SparseMAP: Differentiable Sparse Structured Inference

Presented by Vlad Niculae
Joint work with André FT Martins, Mathieu Blondel, Claire Cardie

poster #66 tonight
github.com/vene/sparsemap
Structured Inference

* I like it

* I like it

* I like it
Structured Inference

...
Structured Inference

PRON  VERB  NOUN
I  like  it

PRON  VERB  PRON
I  like  it

PRON  VERB  ADJ
I  like  it

I  cela
like  me
it  plaît

I  cela
like  me
it  plaît

I  cela
like  me
it  plaît
Structured Inference
(Latent) Structured Inference
\[ M = \text{conv}(Y) \text{ where } Y = \{ \, \} \]
SparseMAP

Efficient & simple to:
- compute
- back-propagate

Useful as:
- hidden layer
- output layer
Deriving SparseMAP
Structured Inference as argmax

input
argmax

input

\[ \frac{\partial p}{\partial x} ? \]
$\partial p / \partial x$?
argmax $\rightarrow$ softmax

$p_i = \exp x_i / Z$

$\partial p / \partial x$ ?

input
argmax $\Rightarrow$ softmax

\[ \frac{\partial p}{\partial x} \]
\( \text{dim}(x) = \text{number of possible structures!} \)

(exponentially large)

\( \frac{\partial p}{\partial x} ? \)

[Martins and Astudillo, 2016]
[Niculae and Blondel, 2017]
\[ x = A^T \eta \quad \in \mathbb{R}^k \quad \in \mathbb{R}^d \quad k \gg d \]
$x = A^T \eta$

* I like it
\[ x = A^T \eta \]
\[ x = A^T \eta \]
$x = A^T \eta$

$\mu = Ap$
\[
\text{argmax } \langle x, p \rangle \\
\text{s.t. } p \in \Delta
\]

\[
p^* = e_i \text{ where } i = \text{argmax}(x)
\]

\[
x = A^T\eta
\]

\[
\Delta
\]

\[
\text{argmax } \langle \eta, \mu \rangle \\
\text{s.t. } \mu \in M
\]

\[
\mu = Ap
\]

\[
M
\]
MAP inference:
Maximum spanning tree (Chu-Liu/Edmonds)

Hungarian algorithm
argmax $\langle x, p \rangle$
\[ \text{s.t. } p \in \Delta \]

argmax $\langle x, p \rangle + H(p)$
\[ \text{s.t. } p \in \Delta \]

softmax, closed-form solution: $p^* = \exp(x)/Z$

argmax $\langle \eta, \mu \rangle$
\[ \text{s.t. } \mu \in M \]

argmax $\langle \eta, \mu \rangle + H(\mu)$
\[ \text{s.t. } \mu \in M \]

Structured attention networks
[Kim et al, 2017],
[Liu et al, 2017]

$x = A^{\top} \eta$

$\mu = Ap$

$\Delta$

$M$
### MAP inference: Maximum spanning tree

### Marginal inference: Matrix-Tree theorem

<table>
<thead>
<tr>
<th>*</th>
<th>*</th>
<th>like</th>
<th>it</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

\[ A = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \]

### Hungarian algorithm

### #$P$ complete

<table>
<thead>
<tr>
<th>I</th>
<th>like</th>
<th>cela</th>
</tr>
</thead>
<tbody>
<tr>
<td>it</td>
<td>me</td>
<td>plait</td>
</tr>
</tbody>
</table>

\[ \eta = \begin{bmatrix} .3 \\ .8 \\ -.5 \\ .2 \\ -.1 \end{bmatrix} \]

<table>
<thead>
<tr>
<th>I</th>
<th>- cela</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>- me</td>
</tr>
<tr>
<td>I</td>
<td>- plait</td>
</tr>
</tbody>
</table>

\[ \eta = \begin{bmatrix} .3 \\ .8 \\ -.5 \\ .2 \\ -.1 \end{bmatrix} \]

<table>
<thead>
<tr>
<th>I</th>
<th>like - cela</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>like - me</td>
</tr>
<tr>
<td>I</td>
<td>like - plait</td>
</tr>
</tbody>
</table>

\[ \eta = \begin{bmatrix} .3 \\ .8 \\ -.5 \\ .2 \\ -.1 \end{bmatrix} \]

<table>
<thead>
<tr>
<th>I</th>
<th>like - it</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>it - cela</td>
</tr>
<tr>
<td>I</td>
<td>it - me</td>
</tr>
<tr>
<td>I</td>
<td>it - plait</td>
</tr>
</tbody>
</table>

\[ \eta = \begin{bmatrix} .3 \\ .8 \\ -.5 \\ .2 \\ -.1 \end{bmatrix} \]
\[
\begin{align*}
\text{argmax } & \langle x, p \rangle \\
\text{s.t. } & p \in \Delta \\
\end{align*}
\]

\[
\begin{align*}
\text{argmax } & \langle x, p \rangle + H(p) \\
\text{s.t. } & p \in \Delta \\
\end{align*}
\]

\[
\begin{align*}
\text{argmax } & \langle \eta, \mu \rangle \\
\text{s.t. } & \mu \in M \\
\end{align*}
\]

\[
\begin{align*}
\text{argmax } & \langle \eta, \mu \rangle + H(\mu) \\
\text{s.t. } & \mu \in M \\
\end{align*}
\]

\[x = A^T \eta\]

\[\mu = Ap\]
argmax $\langle x, p \rangle$
  s.t. $p \in \Delta$

$\arg\max \langle x, p \rangle + H(p)$
  s.t. $p \in \Delta$

argmax $\langle x, p \rangle - \frac{1}{2} \|Ap\|^2$
  s.t. $p \in \Delta$

argmax $\langle \eta, \mu \rangle$
  s.t. $\mu \in M$

$\arg\max \langle \eta, \mu \rangle + H(\mu)$
  s.t. $\mu \in M$

$\arg\max \langle \eta, \mu \rangle - \frac{1}{2} \|\mu\|^2$
  s.t. $\mu \in M$

$x = A^T \eta$

$\mu = Ap$
Efficiently Computing Sparse MAP

\[ \arg\max \langle \eta, \mu \rangle - \frac{1}{2} \| \mu \|^2 \]
\[ \text{s.t. } \mu \in M \]

**Forward Pass:**
- Active Set algorithm
- only accesses \( M \) through MAP calls
- linear & finite convergence

**Backward Pass:**
\[ \frac{\partial \mu^*}{\partial \eta} \]
- Linear in \( \dim(M) \)
- and in \# selected structures

QP with exponentially many vertices!
Sparse Latent Structure
Natural Language Inference

Prem: A gentleman overlooking a neighborhood situation.

Hypo: A police officer watches a situation closely.

(P, H) { entailment, contradiction, neither }
Natural Language Inference

Prem: A gentleman overlooking a neighborhood situation.

Hypo: A police officer watches a situation closely.

(P, H) { entailment contradiction neither

A gentleman overlooking ...
A police officer watches ...

Model: ESIM [Chen & al, 2017]
Natural Language Inference

Prem: A gentleman overlooking a neighborhood situation.

Hypo: A police officer watches a situation closely.

\[(P, H)\]

\[
\begin{align*}
&\text{A gentleman overlooking ...} \\
&\text{A police officer watches ...}
\end{align*}
\]

\{ entailment \text{ contradiction} \neither \}

Model: ESIM [Chen et al, 2017]
# Natural Language Inference

<table>
<thead>
<tr>
<th>a gentleman</th>
<th>a neighborhood</th>
<th>a situation</th>
</tr>
</thead>
<tbody>
<tr>
<td>police officer watches</td>
<td>a situation closely</td>
<td>a situation closely</td>
</tr>
</tbody>
</table>

![Diagram](image.png)
Natural Language Inference with Linear Assignment

Prem: A gentleman overlooking a neighborhood situation.

Hypo: A police officer watches a situation closely.
SNLI

87%
86%
85%
84%
83%
softmax  matching  sequence

MultiNLI

76.5%
76.0%
75.5%
75.0%
74.5%
74.0%
softmax  matching  sequence

(3-class accuracy)
Natural Language Inference
with Linear Assignment
Natural Language Inference with Linear Assignment
Sparse Structured Output Prediction
Sparse Structured Output Prediction

SparseMAP loss

\[ L_A(\eta, \mu) = \max_{\mu \in M} \left\{ \langle \eta, \mu \rangle - \frac{1}{2} \|\mu\|^2 \right\} \]

\[ - \langle \eta, \mu^* \rangle + \frac{1}{2} \|\mu\|^2 \]

margin-SparseMAP loss

\[ L_A^\rho(\eta, \mu) = \max_{\mu \in M} \left\{ \langle \eta, \mu \rangle - \frac{1}{2} \|\mu\|^2 + \rho(\mu, \bar{\mu}) \right\} \]

\[ - \langle \eta, \mu^* \rangle + \frac{1}{2} \|\mu\|^2 \]

Instance of a structured Fenchel-Young loss, like CRF, SVM, etc.

[Blondel, Martins, Niculae ‘18]
Dependency Parsing with bi-LSTM features

[Kiperwasser & Goldberg, 2016]
Sparse Structured Output Prediction
As models train, inference gets sparser!
Sparse Structured Output Prediction

Inference captures linguistic ambiguity!

They did a vehicle wrap for my Toyota Venza that looks amazing.
Sparse Structured Output Prediction

Inference captures linguistic ambiguity!

- the broccoli looks browned around the edges.
Sparse Latent Structure

A gentleman overlooking a neighborhood situation

Sparse Structured Output Prediction

A police officer watches...

* the broccoli looks browned around the edges.

poster #66 tonight @6:15

github.com/vene/sparsemap

https://vene.ro

vnfrombucharest