PROBABILISTIC AND DIFFERENTIABLE PROGRAMMING

V5: Embeddings

Özgür L. Özçep
Universität zu Lübeck
Institut für Informationssysteme
Today’s Agenda

From word embeddings (word2Vec) to ontology embeddings

Or: Advertising Own Work
WORD EMBEDDINGS WITH WORD2VEC
Going deep in two other ways

- Text-related applications (such as document retrieval) require a deep, here: semantical, processing of text
- Word Embeddings achieve this by representing words as vectors in a low-dimensional continuous space
- Semantical similarity of words is reflected by nearness (e.g. cosine) of the words’ vectors
- Prominent example is word2Vec (Mikolov et al. 13)
  - Learning in word2Vec based on a shallow feed forward network
  - But, the idea is deep: reading a window of words, relating each word with surrounding words (= context)

=> The idea of self-supervised learning
Let LeCun speak again

„I now call it „self-supervised learning“ because unsupervised is both a loaded and confusing term. In self-supervised learning, the system learns to predict part of its input from other parts of its input. In other words a portion of the input is used as a supervisory signal to a predictor fed with the remaining portion of the input.“

(somewhere at Facebook, cited according to Hare)
Approaches for Representing Word Semantics

Beyond bags of words

**Distributional Semantics**

*Count*
- Used since the 90’s
- Sparse word-context PMI/PPMI matrix
- Decomposed with SVD

**Word Embeddings (Predict)**
- Inspired by deep learning
  - word2vec (*Mikolov et al., 2013*)
  - GloVe (*Pennington et al., 2014*)

Underlying Theory: The Distributional Hypothesis (*Harris, ’54; Firth, ’57*)
“Similar words occur in similar contexts”

https://www.tensorflow.org/tutorials/word2vec
https://nlp.stanford.edu/projects/glove/
The Contributions of Word Embeddings

**Novel Algorithms**
*objective + training method*
- Skip Grams + Negative Sampling
- CBOW + Hierarchical Softmax
- Noise Contrastive Estimation
- GloVe
- ...

**New Hyperparameters**
*preprocessing, smoothing, etc.*
- Subsampling
- Dynamic Context Windows
- Context Distribution Smoothing
- Adding Context Vectors
- ...

What’s really improving performance?
Represent the meaning of word – word2vec

• 2 basic network models:
  – **Continuous Bag of Word (CBOW)**: use a window of words to predict the middle word
  – **Skip-gram (SG)**: use a word to predict the surrounding ones in window.

We shortly recap the CBOW mode. For SG see e.g. Goldberg/Levy 14
Word2vec – Continuous Bag of Word (CBOW)

- E.g. “The cat sat on floor”
  - Window size = 2
Input layer

Hidden layer

Output layer

Index of cat in vocabulary

one-hot vector

on
We must learn $W$ and $W'$

Input layer

$W_{V \times N}$

V-dim

V-dim

N will be the size of word vector

Hidden layer

N-dim

Output layer

$W'_{N \times V}$

sat

V-dim

We must learn $W$ and $W'$

$W_{V \times N}$

on

V-dim

V-dim
\[ W^T_{V \times N} x_{cat} = v_{cat} \]

\[ x_{on} \]

\[ W^T_{V \times N} x_{on} = v_{on} \]

\[ \hat{v} = \frac{v_{cat} + v_{on}}{2} \]
\[ \begin{align*}
W_{V \times N}^T \times x_{on} &= v_{on} \\
V_{V \times N}^T \times x_{cat} &= v_{cat} \\
\hat{v} &= \frac{v_{cat} + v_{on}}{2}
\end{align*} \]
Input layer

\[ W_{V \times N} \]

Hidden layer

\[ \hat{v} \]

Output layer

\[ \hat{y} = \text{softmax}(z) \]

\[ W'_{N \times V} \times \hat{v} = z \]

N will be the size of word vector
We would prefer \( \hat{y} \) close to \( \hat{y}_{sat} \)

\[
W'_{N \times V} \times \hat{v} = z
\]

\[
\hat{y} = \text{softmax}(z)
\]

N will be the size of word vector
We can consider either $W$ ("context-word role") or $W'$ ("middle word role") as the word’s representation.
Algebra on Words

Word Analogies

Test for linear relationships, examined by Mikolov et al. (2014)

\[ d = \arg \max_x \frac{(w_b - w_a + w_c)^T w_x}{\|w_b - w_a + w_c\|} \]

man:woman :: king:? 

+ king [ 0.30 0.70 ]
- man [ 0.20 0.20 ]
+ woman [ 0.60 0.30 ]

queen [ 0.70 0.80 ]
What is word2vec?

- Word2vec is not a single algorithm
- It is a software package for representing words as vectors, containing:
  - Two distinct models
    - CBoW
    - Skip-Gram (SG)
  - Various training methods
    - Negative Sampling (NS)
    - Hierarchical Softmax
  - A rich preprocessing pipeline
    - Dynamic Context Windows
    - Subsampling
    - Deleting Rare Words
KNOWLEDGE GRAPH EMBEDDING
Knowledge Graph (KG)

- Set of assertions in triple form
- Convenient logical notation

\[
\text{KG} = \{ (\text{subj} \ \text{rel} \ \text{obj}) \}
\]

\[
\text{KG} = \{ \text{rel}(\text{subj}, \text{obj}) \}
\]

Example

\[
\text{KG} = \{ \text{worksIn}(\text{alice}, \text{AI}), \text{subfield}(\text{AI}, \text{CS}), \text{manyPubls}(\text{alice}, \text{bob}) \} \cup \{ \text{worksIn}(\text{bob}, \text{AI}) \}
\]

- Usually KGs highly incomplete
- \(\Rightarrow\) “Learn” new triples assuming regularities in data
Knowledge Graph Embedding

Main Idea (for embedding/interpretation ( )^I )

• Learn
  1. vector representation o^I of objects o in (low-dimensional) continuous space E = \mathbb{R}^n
  2. scoring function r^I = s_r : E \times E \rightarrow \mathbb{R} for relations r

• Completion: If s_r(o_1^I, o_2^I) small, add r(o_1, o_2) to KG.

• Various approaches differing in type of s_r
  - TTransE (Bordes et al. 13), TransR (Lin et al. 15), STransE (Nguyen et al. 16)
  - DistMult (Yang et al. 15)
  - ComplEx (Trouillon et al. 16)
  - Simple (Kazemi/Poole 18)
  - RESCAL (Nickel et al. 11)
Problems of Classical Embeddings

Example (TransE (Bordes et al. 13))

- $s_r(u, v) = \|u + r - v\|$ (\(\| \cdot \|\): Euclidean Norm)
- Limitation: Relations $r =$ vector translations, hence functional

- Similar problems for other embedding approaches: not fully expressive. (Kazemi/Poole 18)
Expressivity Criterion

- Usually one considers low-dimensional spaces
- But nonetheless must be sufficiently high to embed knowledge expressed in KGs
  - $P$: set of valid triples in KG
  - $N$: set of non-valid triples in KG \( (N \cap P = \emptyset) \)

**Definition (Kazemi/Poole 18)**

An embedding model is **fully expressive** iff there are a dimension $n$, an embedding $e$, and a threshold $\lambda_R$ such that:

- for all $R(u, v) \in P$: $s_R(e(u), e(v)) \leq \lambda_R$
- for all $R(u, v) \in N$: $s_R(e(u), e(v)) > \lambda_R$
Solution: Logico-geometrical Semantics

- Represent concepts as **geometrically shaped sets** (sets of vectors, not single vector)

- Represent binary relations as **geometrically shaped sets of pairs** of objects

- Benefit: Can add background knowledge
Adding Background Knowledge

Example

\{ \forall X, Y, Z. \text{subfield}(X, Y) \land \text{worksIn}(Z, X) \rightarrow \text{worksIn}(Z, Y) \}\n
\text{KG=} \{ \text{worksIn}(alice, AI), \text{subfield}(AI, CS), \text{manyPubls}(alice, bob), \\
\text{worksIn}(bob, AI), \quad \text{(by induction)} \\
\text{worksIn}(alice, CS) \quad \text{(by deduction)} \}\n
Adding background knowledge really means a benefit...
In Favour of Controlled Explainable AI

Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead

Cynthia Rudin

NOT: Explain parameter tweaking
BUT: More control via background knowledge/ontology
Research program

Find appropriate pairs

background knowledge language / geometrical structure for

knowledge graph embedding

- Quasi-chained Datalog / convex (Gutierrez-Basulto, Schockaert 18)

- $\mathcal{EL}^{++} /$ spheres (Kulmanov et al. 19)

- $\mathcal{ALC} /$ convexal-cones (this lecture)

- $L_{cone} /$ convex cones (this lecture)
Agenda for the Following

- Description Logics
- Convex Cones
- Faithful Embedding for Al-cones
- Towards a Logic of Convex Cones
DESCRIPTION LOGICS
Definition (Description logics (DLs))

Logics for use in knowledge representation with special attention on a good balance of expressibility and feasibility of reasoning services

- Can be mapped to fragments of FOL

- Usage
  - Ontology representation language
  - Foundation for standard web ontology language (OWL)

- Have been investigated for ca. 30 years now
  - Many theoretical insights on various different purpose DLs
  - Various reasoners
### Description Logics (DLs)

#### Example (ALC Concepts)

- **Students**
  - “Students”

- **Students \( \sqcap \) Male**
  - “Male students”

- **Female \( \sqcap \) Male**
  - “Female or Male”

- **\( \exists \) attends. LogicCourse**
  - “Those attending a logic course”

- **\( \neg \exists \) attends. LogicCourse**
  - “Those not attending a logic course”

---

**Full concept negation allowed in DL ALC.**
Tbox and Abox

• **Terminological box (tbox):** \{ concept inclusions \}

  **Example**

  \{ MasterStudent \sqsubseteq Student, Student \sqcap Professor \sqsubseteq \bot \}\n
• **Assertional box (abox):** \{ assertions \}

  **Example**

  \{ MasterStudent(peter), attends(peter, CS4711) \}

• **Ontology:** tbox \cup abox
Example (Certain Answers for Conjunctive Queries)

- \( T = \{ \top \subseteq \text{Male} \cap \text{Female}, \text{Male} \cap \text{Female} \subseteq \bot \} \)
- \( A = \{ \text{friend}(\text{john}, \text{susan}), \text{friend}(\text{john}, \text{andrea}), \text{female}(\text{susan}), \text{likes}(\text{susan}, \text{andrea}), \text{likes}(\text{andrea}, \text{bill}), \text{Male}(\text{bill}) \} \)
- \( Q(x) = \exists y, z (\text{friend}(x, y) \land \text{Female}(y) \land \text{likes}(y, z) \land \text{Male}(z)) \)

- \( \text{cert}(Q(x), O) =? \)

- We have to consider all possible models of the ontology
- But here there are actually two classes: Andrea is male vs. Andrea is not male.
Example (Certain Answers for Conjunctive Queries)

- $T = \{ \top \subseteq \text{Male} \cap \text{Female}, \ \text{Male} \cap \text{Female} \subseteq \bot \}$
- $A = \{ \text{friend}(\text{john}, \text{susan}), \ \text{friend}(\text{john}, \text{andrea}), \ \text{female}(\text{susan}), \ \text{likes}(\text{susan}, \text{andrea}), \ \text{likes}(\text{andrea}, \text{bill}), \ \text{Male}(\text{bill}) \}$
- $Q(x) = \exists y, z(\text{friend}(x, y) \land \text{Female}(y) \land \text{likes}(y, z) \land \text{Male}(z))$

$\text{cert}(Q(x), O) = \{ \text{john} \}$
CONVEX CONES
Idea: Interpret Concepts by Convex Cones

- $X \subseteq \mathbb{R}^n$ is a convex cone iff for all
  $$v, w \in X, \lambda, \mu \in \mathbb{R}_{\geq 0}: \lambda v + \mu w \in X$$
- $\text{hull}(X) = \text{smallest cone containing } X$
Idea: Interpret Concepts by Convex Cones

- $X \subseteq \mathbb{R}^n$ is a convex cone iff for all
  \[ v, w \in X, \lambda, \mu \in \mathbb{R}_{\geq 0}: \lambda v + \mu w \in X \]
- $\text{hull}(X)$ = smallest cone containing $X$
- Polar cone
  \[ X^\circ = \{ v \in \mathbb{R}^n | \forall w \in X, v \cdot w \leq 0 \} \]

Proposition

For closed convex cones $X, Y$:

- $X^\circ$ is a closed convex cone
- $(X^\circ)^\circ = X$
- $\text{hull}(X \cup Y) = (X^\circ \cap Y^\circ)^\circ$

Why convex cones?

- $\neg X = X^\circ$
  \[ X \cap Y = X \cap Y \]
  \[ X \cup Y = \text{hull}(X \cup Y) \]
- Convex/conic optimization
Not all Cones Appropriate for ALC

\[ A \sqcup B = \text{hull}(A \cup B) \]

\[ C \cap (A \sqcup B) = C \neq (C \cap A) \sqcup (C \cap B) = \bot \]

- Distributivity law not fulfilled
- What should we do?
  1. Restrict family of cones (in the following)
  2. Search for a (the) logic of cones (thereafter)
FAITHFUL EMBEDDING FOR AL-CONES
Searching for ALC-cones (Nomen est Omen)

Definition (Axis-aligned Cone (al-cone))

\[ X \text{ is an al-cone in } \mathbb{R}^n \text{ iff } X = X_1 \times \cdots \times X_n \]
where each \( X_i \in \{ \mathbb{R}, \mathbb{R}^+, \mathbb{R}^-, \{0\} \} \)
Example (Embedding for \( tbox \{ A \sqsubseteq B \} \))
Al-Cones are Appropriate for Boolean $\mathcal{ALC}$

**Proposition (Ö., Leemhuis, Wolter 2020)**

For Boolean$^1$ $\mathcal{ALC}$ -ontologies:

\[
\text{classically satisfiable} = \text{satisfiable with faithful al-cone model}
\]

What does faithfulness mean here?

---

$^1$No roles (binary relations) allowed.
Example (Embedding for tbox \{ A \sqsubseteq B \})
Example (Embedding for $\text{tbox} \{ A \sqsubseteq B \}$)

Our geometric models are **partial**: can model uncertainty in ontology ($a_2$ not known to be $A$ or $\neg A$; in contrast: $a_1$ completely determined)
Example (Embedding for $\text{tbox} \{ A \sqsubseteq B \})$}

```
\begin{align*}
\neg A \land B & \\
\neg A & \\
\neg B & \\
A & \\
\neg A & \\
\neg B & \\
A \sqcup \neg B &
\end{align*}
```

**Faithfulness:** $o^l \in C^l$ iff ontology entails $C(o)$. 
Al-Cone models for full $\mathcal{ALC}$

**Proposition (Ö., Leemhuis, Wolter 2020)**

For possibly non-Boolean $\mathcal{ALC}$-ontologies up to some rank$^1$:

classically satisfiable $= \text{satisfiable with faithful al-cones model}$

- This is an approximative solution
- Relation not interpreted geometrically (not conic)

---

$^1$ Rank = nesting depth of quantifiers in a formula
TOWARDS A LOGIC OF CONES
Searching for an Appropriate Logic

**Definition (Minimal orthologic (Goldblatt 74))**

Minimal orthologic is characterised by $\land, \lor$ fulfilling rules of a lattice and the existence of an orthonegation $\neg$

- If $A \subseteq B$ then $\neg B \subseteq \neg A$
- $\neg \neg A \subseteq A$
- $A \land \neg A \subseteq \bot$

**Proposition (Leemhuis, Ö., Wolter 2020)**

*Convex cones fulfil the rules of minimal orthologic.*

- Hence: $L_{cone}$ must extend minimal orthologic
- Which additional rules hold?
Known Weakenings of Distributivity are Falsified

Definition (Orthomodularity)

If $B \subseteq A$ then $A \cap (\neg A \cup B) \subseteq B$

$B \subseteq A$, but $A \cap (\neg A \cup B) = A \neq B$
Definition (Partial Orthomodularity (pOM))

If $B \subseteq A$ and $A \subseteq B \sqcup (A \cap \neg B)$ then $A \cap (\neg A \sqcup B) \subseteq B$

Theorem (Leemhuis, Ö., Wolter 2020)

Convex cones fulfill (pOM).
Subalgebra Theorem

\( MC_8: \)

\[
\begin{array}{c}
\neg A \sqcup B \\
\neg B \\
A \\
\neg D \\
D \\
\neg A \\
A \cap \neg B \\
B \\
\bot
\end{array}
\]

If \( B \sqsubseteq A \)
and \( A \sqsubseteq B \sqcup (A \cap \neg B) \)
But \( A \cap (\neg A \sqcup B) = D \neq B \)

Theorem (Leemhuis, Ö., Wolter 2020)

A logic fulfils (pOM) iff it does not contain \( MC_8 \).
Open questions

• $L_{cone} = L_{pOM} + ? / \text{convex cones}$

• ML implementation with (al)-cones$^1$

---

1) (Leemhuis, Ö., Wolter ICCS 20)
Uhmm, a lecture with a hopefully useful

APPENDIX
Color Convention in this course

- Formulae, when occurring inline
- Newly introduced terminology and definitions
- Important results (observations, theorems) as well as emphasizing some aspects
- Examples are given with standard orange with possibly light orange frame
- Comments and notes
- Algorithms
Today’s lecture is based on the following

- Jonathon Hare: Lecture 14 of course „COMP6248 Differentiable Programming (and some Deep Learning)“
  http://comp6248.ecs.soton.ac.uk/
- Möller/Özcep: Lecture „Word semantics and Latent Relational Structures“ of Course „Web-Mining Agents“
- Özcep: Knowledge Graph Embeddings, Talk at KI-Kolloqium
  https://www.ifis.uni-luebeck.de/~moeller/KI-Kolloquium/2019-12-02-Oezcep.pdf
References

References


