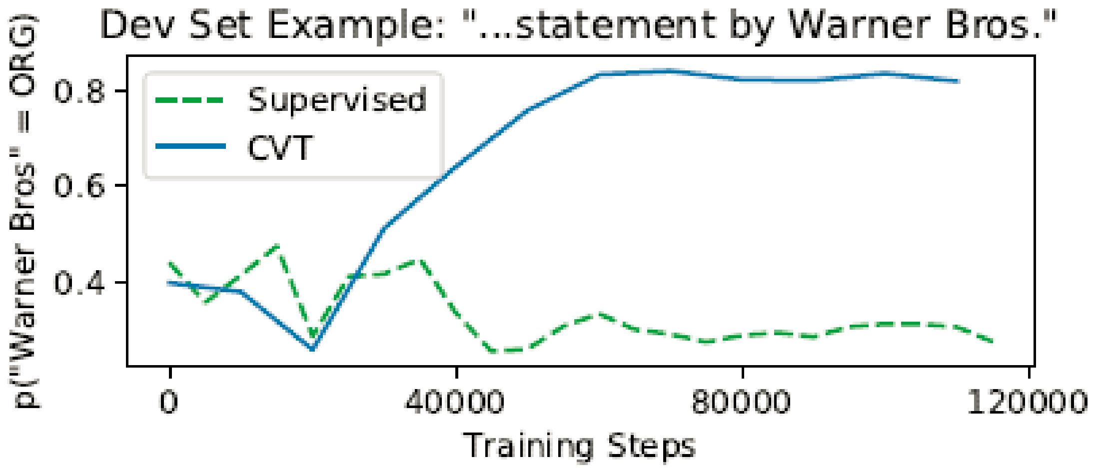
Semi-Supervised Sequence Modeling with Cross-View Training

Kevin Clark, Minh-Thang Luong, Christopher D. Manning, Quoc V. Le
Computer Science Department, Stanford University
Google Brain

EMNLP 2018

Motivation

- Lack of labeled data



- Information from large amounts of unlabeled data can improve the performance of supervised NLP models

- Pre-training does not learn labeled data (generally representations)

Self-Training

Algorithm 1 Self-training.

Initialize:
Given $(X_{train}, y_{train}) = (X_l, y_l)$
while stopping criteria not met **do**
  Train classifier C_{int} from (X_{train}, y_{train}) **Student**
  Use C_{int} to predict class label y_u of X_u **Teacher**
  Select confidence sample (X_{conf}, y_{conf}); $(X_{conf}, y_{conf}) \in (X_u, y_u)$
  Remove selected unlabeled data $X_u \leftarrow X_u - X_{conf}$
  Combine newly labeled data $(X_{train}, y_{train}) \leftarrow (X_l, y_l) \cup (X_{conf}, y_{conf})$
end while

Drawbacks of Self-Training

- The model acts as both a teacher and a student
- The model already produces the predictions it is being trained on
- Adding noise to the student's input, training the model so it is robust to input perturbations on CV
- Applying noise is difficult for discrete inputs like text

Cross-View Training (CVT)

- Multi-view learning: train the model to produce consistent predictions across different views of the input

- CVT adds auxiliary prediction modules (student)

- a restricted view of the input example

- representation learning

Inputs Seen by Auxiliary Prediction Modules

Auxiliary 1: *They traveled to* ____________________

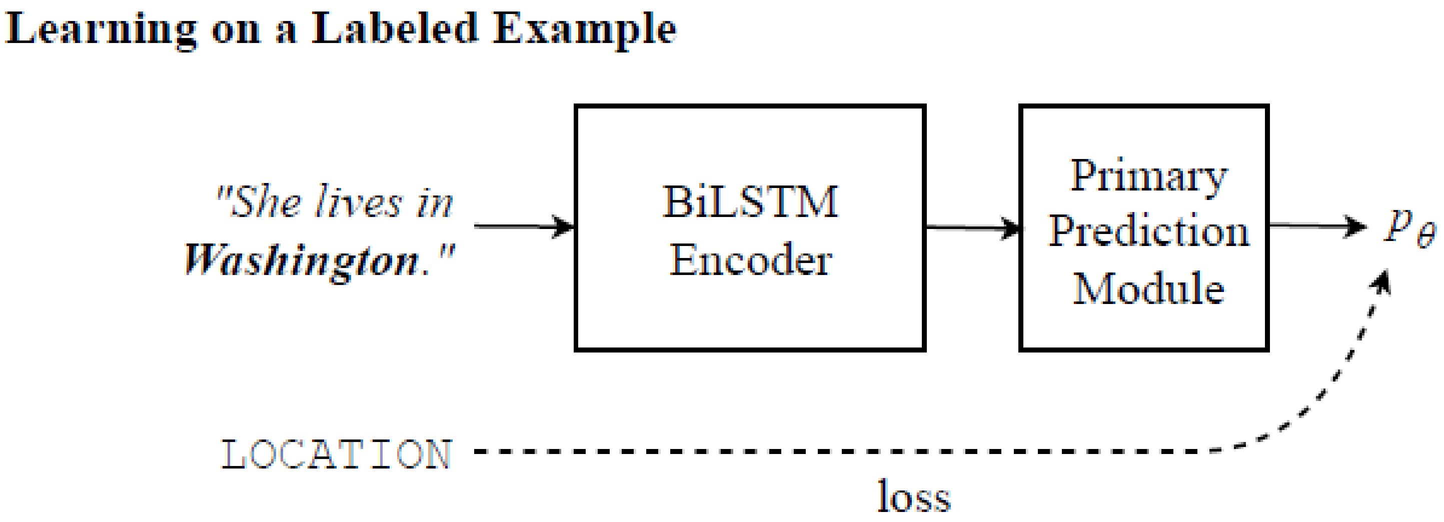
Auxiliary 2: *They traveled to* ***Washington*** ________

Auxiliary 3: ______________ ***Washington*** *by plane*

Auxiliary 4: __________________________ *by plane*

Cross-View Training

- Supervised Learning:

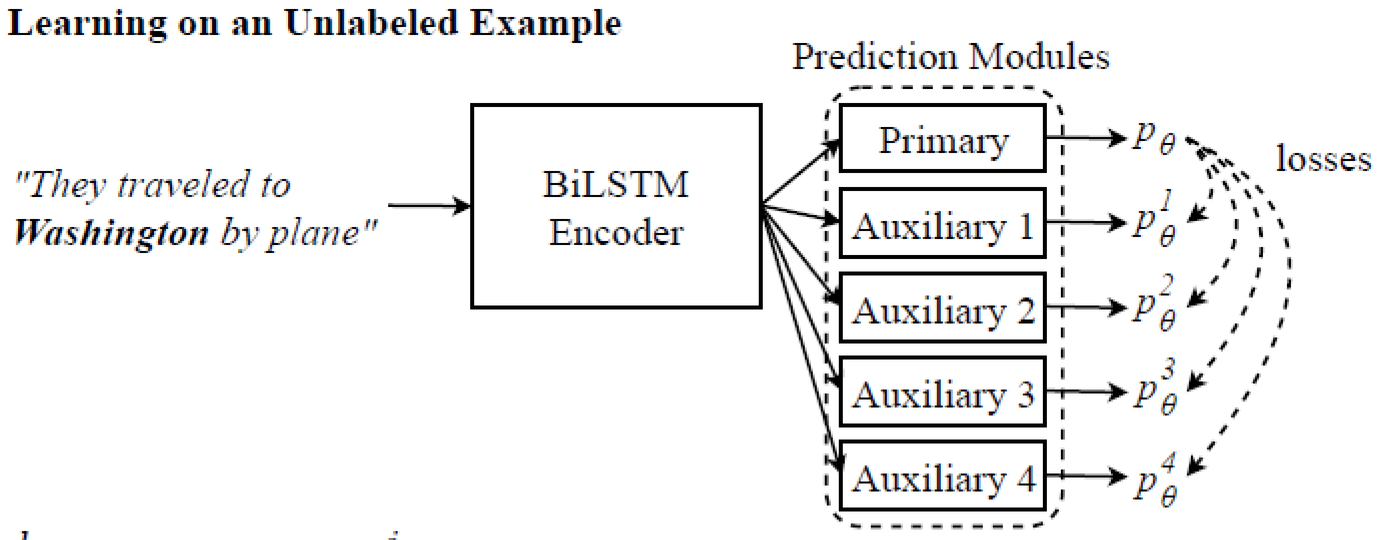


- Supervised Loss:

$$\mathcal{L}_{\text{sup}}(\theta) = \frac{1}{|\mathcal{D}_l|} \sum_{x_i, y_i \in \mathcal{D}_l} CE(y_i, p_\theta(y|x_i))$$

Cross-View Training

- Unsupervised Learning:



- Unsupervised Loss:

$$\mathcal{L}_{\text{CVT}}(\theta) = \frac{1}{|\mathcal{D}_{ul}|} \sum_{x_i \in \mathcal{D}_{ul}} \sum_{j=1}^{k} D(p_\theta(y|x_i), p_\theta^j(y|x_i))$$

- KL divergence

Cross-View Training

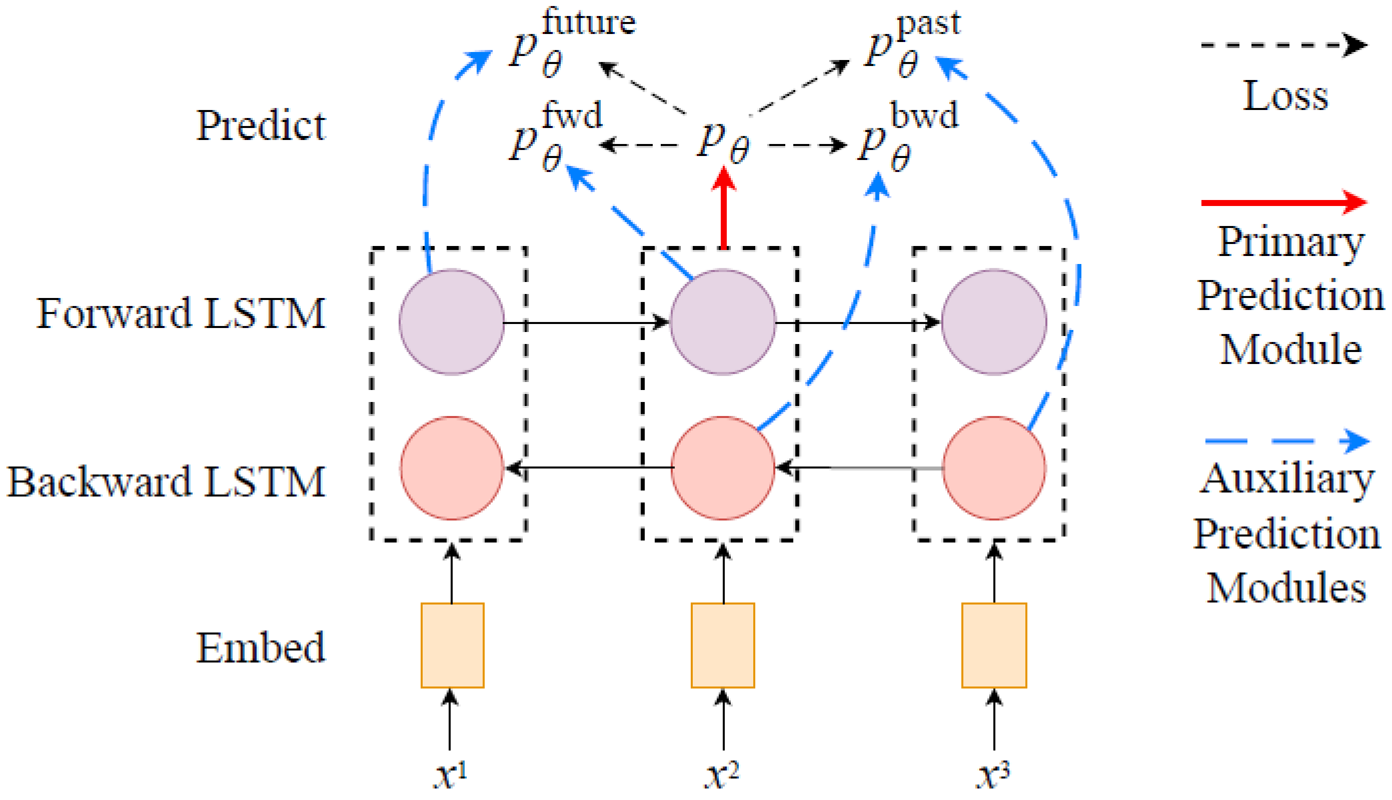
Notes:

- The model alternates learning on a mini-batch of labeled examples and learning on a mini-batch of unlabeled examples

- Hold the primary module’s parameters fixed during unsupervised training

- The auxiliary prediction modules are only used during training

CVT Model on Sequence Tagging



primary prediction module

$$p(y^t|x_i) = \text{NN}(h_1^t \oplus h_2^t)$$
$$= \text{softmax}(U \cdot \text{ReLU}(W(h_1^t \oplus h_2^t)) + b)$$

auxiliary prediction modules

$$p_\theta^{\text{fwd}}(y^t|x_i) = \text{NN}^{\text{fwd}}(\overrightarrow{h}_1^t(x_i))$$
$$p_\theta^{\text{bwd}}(y^t|x_i) = \text{NN}^{\text{bwd}}(\overleftarrow{h}_1^t(x_i))$$
$$p_\theta^{\text{future}}(y^t|x_i) = \text{NN}^{\text{future}}(\overrightarrow{h}_1^{t-1}(x_i))$$
$$p_\theta^{\text{past}}(y^t|x_i) = \text{NN}^{\text{past}}(\overleftarrow{h}_1^{t+1}(x_i))$$

CVT Model on Dependency Parsing

- Each word receives one in-going edge (u, t, r)
- going from word x_i^u (called the “head”) to it (the “dependent”) of type r (the “relation”).
- treats dependency parsing as a classification task

CVT Model on Dependency Parsing

primary prediction module

$$p_\theta((u,t,r)|x_i) \propto e^{s(h_1^u(x_i) \oplus h_2^u(x_i), h_1^t(x_i) \oplus h_2^t(x_i), r)}$$

$$s(z_1, z_2, r) = \text{ReLU}(W_{\text{head}} z_1 + b_{\text{head}})(W_r + W) \\ \text{ReLU}(W_{\text{dep}} z_2 + b_{\text{dep}})$$

auxiliary prediction modules

$$p_\theta^{\text{fwd-fwd}}((u,t,r)|x_i) \propto e^{s^{\text{fwd-fwd}}(\overrightarrow{h}_1^u(x_i), \overrightarrow{h}_1^t(x_i), r)}$$

$$p_\theta^{\text{fwd-bwd}}((u,t,r)|x_i) \propto e^{s^{\text{fwd-bwd}}(\overrightarrow{h}_1^u(x_i), \overleftarrow{h}_1^t(x_i), r)}$$

$$p_\theta^{\text{bwd-fwd}}((u,t,r)|x_i) \propto e^{s^{\text{bwd-fwd}}(\overleftarrow{h}_1^u(x_i), \overrightarrow{h}_1^t(x_i), r)}$$

$$p_\theta^{\text{bwd-bwd}}((u,t,r)|x_i) \propto e^{s^{\text{bwd-bwd}}(\overleftarrow{h}_1^u(x_i), \overleftarrow{h}_1^t(x_i), r)}$$

CVT Model on Sequence-to-Sequence Learning

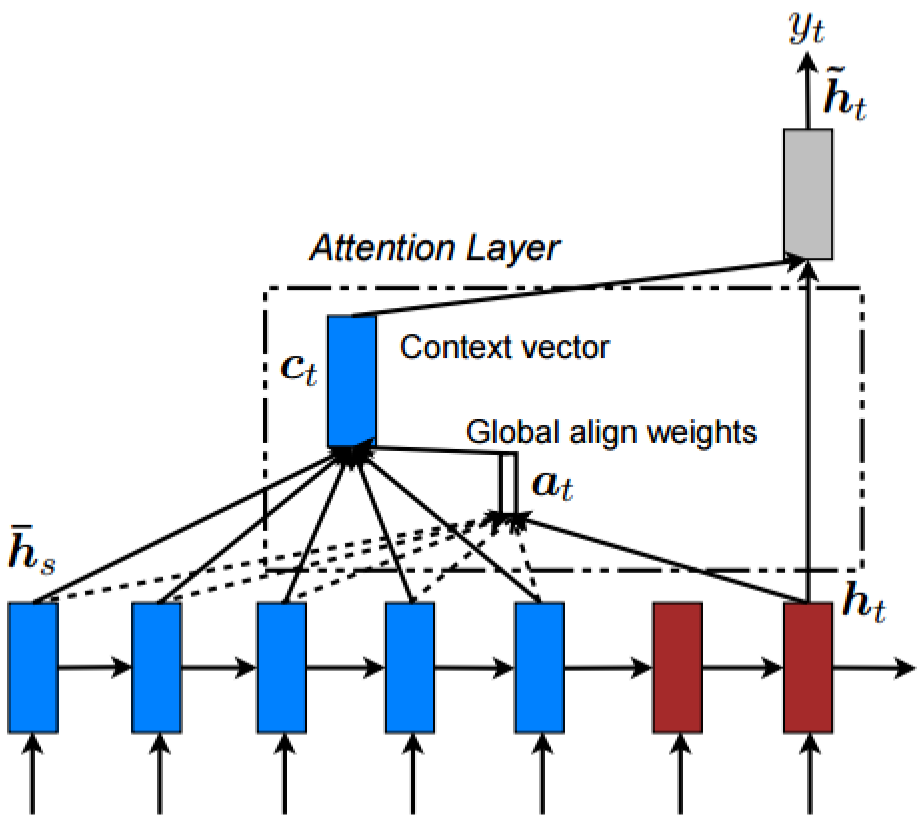
primary prediction decoder:

- LSTM decoder with attention mechanism

auxiliary prediction decoders:

- Apply attention dropout, randomly zeroing out attention weights
- predict the next word in the target sequence

$$p_\theta^{\text{future}}(y_i^t | y_i^{<t}, x_i) = \text{softmax}(W_s^{\text{future}} a_{t-1}^{\text{future}})$$



CVT Model on Sequence-to-Sequence Learning

Unsupervised Loss:

- cannot get an output distribution over the vocabulary from the primary decoder at each time step

- produce hard targets for the auxiliary modules by running the primary decoder with beam search on the input sequence

CVT & Multi-Task Learning

Supervised learning:

- randomly select a task
- update $\mathcal{L}_{sup}$ using a mini-batch of labeled data for that task

Unsupervised learning:

- jointly across all tasks
- update $\mathcal{L}_{CVT}$ using a mini-batch of unlabeled data
- all-tasks-labeled examples

Experiment

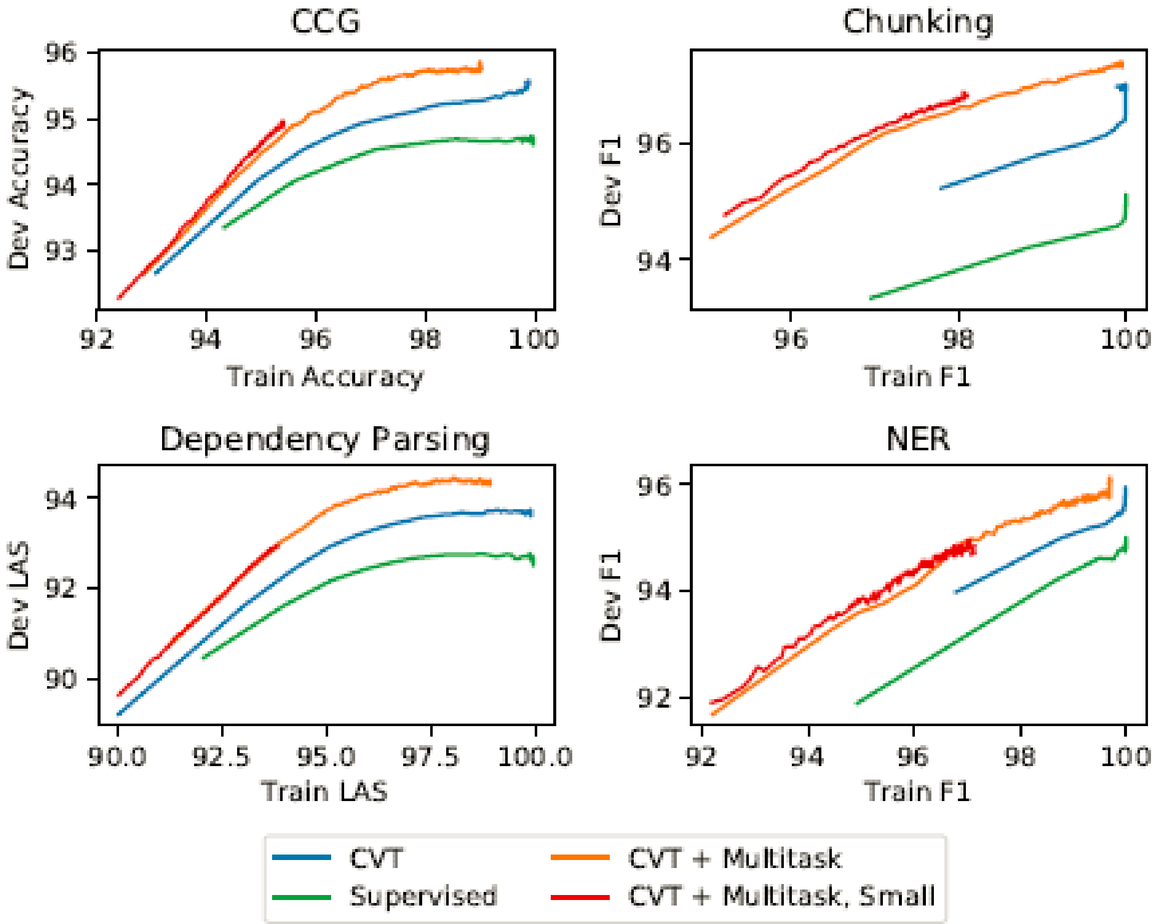
seven tasks:

- Combinatory Categorial Grammar (CCG) Supertagging: CCGBank
- Text Chunking: CoNLL-2000
- Named Entity Recognition (NER): CoNLL-2003
- Fine-Grained NER (FGN): OntoNotes
- Part-of-Speech (POS) Tagging: Wall Street Journal portion of the Penn Treebank
- Dependency Parsing: Penn Treebank converted to Stanford Dependencies version 3.3.0
- Machine Translation: English-Vietnamese translation dataset from IWSLT 2015

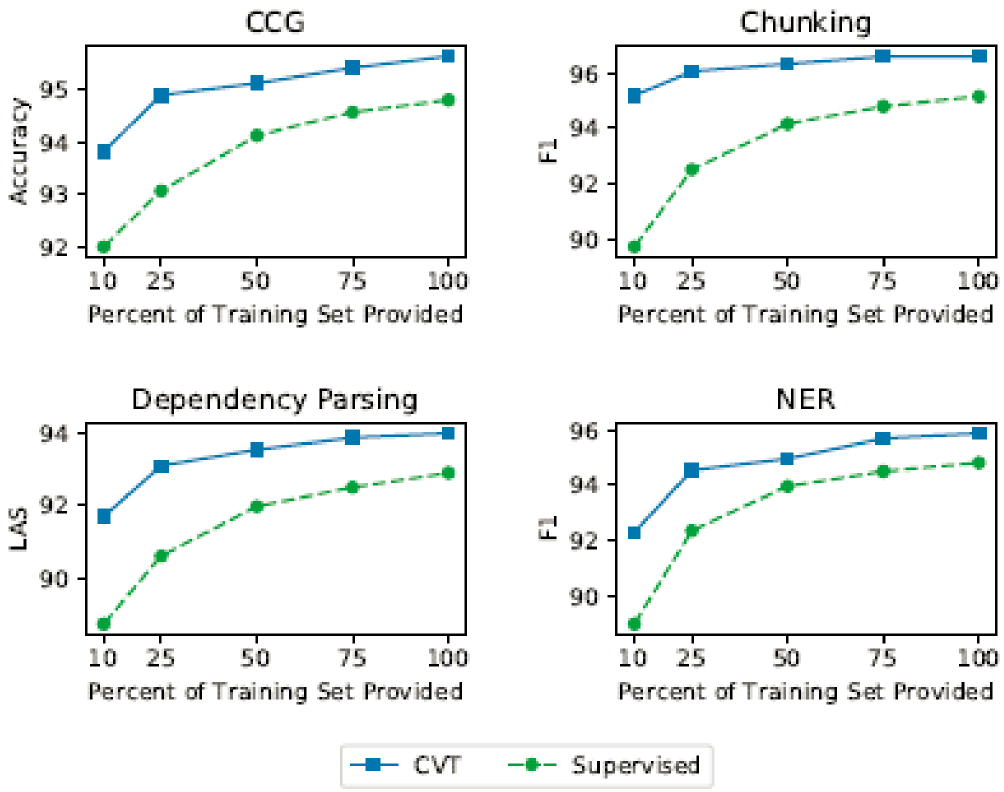
Results

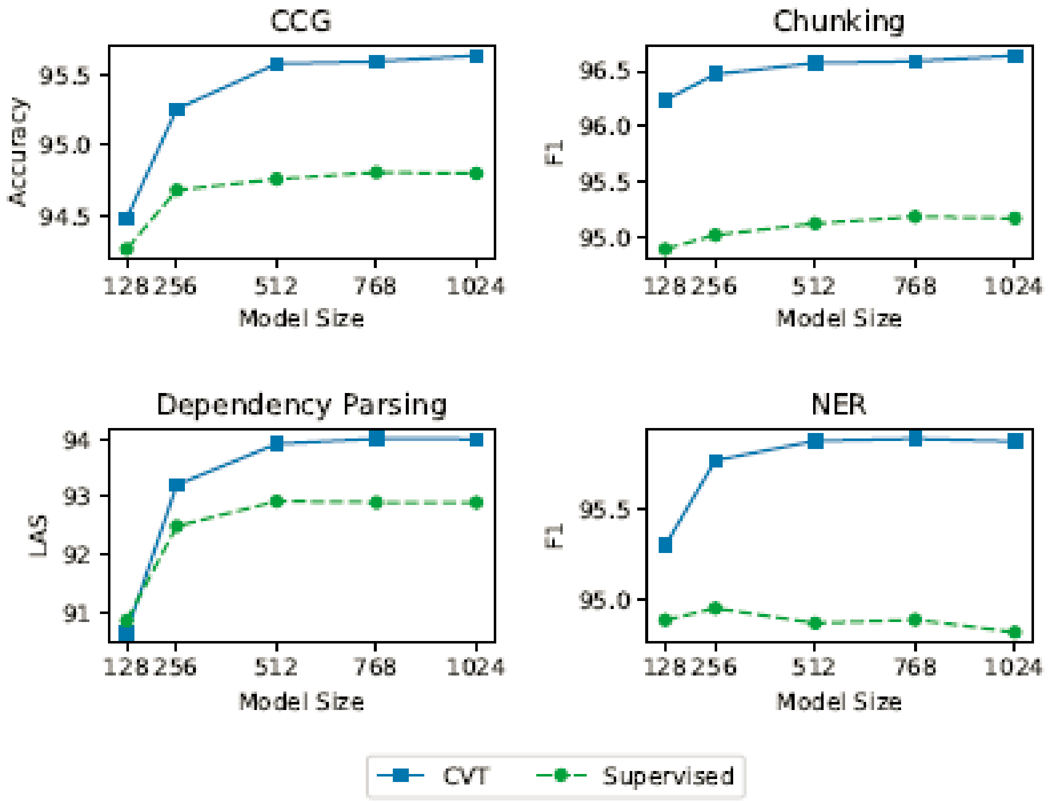
Method	CCG Acc.	Chunk F1	NER F1	FGN F1	POS Acc.	Dep. Parse UAS	Dep. Parse LAS	Translate BLEU
Shortcut LSTM (Wu et al., 2017)	95.1				97.53			
ID-CNN-CRF (Strubell et al., 2017)			90.7	86.8				
JMT† (Hashimoto et al., 2017)		95.8			97.55	94.7	92.9	
TagLM* (Peters et al., 2017)		96.4	91.9					
ELMo* (Peters et al., 2018)			92.2					
Biaffine (Dozat and Manning, 2017)						95.7	94.1	
Stack Pointer (Ma et al., 2018)						95.9	94.2	
Stanford (Luong and Manning, 2015)								23.3
Google (Luong et al., 2017)								26.1
Supervised	94.9	95.1	91.2	87.5	97.60	95.1	93.3	28.9
Virtual Adversarial Training*	95.1	95.1	91.8	87.9	97.64	95.4	93.7	–
Word Dropout*	95.2	95.8	92.1	88.1	97.66	95.6	93.8	29.3
ELMo (our implementation)*	95.8	96.5	92.2	88.5	97.72	96.2	94.4	29.3
ELMo + Multi-task*†	95.9	96.8	92.3	88.4	**97.79**	96.4	94.8	–
CVT*	95.7	96.6	92.3	88.7	97.70	95.9	94.1	**29.6**
CVT + Multi-task*†	96.0	96.9	92.4	88.4	97.76	96.4	94.8	–
CVT + Multi-task + Large*†	**96.1**	**97.0**	**92.6**	**88.8**	97.74	**96.6**	**95.0**	–

Model Generalization



Datasets Size & Model Size





Multi-Task & Auxiliary Module

Model	CCG	Chnk	NER	FGN	POS	Dep.
CVT-MT	95.7	97.4	96.0	86.7	97.74	94.4
w/out parallel	95.4	97.1	95.6	86.3	97.71	94.1

Model	CCG	Chnk	NER	FGN	POS
Supervised	94.8	95.5	95.0	86.0	97.59
CVT	95.6	97.0	95.9	87.3	97.66
no fwd/bwd	–0.1	–0.2	–0.2	–0.1	–0.01
no future/past	–0.3	–0.4	–0.4	–0.3	–0.04

Generalizable Representations

Model	CCG	Chnk	NER	FGN	POS	Dep.
Supervised	94.8	95.6	95.0	86.0	97.59	92.9
CVT-MT frozen	95.1	96.6	94.6	83.2	97.66	92.5
ELMo frozen	94.3	92.2	91.3	80.6	97.50	89.4