are not we using this?

1. Does not necessarily give us the best semantic word vector representations that can be useful for downstream tasks
2. We are quite far in research from getting to semantically meaningful representation and not just syntactic
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.
  - There is no real connection between distinct, distinctive, distinctively, etc.
- Represent unknown words with a few vectors.

Word frequencies in Wikipedia docs (986m tokens)
Luong’s approach – **Context-sensitive morphological RNN**

- **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).

\[ v^T f(W[x_1; x_2; \ldots; x_n]+b) \]

\[ p = f(W_m[x_{stem}; x_{affix}]+b_m) \]
Morphological Recursive NN

• Assumes access to a dictionary of morphemic analyses of words
  • Use Morfessor
  • Distinct morphemes are encoded by column vectors

\[ \text{unfortunately}_{STM} \]
\[ W_m, b_m \]
\[ \text{unfortunate}_{STM} \]
\[ W_m, b_m \]
\[ \text{un}_{PRE} \]
\[ \text{fortunate}_{STM} \]

Computed on the fly

In the lexicon
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>Ø</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>
Improving Word Representations Via Global Context And Multiple 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* \textit{entails} move
  - *waltz* \textit{neutral} tango
  - *tango* \textit{entails} dance
  - *sleep* \textit{contradicts} dance
  - *waltz* \textit{entails} dance

- **Memorization (training set):**
  - *dance* ??? move
  - *waltz* ??? tango

- **Generalization (test set):**
  - sleep ??? waltz
  - tango ??? move
Natural language inference: definitions!

- $x \equiv y$: equivalence
  - Example: $couch \equiv sofa$
- $x \sqsubseteq y$: forward entailment (strict)
  - Example: $crow \sqsubseteq bird$
- $x \sqsupseteq y$: reverse entailment (strict)
  - Example: $European \sqsupseteq French$
- $x \blacksquare y$: negation (exhaustive exclusion)
  - Example: $human \blacksquare nonhuman$
- $x \mid y$: alternation (non-exhaustive exclusion)
  - Example: $cat \mid dog$
- $x \vartriangleright y$: cover (exhaustive non-exclusion)
  - Example: $animal \vartriangleright nonhuman$
- $x \# y$: independence
  - Example: $hungry \# hippo$
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. (composition)
  • This needs to be recursively applicable!

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

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

After each sentence of the question you can buzz in if you know the answer.
Literature Questions are Hard!

Not enough to look at the wikipedia entry and the question – you actually need to read the books!
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
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:

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

Pairs of sentence & image describing the same thing
Pairs of sentence & image NOT describing the same thing
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 horse 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. ☒️

Demo: http://etcml.com:8080/
## Results

### Describing Images

<table>
<thead>
<tr>
<th>Method</th>
<th>Mean Rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random</td>
<td>92.1</td>
</tr>
<tr>
<td>Bag of Words</td>
<td>21.1</td>
</tr>
<tr>
<td>CT-RNN</td>
<td>23.9</td>
</tr>
<tr>
<td>Recurrent Neural Network</td>
<td>27.1</td>
</tr>
<tr>
<td>Kernelized Canonical Correlation Analysis</td>
<td>18.0</td>
</tr>
<tr>
<td>DT-RNN</td>
<td><strong>16.9</strong></td>
</tr>
</tbody>
</table>

### Image Search

<table>
<thead>
<tr>
<th>Method</th>
<th>Mean Rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random</td>
<td>52.1</td>
</tr>
<tr>
<td>Bag of Words</td>
<td>14.6</td>
</tr>
<tr>
<td>CT-RNN</td>
<td>16.1</td>
</tr>
<tr>
<td>Recurrent Neural Network</td>
<td>19.2</td>
</tr>
<tr>
<td>Kernelized Canonical Correlation Analysis</td>
<td>15.9</td>
</tr>
<tr>
<td>DT-RNN</td>
<td><strong>12.5</strong></td>
</tr>
</tbody>
</table>
Image – Sentence Generation (!)

- 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 \( \rightarrow \) 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."
Here come the Chat Bots!

- Greatest buzz in the VC community: Conversational Agents /Dialog Systems
- Microsoft, Facebook (M), Apple (Siri), Google, WeChat, and Slack are all in it with a wave of other startups: Operator, x.ai etc
- Platforms: Chatfuel, Howdy’s Botkit and Microsoft Bot Framework

Chat Bots Models

- Retrieval-based models (easier)
  - Use a repository of predefined responses and some kind of heuristic to pick an appropriate response based on the input and context.

- Generative models (harder)
  - Generate new responses from scratch, usually based on the “encoder-decoder” framework we studied
  - Current research is on the sequence to sequence techniques
Chat Bots Models

- **Short-Text Conversations (easier)**
  - create a single response to a single input

- **Long conversations (harder)**
  - go through multiple turns and need to keep track of what has been said. Customer support conversations are typically long conversational threads with multiple questions.
Chat Bots Models

- **Closed domain (easier)**
  - the space of possible inputs and outputs is somewhat limited because the system is trying to achieve a very specific goal. E.g., Technical Customer Support or Shopping Assistants

- **Open domain (harder)**
  - the user can take the conversation anywhere, e.g., conversations on social media sites like Twitter and Reddit
  - The infinite number of topics and the fact that a certain amount of world knowledge is required to create reasonable responses makes this a hard problem.
The Challenges

• **Incorporating Context**
  • Incorporate both *linguistic context* and *physical context*.
  • General Method: embed the conversation into a vector
    • Hard for long conversations
  • **Building End-To-End Dialogue Systems Using Generative Hierarchical Neural Network Models**
  • **Attention with Intention for a Neural Network Conversation Model**
The Challenges

• **Coherent Personality**
  • Agent should ideally produce consistent answers to semantically identical inputs: “How old are you?” “What is your age?”
  • Systems learn to generate linguistic plausible responses, but they are not trained to generate semantically consistent ones (usually trained on data from multiple users!)
• **A Persona-Based Neural Conversation Model**

<table>
<thead>
<tr>
<th>message</th>
<th>response</th>
</tr>
</thead>
<tbody>
<tr>
<td>Where were you born?</td>
<td>I was born in Canada.</td>
</tr>
<tr>
<td>Where are you from?</td>
<td>England, you?</td>
</tr>
<tr>
<td>Where did you grow up?</td>
<td>I grew up in Texas.</td>
</tr>
<tr>
<td>How old are you?</td>
<td>16 and you?</td>
</tr>
<tr>
<td>What’s your age?</td>
<td>18.</td>
</tr>
</tbody>
</table>
The Challenges

• Evaluation
  • Ideally - evaluate a conversational agent by measuring whether or not it is fulfilling its task, e.g. solve a customer support problem. But too expensive to get labels....
  • Common metrics such as BLEU that are used for Machine Translation and are based on text matching aren’t well suited because sensible responses can contain completely different words or phrases
  • None of the commonly used metrics really correlate with human judgment: “How NOT To Evaluate Your Dialogue System: An Empirical Study of Unsupervised Evaluation Metrics for Dialogue Response Generation” paper
The Challenges

• **Intention and Diversity**
  - Tend to produce generic responses like “That’s great!” or “I don’t know”
  - Early versions of Google’s Smart Reply tended to respond with “I love you” to almost anything.
  - Some early work on diversity:
    - Use Minimum Mutual Information (MMI) as an optimization objective that measures the mutual dependence between inputs and outputs instead of the usual optimization of log-likelihood of target T given source S
“Most of the value of deep learning today is in narrow domains where you can get a lot of data. Here’s one example of something it cannot do: have a meaningful conversation. There are demos, and if you cherry-pick the conversation, it looks like it’s having a meaningful conversation, but if you actually try it yourself, it quickly goes off the rails.” (Andrew Ng)

• Many companies start off by outsourcing their conversations to human workers and promise that they can “automate” it once they’ve collected enough data. Might work well in closed domains.

• Learn to control your bot: “Microsoft is deleting its AI chatbot's incredibly racist tweets”