Primal-Dual Formulation for Deep Learning with Constraints

Yatin Nandwani, Abhishek Pathak, Mausam and Parag Singla
Department of Computer Science and Engineering
Indian Institute of Technology Delhi

Deep Learning with Constraints

- Augmenting deep neural models (**DNN**) with Domain Knowledge (**DK**)

- **Domain Knowledge** expressed in the form of ***Constraints*** **(C)**

  - **Learning with constraints:** Learn **DNN** weights s.t. output satisfies **constraints C**

Learning with Constraints: *Motivation*

- Why bother about learning with constraints over constrained inference?
    - Inference time speed up
    - Better performance, *verified experimentally*
    - Tool to exploit unlabeled data

Learning with Constraints: *Motivation*

- Why bother about learning with constraints over constrained inference?
  - Inference time speed up
  - Better performance, *verified experimentally*
  - Tool to exploit unlabeled data

- **Learning with constraints:** A framework for directly solving constrained optimization problem, instead of an approximation: Based on Primal Dual Formulation.

Learning with Constraints: *Formulation*

Unconstrained Problem

$$w^* = \arg\min_w L(w);$$

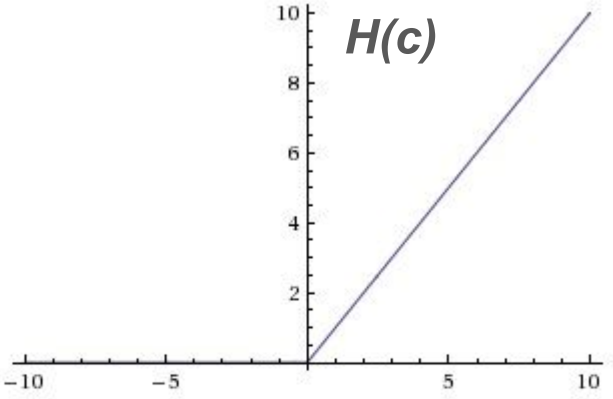
$$L(w) = \frac{1}{m} \sum_{i=1}^{m} \sum_{j=1}^{r} \mathbb{1}\{y_j^{(i)} = v\} \log(P_w(y_j^{(i)} = v)|x)$$

Constrained Problem

$$\arg\min_w L(w) \quad \text{subject to} \quad f_k^i(w) \leq 0; \;\; \forall 1 \leq i \leq m; \;\; \forall 1 \leq k \leq K$$

$O(mK)$ number of constraints

Learning with Constraints: *Reduce # Constraints*



$$H(c) = c \;\; for \;\; c \geq 0, \;\; \text{and} \;\; 0 \;\; for \;\; c < 0$$

$f_k^i(w) \leq 0$ **Equivalent To:** $H(f_k^i(w)) = 0$

$$\arg\min_w L(w) \;\; \text{subject to} \quad H(f_k^i(w)) = 0; \;\; \forall 1 \leq i \leq m; \;\; \forall 1 \leq k \leq K$$

$\forall i : H(f_k^i(w)) = 0$ **Equivalent To:** $\sum_i H(f_k^i(w)) = 0$ Defining $h_k(w) = \sum_i H(f_k^i(w))$

$$\arg\min_w L(w) \;\; \text{subject to} \quad h_k(w) = 0; \;\; \forall 1 \leq k \leq K$$

Learning with Constraints: *Primal-Dual Formulation*

$$\arg\min_w L(w) \text{ subject to } \quad h_k(w) = 0; \ \forall 1 \leq k \leq K$$

Lagrangian

$$\mathcal{L}(w; \Lambda) = L(w) + \sum_{k=1}^{K} \lambda_k h_k(w)$$

Primal

$$\min_w \max_\Lambda \mathcal{L}(w, \Lambda)$$

Dual

$$\max_\Lambda \min_w \mathcal{L}(w, \Lambda)$$

Learning with Constraints: *Update Equations*

$$\mathcal{L}(w;\Lambda) = L(w) + \sum_{k=1}^{K} \lambda_k h_k(w)$$

Derivative w.r.t. ***weights***

$$\nabla_w \mathcal{L}(w;\Lambda) = \nabla_w L(w) + \sum_{k=1}^{K} \lambda_k \nabla_w h_k(w)$$

Derivative w.r.t. ***Lambda***

$$\frac{\partial \mathcal{L}(w;\Lambda)}{\partial \lambda_k} = h_k(w), \forall k$$

Update eq. for ***weights***

$$w^{(t_1+1)} \leftarrow w^{(t_1)} - \alpha_w \nabla_w \mathcal{L}(w;\Lambda)$$

Update eq. for ***Lambda***

$$\Lambda^{(t_2+1)} \leftarrow \Lambda^{(t_2)} + \alpha_\Lambda \nabla_\Lambda \mathcal{L}(w;\Lambda)$$

Learning with Constraints: *Training Algorithm*

Algorithm 1 Training of a Deep Net with Constraints. Hyperparameters: $warmup$, d, β, α_Λ^0, α_w

```
Initialize: w randomly; λ_k = 0, ∀k = 1 ... K
for warmup iterations do
    Update w: Take an SGD step wrt w on L(w; Λ) on a mini-batch
end
Initialize: l = 1; t = 1; t_1 = 1; α_Λ = α_Λ^0
while not converged do
    Update Λ: Take an SGA step wrt Λ on L(w; Λ) on a mini-batch
    Increment t = t + 1
    for l steps do
        Update w: Take an SGD step wrt w on L(w; Λ) on a mini-batch
        Increment t_1 = t_1 + 1
    end
    Update l = l + d
    Set learning rates: α_Λ = α_Λ^0 · 1/(1+βt)
end
```

Learning with Constraints: *Experiments*

SRL

	F1 Score			Total Constraint Violations		
Scenario	**1% Data**	**5% Data**	**10% Data**	**1% Data**	**5% Data**	**10% Data**
B	62.99	72.64	76.04	14,857	9,708	7,704
CL	66.21	74.27	77.19	9,406	7,461	5,836
B+CI*	67.90	75.96	78.63	5,737	4,247	3,654
CL + CI*	68.71	76.51	78.72	5,039	3,963	3,476

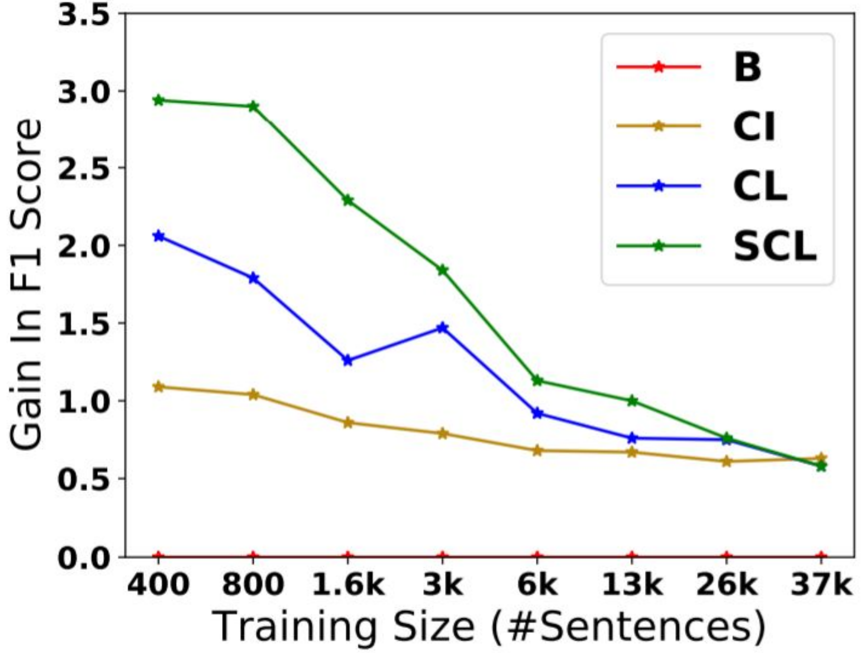
Learning with Constraints: *Experiments*

Typenet

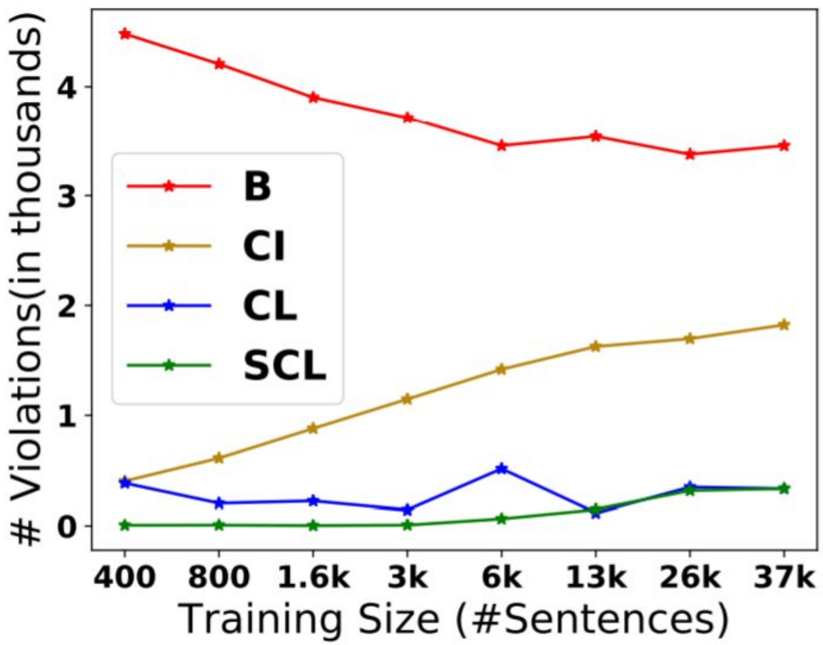
	MAP Scores			Constraint Violations		
Scenario	**5% Data**	**10% Data**	**100% Data**	**5% Data**	**10% Data**	**100% Data**
B	68.62	69.21	70.47	22,715	21,451	22,359
B+H	68.71	69.31	71.77	22,928	21,157	24,650
CL	80.13	81.36	82.80	25	45	12
SCL	82.22	83.81		41	26	

Learning with Constraints: *Experiments*

NER



(a) Avg. Gain in F1 Score Over Baseline.



(b) Avg. number of Constrained Violations

Thank You!