Contextual Distributed Models for Sequences
aka Latent parameterizations, Family Learning, ...

Benjamin Piwowarski
LIP6, Université Pierre et Marie Curie / CNRS

9th of November, 2015
Introduction
Importance of context

- Non generative process
  - *User, Location* in Personalized Web Search

- Generative process
  - *Speaker* in spoken language understanding
  - *Topic and style* in document generation
  - *User* in web browsing
  - *User and Activity* in music listening
  - ...
Distributed Representations

- Allow to capture geometrically dependencies between entities (users, objects, etc.)
- Used successfully in image, text, etc.
- Able to take into account various sources of information
This talk

- Taking into account context
  - Learn $p(x|\mathcal{D})$ on a dataset $\mathcal{D}$
  - Learn $p(x|\mathcal{D}_c)$ for a given dataset $\mathcal{D}_c \ll \mathcal{D}$, taking into account $p(x|\mathcal{D})$

- ...using distributed representations

- Goal: learn a representation that can be reused
  - Predicting search results for a user in a given context
  - Predicting the satisfaction of a user
  - Predicting the next recommendation to a user
Taking into account Context
Overview

Different ways to take into account parameters

- Explicit factors
- Families of parameters
- Prior on parameters
Probabilistic models

Latent variables describing the context

Latent Dirichlet Allocation (text)

Word generation is governed by the themes the document deals with:

\[ \theta \sim \text{Dirichlet} (\alpha) \]
\[ z \sim \text{Multinomial}(\theta) \]
\[ w \sim \text{Multinomial}(z) \]

Problem of estimating (the maximum likelihood) of

\[ p(\theta|d_1 \ldots d_n) \]

“Latent dirichlet allocation,” Ng, Blei, and Jordan (2003)
Family Discovery

- **Mixture of experts**
  1. Learn a mixture of experts over the dataset
     \[ p(x) = \sum_i p(x|M_i) p(M_i) \]
  2. Estimate \( p(M_i|D_c) \)

- Learn families of models as a subset \( S \) of the parameter space
  1. Single Model
  2. Separate Models
  3. Affine subspace (determined by PCA)
  4. Affine patch
  5. Coupled Map (auto-encoder)

and project a new solution onto the family space

“Family Discovery,” *Omohundro (1995)*
Parameterizing probabilistic models

- Global idea:
  - Parametric model
  - Modify some parameters using context-specific parameters
Parameterizing probabilistic models

Gesture recognition with HMMs
The gesture is parameterized with $\theta$

$$\hat{\mu}_j = W_j \theta + \mu$$

$$p(x_t|q_t = j, \theta) = \mathcal{N}(x_t; \hat{\mu}_j(\theta), \Sigma_j)$$

where $x_t$ is a 6-dimensional vector (position of each hand); $\theta$ is re-estimated for each example


Extension of Wilson and Bobick to parameterization of variance $\Sigma_j(\theta)$ in handwriting recognition

“Handling signal variability with contextual markovian models,” Radenen and Artieres (2014)
Regularization

- Learning a model on a dataset $\mathcal{D}$
  
  \[ \theta^* = \arg\min_{\theta} R(\theta; \mathcal{D}) \]

- Contextualization: the parameters are used as a “prior” for the new optimization
  
  \[ \theta_c = \arg\min_{\theta} R(\theta; \mathcal{D}) + \Delta(\theta, \theta^*) \]
Regularization: example

**Truncated gradient**

Adapting Deep RankNet for Personalized Search

\[ w \leftarrow w - \eta T_1 \left( \frac{\partial C}{\partial w}, a, \tau \right) \]

where \( C \) is the cost, \( a \) is the output of the neuron, and \( \tau \) is a threshold.

\[ T_1 (v, a, \theta) = \begin{cases} 
\max (0, v - a) & v \in [0, \tau] \\
\min (0, v + a) & v \in [-\tau, 0] \\
v & \text{otherwise}
\end{cases} \]

Adapting deep RankNet for personalized search, Song, Wang, and He (2014)
Representation Learning and Sequences
**Representation Learning**

- Projecting points from the original point to a manifold
- Hypotheses
  
  
  - *The semi-supervised learning hypothesis*
    Learning $p(x)$ can improve our classification $p(y|x)$
  
  - *The (unsupervised) manifold hypothesis*
    Real world data presented concentrate in the vicinity of non-linear sub-manifolds
  
  - *The manifold hypothesis for classification*
    Points of different classes concentrate along different sub-manifolds
Representation Learning

- Represent complex objects (as vectors in $\mathbb{R}^n$)
  - Nodes in a graph
  - Sequences (state as a vector)
  - Context

Examples:
- Language Models
- Spoken Language Understanding
- Handwriting
- ...
Distributed States (traffic prediction)

\[
\mathcal{L}(\theta) = \frac{1}{O} \sum_{it} m_i \Delta \left( f_\theta \left( z_i^{(t)} \right), x_i^{(t)} \right) \\
+ \lambda_1 \sum_{it} \left\| z_i^{(t+1)} - h_\gamma \left( z_i^{(t)} \right) \right\| \\
+ \lambda_2 \sum_{ijt} e_{ij} \left\| z_i^{(t)} - z_j^{(t)} \right\|
\]

Handwriting sequences

\[ p(x_{t+1}|s_t) = \sum_{j=1}^{M} \pi^j_t \mathcal{N}(x_{t+1}|\mu^j_t, \sigma^j_t, \rho^j_t) \]

\[ \hat{y}_t = (\hat{e}_t, \{\hat{\pi}^j_t, \hat{\mu}^j_t, \hat{\sigma}^j_t, \hat{\rho}^j_t\}) = b_y + \sum Ws \]

\[ y_t = f(\hat{y}_t) \]

“Generating Sequences With Recurrent Neural Networks,” Graves (2013)
Neural Network Language Models

1. Lookup table: a word $w \leftrightarrow a_w \in \mathbb{R}^p$

2. Computing the state
   1. Convolution $s_k = f (w_{k-d} \ldots w_{k-1}) \in \mathbb{R}^k$
   2. Recursion $s_k = f (s_{k-1}, w_{k-1})$

3. Probability distribution over words $p (w|s_k) \propto \exp \left( \langle b_w, s_k \rangle \right)$

“Natural language processing (almost) from scratch,” Collobert et al. (2011)
Taking context into account in distributed models
Priming

Take the breath away when they are
when the network is primed
and biased, it writes
in a cleaned up version
of the original style

She looked closely as she
when the network is primed
and biased, it writes
in a cleaned up version
of the original style

```
B. Piwowarski (LIP6/CNRS)
```

“Generating Sequences With Recurrent Neural Networks,” Graves (2013)
Taking context into account in distributed models

Two approaches

Parameterized LM

Estimates

\[ p(w_k | w_{k-l}, \ldots, w_{k-1}, \theta_d) \]

“Distributed Representations of Sentences and Documents,” *Le and Mikolov (2014)*
Problem

- Given a family of functions $F$ and a context space $C$, we want to find a function $\Phi$ such that

$$\Phi : F \times C \rightarrow F$$

$$f \times \theta_c \rightarrow f_c$$

- Neutral element $\theta_e$ such that $\Phi (f, \theta_e) = f$ (we want to learn a general model)

- Problems:
  - What class of functions
  - Balance between complexity and generalization
Inputs

- Sets of documents

  ... “I don’t have a vendetta against Bambi. I really don’t give a darn. It was just my personal opinion,” Louise Bates Ames, associate director of the Gesell Institute of Human Development, said Wednesday.

- A query

  Document will report judicial proceedings and opinions on contracts for surrogate motherhood.

Goal: rank relevant documents before non relevant ones

Measure: Mean Average Precision over queries

\[
AP = \sum_{k=1}^{N} \delta \left( d_k \text{ is relevant} \right) \frac{\sum_{i \leq k} \delta \left( d_i \text{ is relevant} \right)}{k}
\]

precision at rank \( k \)
Language Models

Language Models Hypothesis (Information Retrieval)

The query should be generated by the document language model

\[ p(q_1 \ldots q_n | M_{d_1}) > p(q_1 \ldots q_n | M_{d_2}) \]
\[ \iff d_1 \ is \ more \ relevant \ than \ d_2 \]

- Unigram model

\[ p(w_1 \ldots w_n | M_d) = H \prod_i p(w_i | M_d) = ML \prod_i \frac{tf(w_i, d)}{tf(d)} \]
Limits

- Dependencies are not taken into account
  Extending to higher order Markov chain is not straightforward

\[ p(w_1 \ldots w_n | \mathcal{M}_i) = \prod_i p(w_i | w_{i-1}, \ldots, w_{i-k}, \mathcal{M}) \]

- Problem of vocabulary mismatch

Both problems can be tackled using distributed representations
Parameterized LM

\[ \phi \{ P(w | w_1, \ldots, w_{n-1}) \} \]

Possible transformations:
- (Le and Mikolov) Component-Wise Sum
- Component-Wise Product
Procedure

- Initialization with Mikolov word2vec embeddings
- Train the general/collection language model

\[ \theta^* = \arg\max_{\theta} \sum \log p_{nn}(w_1 \ldots w_n | \theta, z_d = e) \]

- For each document \( d \) (pre-selection with a standard IR model)
  - Compute its representation

\[ z_d^* = \arg\max_{\theta} \sum \log p_{nn}(d_1 \ldots d_n | \theta, z_d = e) \]

  - Compute the score for the query

\[ RSV(q, d) = \lambda p_{nn}(q_1 \ldots q_n | \theta, z_d) + (1 - \lambda) p_{unigram}(q_1 \ldots q_n | d) \]
Results

- Qualitative analysis: *underestimated* probability of document-specific terms
- Looking at alternative transformations:

TREC1 (similar for others) / sum or product
Translations

- Translation Transformations

\[ ld + \sum_{i=1}^{K} a_i b_i^\top \]

Example: \[ ld + \frac{1}{\|a\|} (t - a) a^\top \]

- Preliminary experiments have shown that translation > product/sum
Conclusion
Conclusion

- Distributed representations are useful to capture context
- Context = manipulation of the state in sequences

Future work

- Understand what the different transformations do
- Recurrent models (the modified state is taken into account)
- Other applications: user modeling (Web)


