Scaling Matters in Deep Structured Prediction Models

Alexander Shevchenko, Anton Osokin

Structured Prediction

Dataset: $\mathcal{D} = \{(X^j, Y^j)\}_{j=1}^N$

Structure: $Y = \{y_1, \ldots, y_L\}$

y_i highly correlated!

Tasks: sequence tagging, semantic segmentation, human pose estimation

Structured Prediction

Dataset: $\mathcal{D} = \{(X^j, Y^j)\}_{j=1}^N$

Structure: $Y = \{y_1, \ldots, y_L\}$

y_i highly correlated!

Tasks: sequence tagging, semantic segmentation, human pose estimation

How to encode structure?

Energy-based Models

Score (negative energy): $F(Y \mid \Psi(X))$

Energy-based Models

Score (negative energy): $F(Y \mid \Psi(X))$

Associated likelihood: $p(Y \mid X) \propto \exp\{F(Y \mid \Psi(X))\}$

$$Z(X) = \sum_{Y'} \exp\{F(Y' \mid \Psi(X))\}$$

Energy-based Models

Score (negative energy): $F(Y \mid \Psi(X))$

Associated likelihood: $p(Y \mid X) \propto \exp\{F(Y \mid \Psi(X))\}$

$$Z(X) = \sum_{Y'} \exp\{F(Y' \mid \Psi(X))\}$$

Potentials $\Psi(X)$

- $U(X)$ unary potentials
- $W(X)$ pairwise potentials

Energy-based Models Training

Consider the parameterization of potentials Ψ_θ

Training objective $$\frac{1}{N}\sum_{j=1}^{N}\mathcal{L}(X^j, Y^j, F_\theta(\cdot)) \to \min_\theta$$

Energy-based Models Training

Consider the parameterization of potentials Ψ_θ

Training objective $$\frac{1}{N}\sum_{j=1}^{N}\mathcal{L}(X^j, Y^j, F_\theta(\cdot)) \to \min_\theta$$

Stage training

+ with proper tuning works almost always
- time-consuming
- not stable due to phase switching

Energy-based Models Training

Consider the parameterization of potentials Ψ_θ

Training objective $$\frac{1}{N}\sum_{j=1}^{N}\mathcal{L}(X^j, Y^j, F_\theta(\cdot)) \to \min_\theta$$

Stage training

+ with proper tuning works almost always
- time-consuming
- not stable due to phase switching

Joint training

- increased optimization complexity
+ significantly faster
+ often results in better model

Energy-based Models Training

Consider the parameterization of potentials Ψ_θ

Training objective $$\frac{1}{N}\sum_{j=1}^{N}\mathcal{L}(X^j, Y^j, F_\theta(\cdot)) \to \min_\theta$$

Stage training

+ with proper tuning works almost always
- time-consuming
- not stable due to phase switching

Joint training

- increased optimization complexity
+ significantly faster
+ often results in better model

Reason to fail: improper relative scaling

Motivation: OCR case study

OBJECTIVE	STAGE	JOINT
Cross-entropy	**97.18** $\pm$ 0.12	96.48 $\pm$ 0.23
Structured SVM	**96.97** $\pm$ 0.27	96.43 $\pm$ 0.39
Log-likelihood	**97.15** $\pm$ 0.13	96.46 $\pm$ 0.25

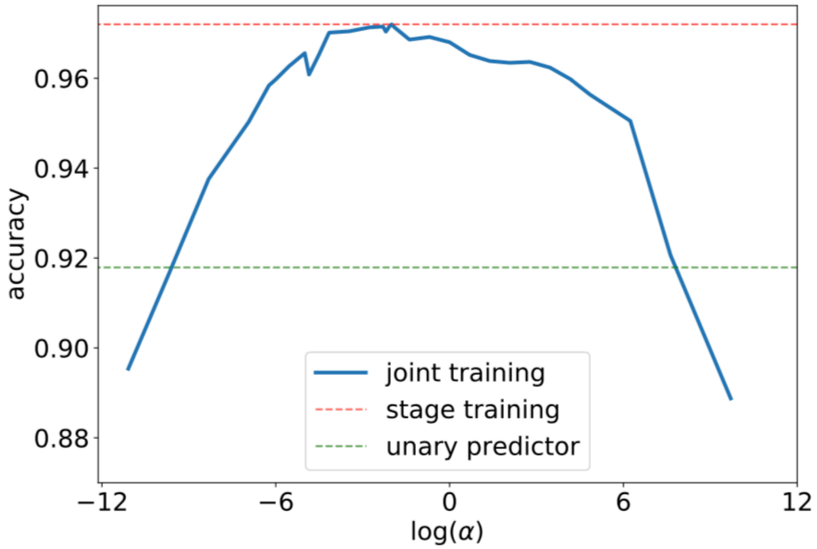
Motivation: OCR case study

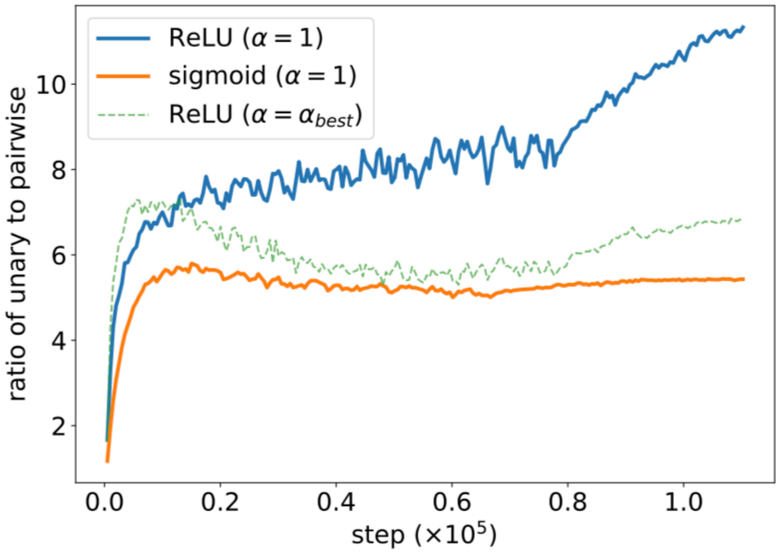
Rescale unary

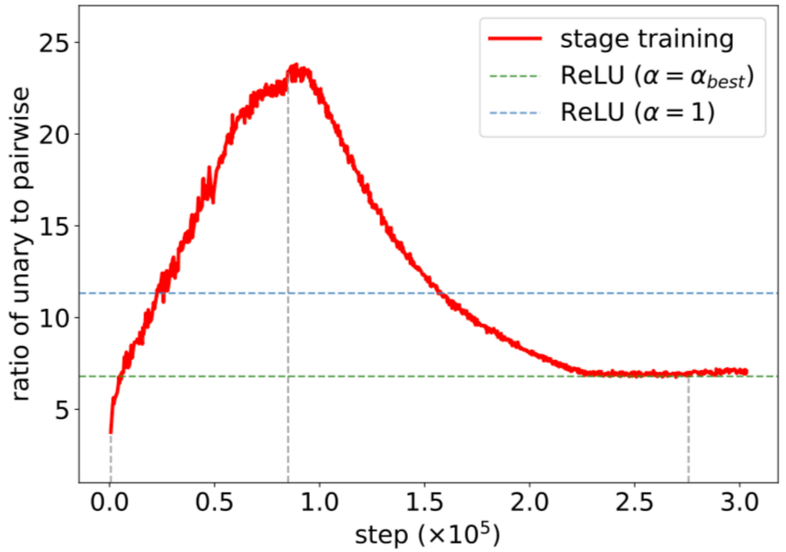
$$\hat{U}_\theta(X) = \alpha \cdot U(X)$$

OBJECTIVE	STAGE	JOINT
Cross-entropy	**97.18** $\pm$ 0.12	96.48 $\pm$ 0.23
Structured SVM	**96.97** $\pm$ 0.27	96.43 $\pm$ 0.39
Log-likelihood	**97.15** $\pm$ 0.13	96.46 $\pm$ 0.25

Motivation: OCR case study







Rescale unary

$$\hat{U}_\theta(X) = \alpha \cdot U(X)$$

OBJECTIVE	STAGE	JOINT
Cross-entropy	**97.18** ± 0.12	96.48 ± 0.23
Structured SVM	**96.97** ± 0.27	96.43 ± 0.39
Log-likelihood	**97.15** ± 0.13	96.46 ± 0.25

Linear-Chain CRF (LCCRF)

LCCRF score function

$$F(Y \mid X, \theta) = \sum_{i=1}^{L} U_\theta(x_i, y_i) + \sum_{i=1}^{L-1} W_\theta(y_i, y_{i+1})$$

Linear-Chain CRF (LCCRF)

LCCRF score function

$$F(Y \mid X, \theta) = \sum_{i=1}^{L} U_\theta(x_i, y_i) + \sum_{i=1}^{L-1} W_\theta(y_i, y_{i+1})$$

Likelihood and marginals

$$P(Y \mid X, \theta) = \frac{\exp\{F(Y \mid X, \theta)\}}{Z(\theta)}$$

$$p_i(y_i \mid X, \theta) = \sum_{Y \setminus y_i} P(Y \mid X, \theta)$$

Linear-Chain CRF (LCCRF)

LCCRF score function

$$F(Y \mid X, \theta) = \sum_{i=1}^{L} U_\theta(x_i, y_i) + \sum_{i=1}^{L-1} W_\theta(y_i, y_{i+1})$$

Likelihood and marginals

$$P(Y \mid X, \theta) = \frac{\exp\{F(Y \mid X, \theta)\}}{Z(\theta)}$$

$$p_i(y_i \mid X, \theta) = \sum_{Y \setminus y_i} P(Y \mid X, \theta)$$

Inference

- ✦ Partition function, marginals (sum-product BP)
- ✦ MAP estimate (max-sum BP)

Linear-Chain CRF (LCCRF)

LCCRF score function

$$F(Y \mid X, \theta) = \sum_{i=1}^{L} U_\theta(x_i, y_i) + \sum_{i=1}^{L-1} W_\theta(y_i, y_{i+1})$$

Likelihood and marginals

$$P(Y \mid X, \theta) = \frac{\exp\{F(Y \mid X, \theta)\}}{Z(\theta)}$$

$$p_i(y_i \mid X, \theta) = \sum_{Y \setminus y_i} P(Y \mid X, \theta)$$

Inference

- ✦ Partition function, marginals (sum-product BP)
- ✦ MAP estimate (max-sum BP)

Training objectives

$$\mathcal{L}_{\text{MLE}}(X, Y \mid \theta) = -\log P(Y \mid X, \theta)$$

Linear-Chain CRF (LCCRF)

LCCRF score function

$$F(Y \mid X, \theta) = \sum_{i=1}^{L} U_\theta(x_i, y_i) + \sum_{i=1}^{L-1} W_\theta(y_i, y_{i+1})$$

Likelihood and marginals

$$P(Y \mid X, \theta) = \frac{\exp\{F(Y \mid X, \theta)\}}{Z(\theta)}$$

$$p_i(y_i \mid X, \theta) = \sum_{Y \setminus y_i} P(Y \mid X, \theta)$$

Inference

- ✦ Partition function, marginals (sum-product BP)
- ✦ MAP estimate (max-sum BP)

Training objectives

$$\mathcal{L}_{\text{MLE}}(X, Y \mid \theta) = -\log P(Y \mid X, \theta)$$

$$\mathcal{L}_{\text{cross}}(X, Y \mid \theta) = -\frac{1}{L} \sum_{i=1}^{L} \log p_i(y_i \mid X, \theta)$$

Linear-Chain CRF (LCCRF)

LCCRF score function

$$F(Y \mid X, \theta) = \sum_{i=1}^{L} U_\theta(x_i, y_i) + \sum_{i=1}^{L-1} W_\theta(y_i, y_{i+1})$$

Likelihood and marginals

$$P(Y \mid X, \theta) = \frac{\exp\{F(Y \mid X, \theta)\}}{Z(\theta)}$$

$$p_i(y_i \mid X, \theta) = \sum_{Y \setminus y_i} P(Y \mid X, \theta)$$

Inference

- ✦ Partition function, marginals (sum-product BP)
- ✦ MAP estimate (max-sum BP)

Training objectives

$$\mathcal{L}_{\text{MLE}}(X, Y \mid \theta) = -\log P(Y \mid X, \theta)$$

$$\mathcal{L}_{\text{cross}}(X, Y \mid \theta) = -\frac{1}{L} \sum_{i=1}^{L} \log p_i(y_i \mid X, \theta)$$

$$\mathcal{L}_{\text{SSVM}}(X, Y \mid \theta) = \max_{Y'} \{F(Y' \mid X, \theta) + \Delta(Y', Y)\} - F(Y \mid X, \theta)$$

Hamming distance

$$\Delta(Y', Y) = \frac{1}{L} \sum_{i=1}^{L} \mathbb{I}[y'_i \neq y_i]$$

Gaussian CRF (GCRF)

The energy function has different nature

GCRF score function

$$F(s \mid X, \theta) = -\frac{1}{2} s^T (W_\theta(X) + \lambda I) s + U_\theta(x) s$$

$$s = [s_{1,1}, \ldots, s_{1,L} | \ldots | s_{M,1}, \ldots, s_{M,L}]$$

$$U_\theta(X) = [u_\theta^1(X) | \ldots | u_\theta^M(X)] \qquad W_\theta(X) \in \mathbb{R}^{LM \times LM}$$

Gaussian CRF (GCRF)

The energy function has different nature

GCRF score function

$$F(s \mid X, \theta) = -\frac{1}{2} s^T (W_\theta(X) + \lambda I) s + U_\theta(x) s$$

$$s = [s_{1,1}, \ldots, s_{1,L} | \ldots | s_{M,1}, \ldots, s_{M,L}]$$

$$U_\theta(X) = [u_\theta^1(X) | \ldots | u_\theta^M(X)] \qquad W_\theta(X) \in \mathbb{R}^{LM \times LM}$$

Optimal scores and marginals

$$s^* = (W_\theta(x) + \lambda I)^{-1} U_\theta(X)$$ done efficiently via CG

$$p_j(y_j \mid X, \theta) = \mathrm{softmax}_{y_j}(s^*_{:,j})$$

Gaussian CRF (GCRF)

The energy function has different nature

GCRF score function

$$F(s \mid X, \theta) = -\frac{1}{2} s^T (W_\theta(X) + \lambda I) s + U_\theta(x) s$$

$$s = [s_{1,1}, \ldots, s_{1,L} | \ldots | s_{M,1}, \ldots, s_{M,L}]$$

$$U_\theta(X) = [u_\theta^1(X) | \ldots | u_\theta^M(X)] \qquad W_\theta(X) \in \mathbb{R}^{LM \times LM}$$

Optimal scores and marginals

$$s^* = (W_\theta(x) + \lambda I)^{-1} U_\theta(X)$$ done efficiently via CG

$$p_j(y_j \mid X, \theta) = \mathrm{softmax}_{y_j}(s^*_{:,j})$$

Prediction

$$\hat{Y} := [\arg\max_k s^*_{k,1}, \ldots, \arg\max_k s^*_{k,L}]$$

Gaussian CRF (GCRF)

The energy function has different nature

GCRF score function

$$F(s \mid X, \theta) = -\frac{1}{2} s^T (W_\theta(X) + \lambda I) s + U_\theta(x) s$$

$$s = [s_{1,1}, \ldots, s_{1,L} | \ldots | s_{M,1}, \ldots, s_{M,L}]$$

$$U_\theta(X) = [u_\theta^1(X) | \ldots | u_\theta^M(X)] \qquad W_\theta(X) \in \mathbb{R}^{LM \times LM}$$

Optimal scores and marginals

$$s^* = (W_\theta(x) + \lambda I)^{-1} U_\theta(X)$$ done efficiently via CG

$$p_j(y_j \mid X, \theta) = \mathrm{softmax}_{y_j}(s^*_{:,j})$$

Prediction

$$\hat{Y} := [\arg\max_k s^*_{k,1}, \ldots, \arg\max_k s^*_{k,L}]$$

Backward pass

$$\frac{\partial \mathcal{L}}{\partial U} = (W + \lambda I)^{-1} \frac{\partial \mathcal{L}}{\partial s^*}, \quad \frac{\partial \mathcal{L}}{\partial \mathrm{vec}(W)} = -\frac{\partial \mathcal{L}}{\partial U} \otimes s^*$$

Gaussian CRF (GCRF)

The energy function has different nature

GCRF score function

$$F(s \mid X, \theta) = -\frac{1}{2} s^T (W_\theta(X) + \lambda I) s + U_\theta(x) s$$

$$s = [s_{1,1}, \ldots, s_{1,L} | \ldots | s_{M,1}, \ldots, s_{M,L}]$$

$$U_\theta(X) = [u_\theta^1(X) | \ldots | u_\theta^M(X)] \qquad W_\theta(X) \in \mathbb{R}^{LM \times LM}$$

Optimal scores and marginals

$$s^* = (W_\theta(x) + \lambda I)^{-1} U_\theta(X)$$ done efficiently via CG

$$p_j(y_j \mid X, \theta) = \mathrm{softmax}_{y_j}(s^*_{:,j})$$

Prediction

$$\hat{Y} := [\arg\max_k s^*_{k,1}, \ldots, \arg\max_k s^*_{k,L}]$$

Backward pass

$$\frac{\partial \mathcal{L}}{\partial U} = (W + \lambda I)^{-1} \frac{\partial \mathcal{L}}{\partial s^*}, \quad \frac{\partial \mathcal{L}}{\partial \mathrm{vec}(W)} = -\frac{\partial \mathcal{L}}{\partial U} \otimes s^*$$

Pairwise potentials

Class-agnostic Potts model

$\mathcal{A}_i$ - pixel embedding

$$W = \begin{bmatrix} 0 & \mathcal{A}\mathcal{A}^T \\ \mathcal{A}\mathcal{A}^T & 0 \end{bmatrix}$$

Online Scaling

Idea

Modify the score function:

$$F_\alpha(Y \mid X, \theta) = F(Y \mid \alpha U_\theta(x), W_\theta(x))$$

Online Scaling

Idea

Modify the score function:

$$F_\alpha(Y \mid X, \theta) = F(Y \mid \alpha U_\theta(x), W_\theta(x))$$

Update scaling factor via grid-search on:

$$\mathcal{L}_\alpha(\mathcal{D}) = \frac{1}{|\mathcal{D}|} \sum_{(X,Y)\in\mathcal{D}} \mathcal{L}(Y, \alpha U_\theta(X), W_\theta(X))$$

Online Scaling

Idea

Modify the score function:

$$F_\alpha(Y \mid X, \theta) = F(Y \mid \alpha U_\theta(x), W_\theta(x))$$

Update scaling factor via grid-search on:

$$\mathcal{L}_\alpha(\mathcal{D}) = \frac{1}{|\mathcal{D}|} \sum_{(X,Y)\in\mathcal{D}} \mathcal{L}(Y, \alpha U_\theta(X), W_\theta(X))$$

Algorithm 1 Online Scaling

initialize θ
epoch = 0, $\alpha = 1$
for epoch $<$ n_epoch **do**
  for (X, Y) in training set **do**
    compute $(U_\theta(X), W_\theta(X))$
    compute $\mathcal{L}(Y, \alpha U_\theta(X), W_\theta(X))$
    update θ via an SGD step
  end for
  epoch += 1
  $\mathcal{D}$:= training set or its subset
  pick the best α_{best} minimizing $\mathcal{L}_\alpha(\mathcal{D})$ on an α-grid
  $\alpha = \alpha_{\text{best}}$
end for

Online Scaling

Idea

Modify the score function:

$$F_\alpha(Y \mid X, \theta) = F(Y \mid \alpha U_\theta(x), W_\theta(x))$$

Update scaling factor via grid-search on:

$$\mathcal{L}_\alpha(\mathcal{D}) = \frac{1}{|\mathcal{D}|} \sum_{(X,Y)\in\mathcal{D}} \mathcal{L}(Y, \alpha U_\theta(X), W_\theta(X))$$

Remark:

$$\text{constant scaling} \Leftrightarrow \text{init} + \text{lr for last layer}$$

Algorithm 1 Online Scaling

initialize θ
epoch = 0, $\alpha = 1$
for epoch $<$ n_epoch **do**
  for (X, Y) in training set **do**
    compute $(U_\theta(X), W_\theta(X))$
    compute $\mathcal{L}(Y, \alpha U_\theta(X), W_\theta(X))$
    update θ via an SGD step
  end for
  epoch += 1
  $\mathcal{D}$:= training set or its subset
  pick the best α_{best} minimizing $\mathcal{L}_\alpha(\mathcal{D})$ on an α-grid
  $\alpha = \alpha_{\text{best}}$
end for

Offline Scaling

Explicit relative scaling

$$\hat{U}_\theta(X) = \frac{U_\theta(X)}{\|U_\theta(X)\|} \cdot \alpha, \quad \hat{W}_\theta(X) = \frac{W_\theta(X)}{\|W_\theta(X)\|}$$

Offline Scaling

Explicit relative scaling

$$\hat{U}_\theta(X) = \frac{U_\theta(X)}{\|U_\theta(X)\|} \cdot \alpha, \quad \hat{W}_\theta(X) = \frac{W_\theta(X)}{\|W_\theta(X)\|}$$

or implicit regularizer version

$$\mathcal{R} = \lambda \left(\frac{\|U_\theta(X)\|}{\|W_\theta(X)\|} - \alpha \right)^2$$

Tasks and architectures

Optical Character Recognition

Unary: LeNet-5 like NN

Pairwise: real-valued matrix of size 26

Tasks and architectures

Optical Character Recognition

Unary: LeNet-5 like NN

Pairwise: real-valued matrix of size 26

Text Chunking

[NP He] [VP reckons] [NP the current account deficit] [VP will narrow] [PP to] [NP only # 1.8 billion] [PP in] [NP September] .

Figure 6. Chunking example for: He reckons the current account deficit will narrow to only #1.8 billion in September.

Unary: char-level BiLSTM + SENNA + POS + FC

Pairwise: real-value matrix of size N_{tags}

Tasks and architectures

Optical Character Recognition

Unary: LeNet-5 like NN

Pairwise: real-valued matrix of size 26

Binary Segmentation

Unary and Pairwise: Unet-like FCNNs

Text Chunking

[NP He] [VP reckons] [NP the current account deficit] [VP will narrow] [PP to] [NP only # 1.8 billion] [PP in] [NP September] .

Figure 6. Chunking example for: He reckons the current account deficit will narrow to only #1.8 billion in September.

Unary: char-level BiLSTM + SENNA + POS + FC

Pairwise: real-value matrix of size N_{tags}

Results

TASK	OBJECTIVE	STAGE	JOINT	ONLINE	OFFLINE	OFFLINE (Reg)
OCR	Cross-entropy	97.18 ± 0.12	96.48 ± 0.23	97.25 ± 0.11	97.20 ± 0.12	97.14 ± 0.18
	Structured SVM	96.97 ± 0.27	96.43 ± 0.39	97.01 ± 0.24	96.99 ± 0.29	96.91 ± 0.35
	Log-likelihood	97.15 ± 0.13	96.46 ± 0.25	97.17 ± 0.11	97.14 ± 0.13	97.08 ± 0.17
	Cross-entropy (MF)	97.06 ± 0.09	96.62 ± 0.18	97.11 ± 0.08	97.10 ± 0.11	97.01 ± 0.14
Chunking	Cross-entropy	89.52 ± 0.29	87.92 ± 0.38	89.56 ± 0.26	89.53 ± 0.28	89.45 ± 0.34
	Structured SVM	89.21 ± 0.49	87.64 ± 0.57	89.25 ± 0.44	89.19 ± 0.51	89.13 ± 0.55
	Log-likelihood	89.43 ± 0.31	87.85 ± 0.37	89.48 ± 0.27	89.45 ± 0.29	89.35 ± 0.36
	Cross-entropy (MF)	89.25 ± 0.26	87.68 ± 0.29	89.34 ± 0.22	89.27 ± 0.30	89.15 ± 0.28
Bin. segm.	Cross-entropy	86.51 ± 0.19	85.62 ± 0.31	86.56 ± 0.17	86.60 ± 0.21	86.48 ± 0.25

Conclusion

- Conjecture: scaling matters
- We propose online and offline scaling algorithms
- And demonstrate their efficiency on three different tasks