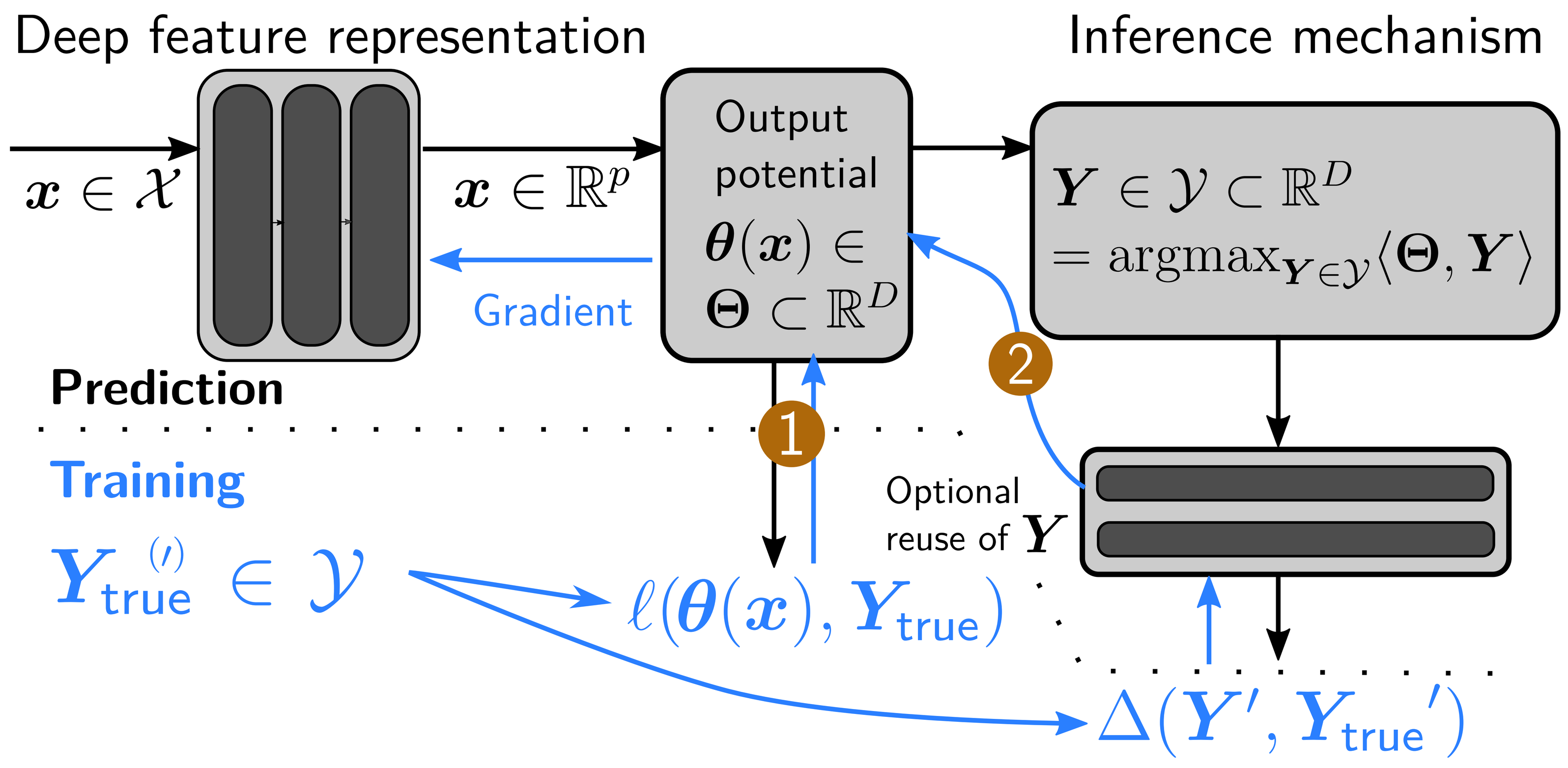
Differentiable Dynamic Programing for Structured Prediction and Attention

Arthur Mensch[1] Mathieu Blondel[2]

[1]Parietal team, Inria, Saclay, France

[2]NTT Communication Science Laboratories, Kyoto, Japan

1 – Context – End-to-end deep structured prediction



- $x \in \mathcal{X}$: text sequence, audio time-series
- $Y \in \mathcal{Y}$: tag sequence, alignment between time-series, **attention map**
- Losses typically involve LP solution — baseline structured perceptron

$$\ell(\theta(x), Y_{\text{true}}) = \max_{Y \in \mathcal{Y}} \langle Y, \theta \rangle - \langle Y_{\text{true}}, \theta \rangle$$

- Easiest case: $\theta(x) = Wx + b$, $\mathcal{Y} = [1, k]$
- Often in machine learning: max and argmax from dynamic programing

Dynamic programing

$(\text{arg})\max_{Y \in \mathcal{Y}} \langle \theta(x), Y \rangle$ breaking $\mathcal{Y}$ into smaller sets

- Linear vs exp complexity
- Inference on tree CRF / DTW

End-to-end training

1. Loss functions

$\Rightarrow$ Backpropagate through LP value (max)

2. Differentiable inference mechanism

$\Rightarrow$ Backpropagate through LP solution (argmax)

2 – Contrib – Smoothing DP for end-to-end training

- **Dynamic programing value/solution become differentiable layer**
- End-to-end training for many supervised prediction problems
- Easy to implement custom backward module, with probabilistic interpreta°
- Optional **sparsity** in Y $\Rightarrow$ small set of best guesses / sparse attention

3 – Generic DP on a directed acyclic graph

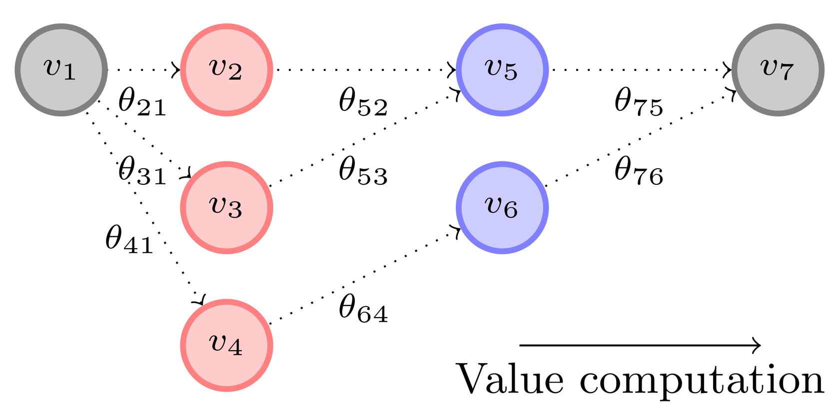
- Find shortest path on a graph G. Edge (i, j) has weight $\theta_{i,j}$ — j parent, i child
- Path $Y \in \mathcal{Y} \subset \{0,1\}^{N \times N}$: $y_{i,j} = 1$ iff (i, j) is taken, $\theta \in \mathbb{R}^{N \times N}$ adj. matrix

- **Bellman equation**

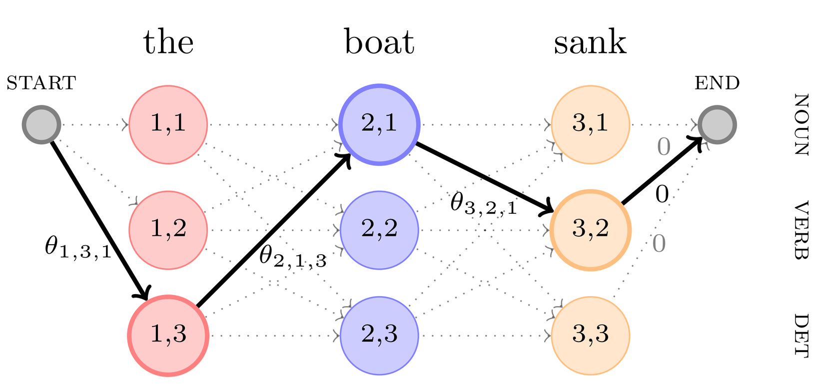
$$v_i(\theta) = \max_{j \in \mathcal{P}_i} \theta_{i,j} + v_j(\theta), \mathrm{DP}(\theta) \triangleq v_n(\theta)$$

Highest score among all paths:

$$\mathrm{LP}(\theta) = \max_{Y \in \mathcal{Y}} \langle Y, \theta \rangle = \mathrm{DP}(\theta)$$



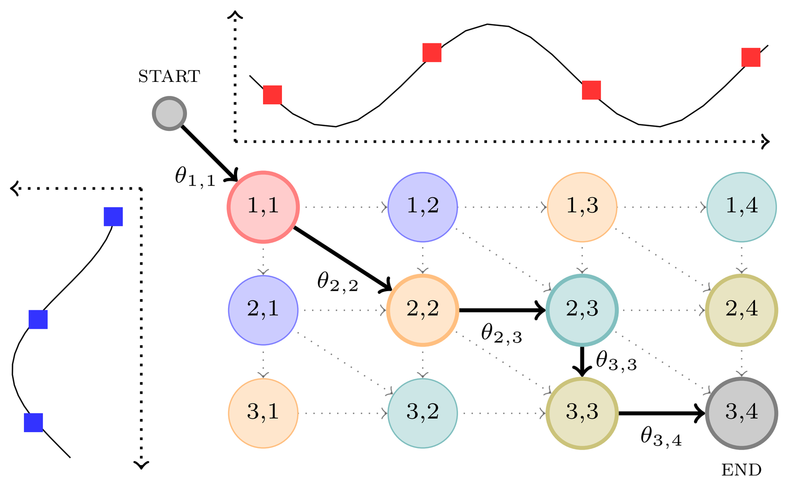
Viterbi decoder



$$\langle Y, \theta \rangle = \theta_{1,3,1} + \theta_{2,1,3} + \theta_{3,2,1}$$

$$\mathrm{DP}(\theta) = \mathrm{Vit}(\theta)$$

Dynamic time warping



$$\langle Y, \theta \rangle = \theta_{1,1} + \theta_{2,2} + \theta_{2,3} + \theta_{3,3} + \theta_{3,4}$$

$$\mathrm{DP}(\theta) = \mathrm{DTW}(\theta)$$

Lampe, Guillaume et al. (2016). "Neural Architectures for Named Entity Recognition". In: *Proc. of NAACL*, pp. 260–270.
LeCun, Yann et al. (2006). "A tutorial on energy-based learning". In: *Predicting structured data* 1.0.
Martins, André F.T. and Ramón Fernandez Astudillo (2016). "From softmax to sparsemax: A sparse model of attention and multi-label classification". In: *Proc. of ICML*, pp. 1614–1623.
Mensch, Arthur and Mathieu Blondel (2018). "Differentiable Dynamic Programming for Structured Prediction and Attention". In: *Proceedings of the International Conference on Machine Learning*.
Nesterov, Yurii (2005). "Smooth minimization of non-smooth functions". In: *Mathematical Programming* 103.1, pp. 127–152.
Niculae, Vlad et al. (2018). "SparseMAP: Differentiable sparse structured inference". In: *Proceedings of the International Conference on Machine Learning*.
Pearlmutter, Barak A. (1994). "Fast exact multiplication by the Hessian". In: *Neural computation* 6.1, pp. 147–160.

Backpropagation

Danskin theorem

$$\partial \mathrm{DP}(\theta) = \mathrm{conv}(\mathrm{argmax}_{Y \in \mathcal{Y}} \langle Y, \theta \rangle)$$

- Not defined everywhere, not smooth, piecewise constant
- Need regularization

4 – DP smoothing

- $\Omega : \mathbb{R} \to \mathbb{R}$ str-convex f (Nesterov, 2005).

$\max_\Omega(x) = \max_{y \in \Delta^d} \langle x, y \rangle - \sum_{i=1}^d \Omega(y_i)$

- **Bottom-up Bellman smoothing**

$\mathrm{DP}(\theta) \to \mathrm{DP}_\Omega(\theta)$, $\partial \mathrm{DP}(\theta) \to \nabla \mathrm{DP}_\Omega(\theta)$

- Vs generally intractable $\mathrm{LP}_\Omega(\theta)$
- Argmax backprop: $\nabla^2 \mathrm{DP}_\Omega(\theta)$

Regularizations have different properties

- **Negentropy** $-H$: $\Omega(p) = p \log(p)$
- $\nabla \max_\Omega(x) = \mathrm{softmax}(x)$
- $\nabla \mathrm{DP}_\Omega$ expected Gibbs distribution

- ℓ_2^2: $\Omega(p) = p^2$ (Martins et al., 2016)
- $\nabla \max_\Omega(x) = \mathrm{sparsemax}(x)$
- Projection on $\triangle^d$: $\nabla \mathrm{DP}_\Omega$ **sparse**

5 – Theoretical properties

DP value layer

- DP_Ω is convex, $\mathrm{LP} - \mathrm{DP}_\Omega$ bounded
- $\mathrm{DP}_\Omega(\theta) = \mathrm{LP}_\Omega(\theta)$ **if and only if** $\Omega = -\gamma H$, where $\gamma \geq 0$

DP gradient layer

- Path distribution $p_{\theta,\Omega}$ (over $\mathcal{Y}$) s.t. $\nabla \mathrm{DP}_\Omega(\theta) = \mathbb{E}_{\theta,\Omega}[Y] \in \mathrm{conv}(\mathcal{Y})$

$=$ Graph random walk, easy to sample

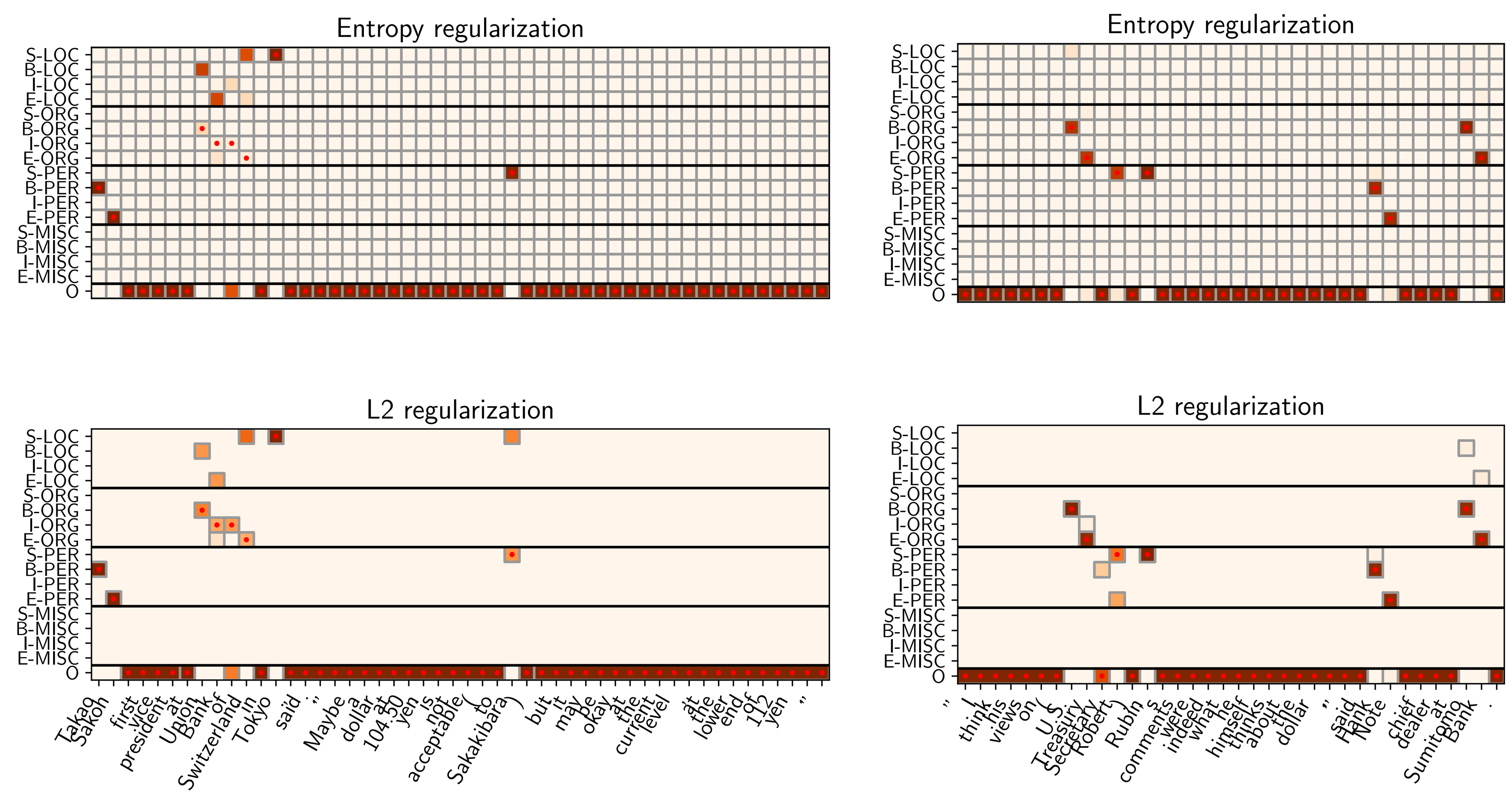
6 – Experiments – Named entity recognition (CoNLL'03)

- Word embedding + word LSTM
- Linear chain CRF-like potential

$$\theta(X)_{t,i,j} = w_i^T x_t + b_i + t_{i,j}$$

- Sparse prediction $\nabla \mathrm{DP}_{\ell_2^2}(\theta)$

Ω	Loss	English	Spanish	German	Dutch
Negent.	Surrogate	90.80	**86.68**	77.35	**87.56**
	Relaxed	90.47	86.20	**77.56**	87.37
ℓ_2^2	Surrogate	**90.86**	85.51	76.01	86.58
	Relaxed	89.49	84.07	76.91	85.90
0	Struct. perceptron	86.52	81.48	68.81	80.49
Lample et al., 2016		*90.96*	*85.75*	*78.76*	*81.74*

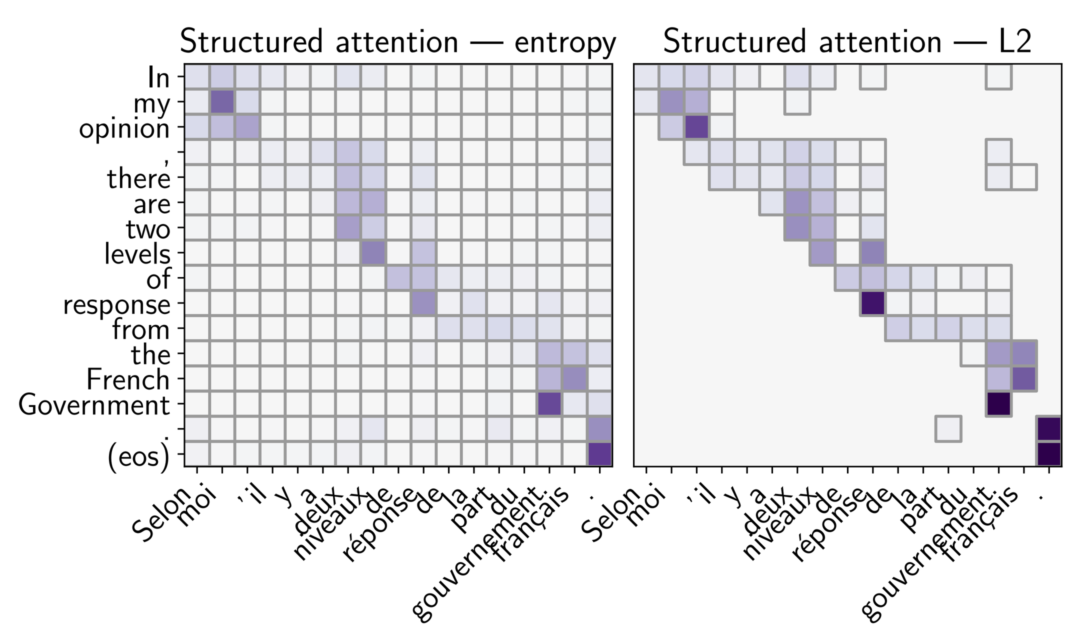


Exp – Structured and sparse attention (WMT14 fr-en)

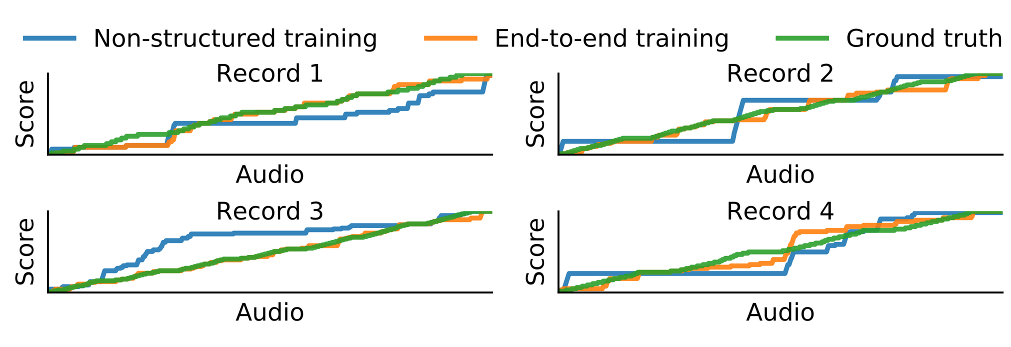
- Compute attention c by marginalizing a graphical model (Vit_Ω), with sparse marginal computation $\Omega = \ell_2^2$. vs simple softmax in original version

- Regularization is needed to backprop through argmax
- Vit_Ω double-backward using Pearlmutter (1994) AD

	Attention	WMT14 1M BLEU
	Softmax	**27.96**
fr→en	Entropy reg.	**27.96**
	ℓ_2^2 reg.	27.21
	Softmax	**28.08**
en→fr	Entropy reg.	27.98
	ℓ_2^2 reg.	27.28



Exp – Audio to score alignment



7 – Code + further

- PyTorch + AllenNLP implem:

github.com/arthurmensch/didyprog

- More on algo + theory in paper
- See also SparseMAP (Niculae et al.) for a different approach at ICML'18