Variational Sequential Labelers for Semi-Supervised Learning

Mingda Chen, Qingming Tang, Karen Livescu, Kevin Gimpel

TOYOTA TECHNOLOGICAL INSTITUTE AT CHICAGO

Sequence Labeling

Part-of-Speech (POS) Tagging

determiner	noun	verb	determiner	adjective	noun	coordinating conjunction	adverb	verb	punctuation
This	item	is	a	small	one	and	easily	missed	.

Named Entity Recognition (NER)

B-ORG	O	B-MISC	O	O	O	B-MISC	O	O
EU	rejects	German	call	to	boycott	British	lamb	.

Overview

- Latent-variable generative models for sequence labeling
- 0.8 ~ 1% absolute improvements over 8 datasets without structured inference
- 0.1 ~ 0.3% absolute improvements from adding unlabeled data

Why latent-variable models?

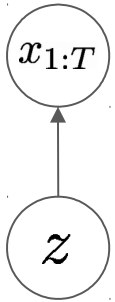
- Natural way to incorporate unlabeled data
- Ability to disentangle representations via the configuration of latent variables
- Allow us to use neural variational methods

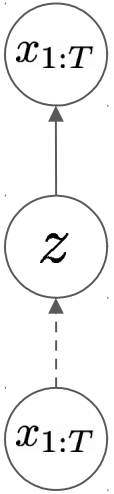
Variational Autoencoder (VAE)

[Kingma and Welling, ICLR'14; Rezende and Mohamed, ICML'15]

Observation

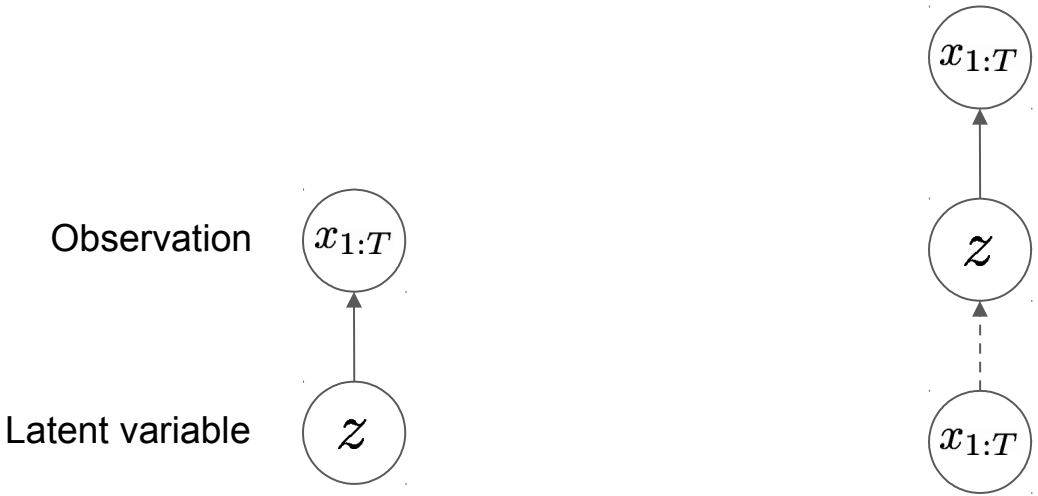
Latent variable





Variational Autoencoder (VAE)

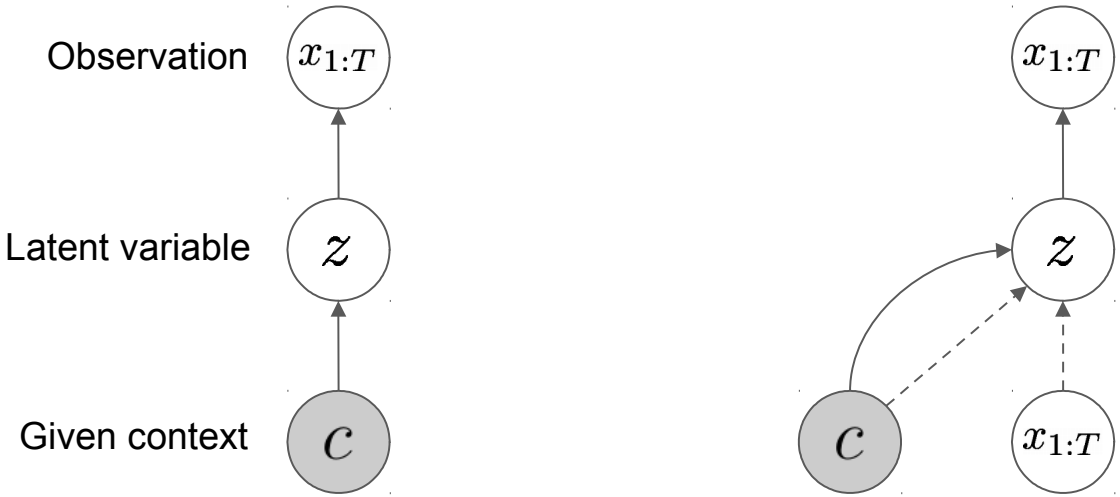
[Kingma and Welling, ICLR'14; Rezende and Mohamed, ICML'15]



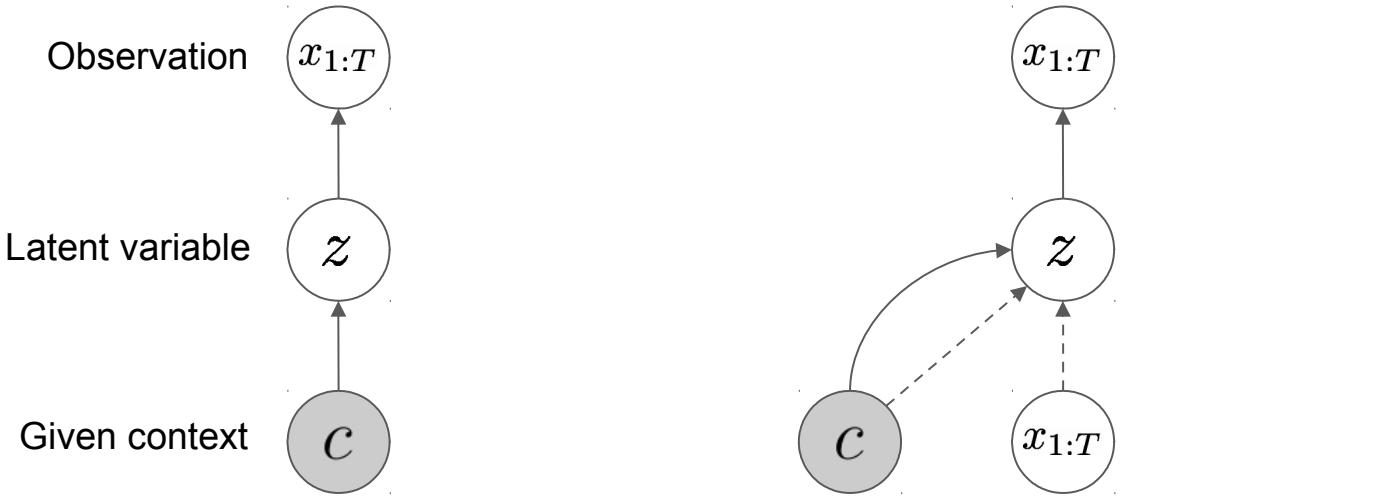
$$\log\ p_\theta(x_{1:T}) \geq \underbrace{\mathop{\mathbb{E}}_{z\sim q_\phi(\cdot|x_{1:T})}[\log p_\theta(x_{1:T}|z)]}_{\text{Reconstruction Loss}} - \underbrace{KL(q_\phi(z|x_{1:T})\|\ p_\theta(z))}_{\text{KL divergence}}$$

Evidence Lower Bound (ELBO)

Conditional Variational Autoencoder



Conditional Variational Autoencoder



$$\log\ p_\theta(x_{1:T}|c) \geq \mathop{\mathbb{E}}_{z\sim q_\phi(\cdot|x_{1:T},c)}[\log p_\theta(x_{1:T}|z)] - KL(q_\phi(z|x_{1:T},c)\|\ p_\theta(z|c))$$

$$x_{-t}$$

The input words other than the word at position t

$$x_{-t}$$

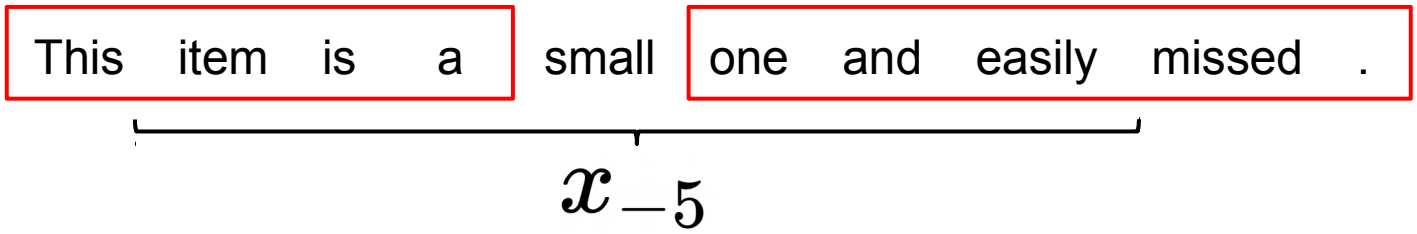
The input words other than the word at position t

This | item | is | a | small | one | and | easily | missed | .

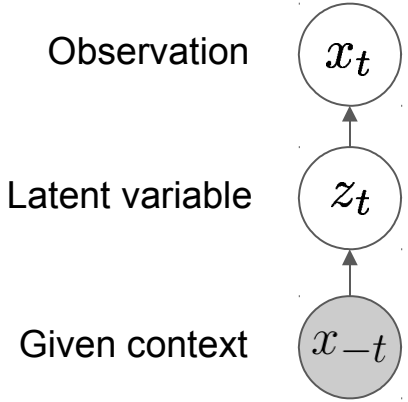
$$x_{-1}$$

$$x_{-t}$$

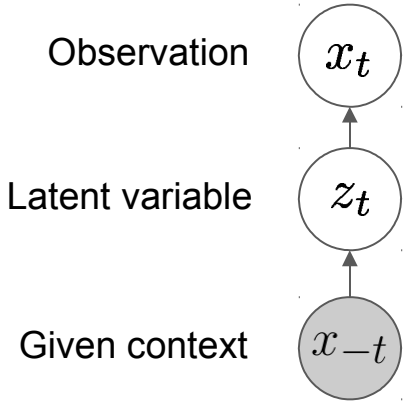
The input words other than the word at position t



Variational Sequential Labeler (VSL)

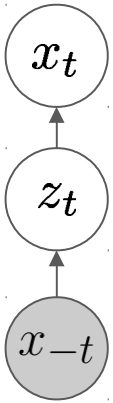


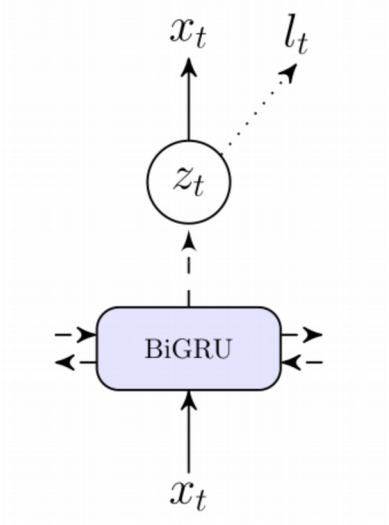
Variational Sequential Labeler (VSL)



$$\log\ p_\theta(x_t|x_{-t}) \geq \underbrace{\mathop{\mathbb{E}}_{z_t \sim q_\phi(\cdot|\ x_{1:T}, t)} [\log p_\theta(x_t|\ z_t)] - KL(q_\phi(z_t|\ x_{1:T}, t) \|\ p_\theta(z_t|\ x_{-t}))}_{\text{ELBO}}$$

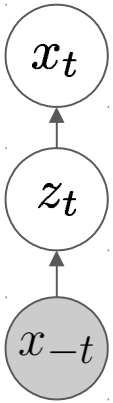
Variational Sequential Labeler (VSL)

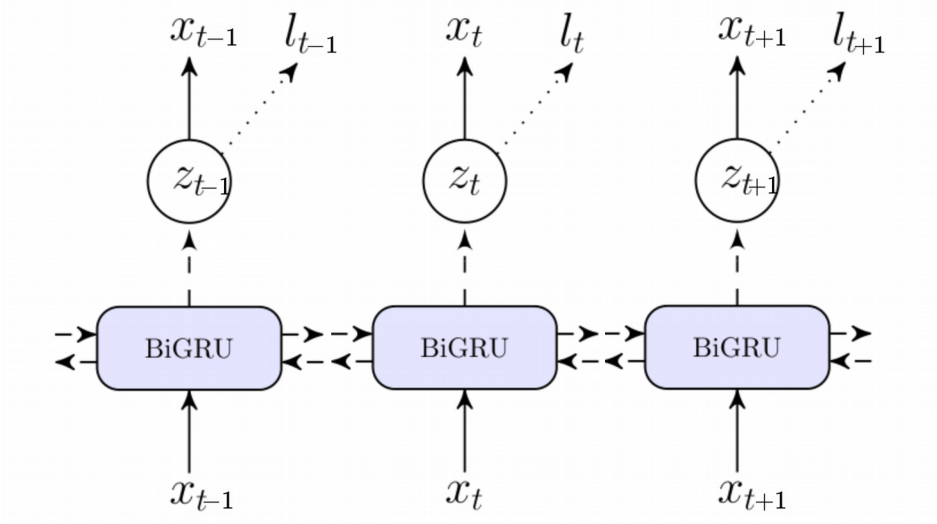




$$\log\ p_\theta(x_t|x_{-t}) \geq \mathop{\mathbb{E}}_{z_t \sim q_\phi(\cdot|\ x_{1:T}, t)} [\log p_\theta(x_t|\ z_t)] - KL(q_\phi(z_t|\ x_{1:T}, t) \|\ p_\theta(z_t|\ x_{-t}))$$

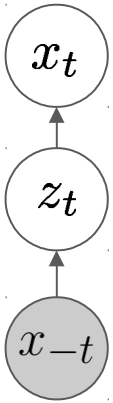
Variational Sequential Labeler (VSL)

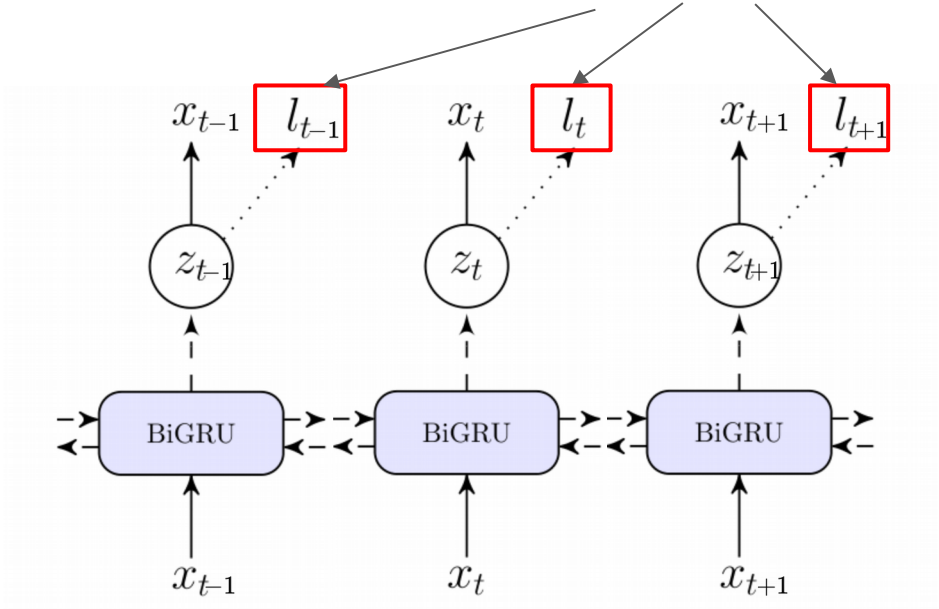




$$\log\ p_\theta(x_t|x_{-t}) \geq \mathop{\mathbb{E}}_{z_t \sim q_\phi(\cdot|\ x_{1:T}, t)} [\log p_\theta(x_t|\ z_t)] - KL(q_\phi(z_t|\ x_{1:T}, t) \|\ p_\theta(z_t|\ x_{-t}))$$

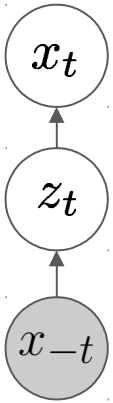
Variational Sequential Labeler (VSL)

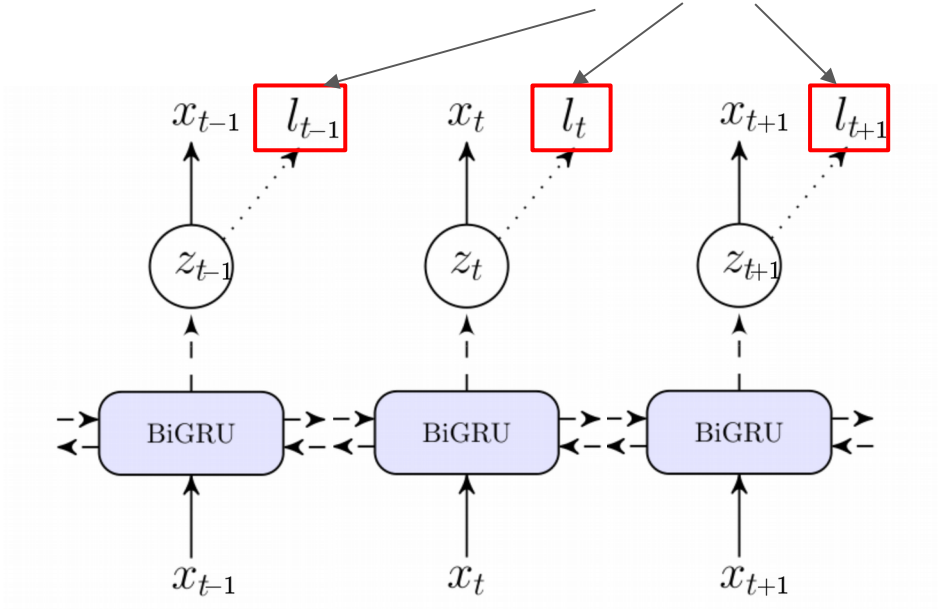




$$\log\ p_\theta(x_t|x_{-t}) \geq \mathop{\mathbb{E}}_{z_t \sim q_\phi(\cdot|\ x_{1:T}, t)} [\log p_\theta(x_t|\ z_t)] - KL(q_\phi(z_t|\ x_{1:T}, t) \|\ p_\theta(z_t|\ x_{-t}))$$

Variational Sequential Labeler (VSL)





$$\log\ p_\theta(x_t|x_{-t}) \geq \mathop{\mathbb{E}}_{z_t \sim q_\phi(\cdot|\ x_{1:T},t)} [\log p_\theta(x_t|\ z_t)] - KL(q_\phi(z_t|\ x_{1:T},t) \| \boxed{p_\theta(z_t|\ x_{-t})})$$

VSL: Training and Testing

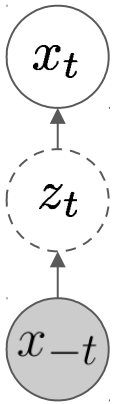
Training

- Maximize $\mathrm{ELBO} - \alpha \cdot \mathrm{CL}$ where α is a hyperparameter
- Use one sample from Gaussian distribution using reparameterization trick

Testing

- Use the mean of Gaussian distribution

Variants of VSL

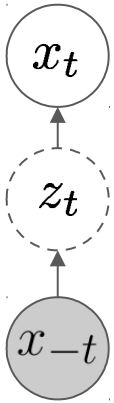


VSL-G

Position of classifier

Variants of VSL

Position of classifier

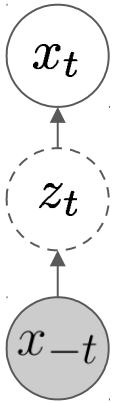


VSL-G

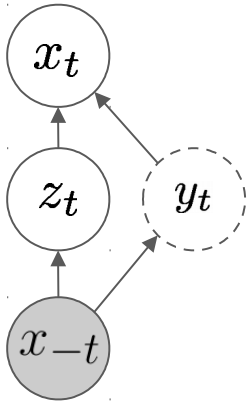
Stands for “Gaussian”

Variants of VSL

Position of classifier



VSL-G

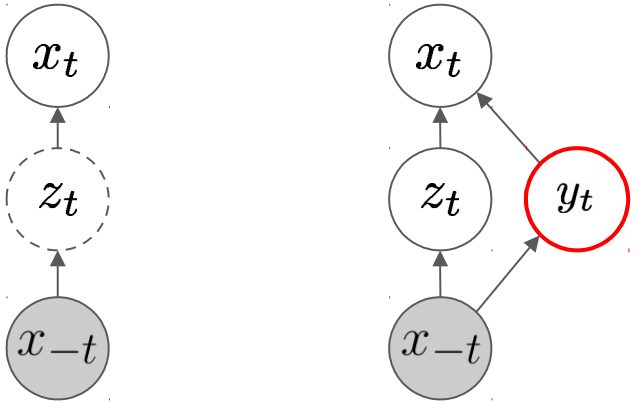


VSL-GG-Flat

Stands for “Gaussian”

Variants of VSL

Position of classifier



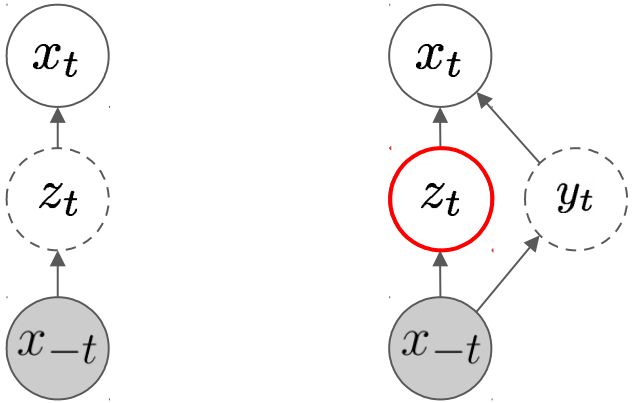
VSL-G

VSL-GG-Flat

Stands for “Gaussian”

Variants of VSL

Position of classifier



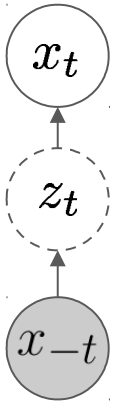
VSL-G

VSL-GG-Flat

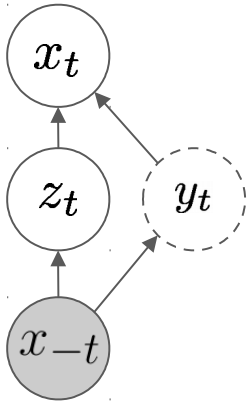
Stands for “Gaussian”

Variants of VSL

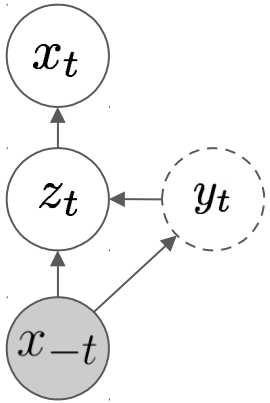
Position of classifier



VSL-G



VSL-GG-Flat

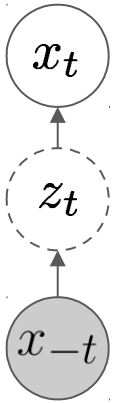


VSL-GG-Hier

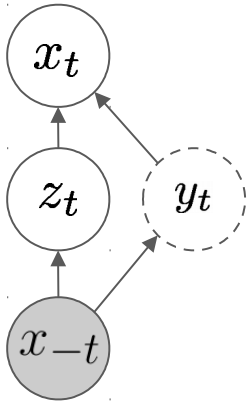
Stands for “Gaussian”

Variants of VSL

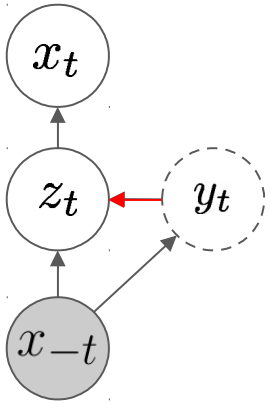
Position of classifier



VSL-G



VSL-GG-Flat



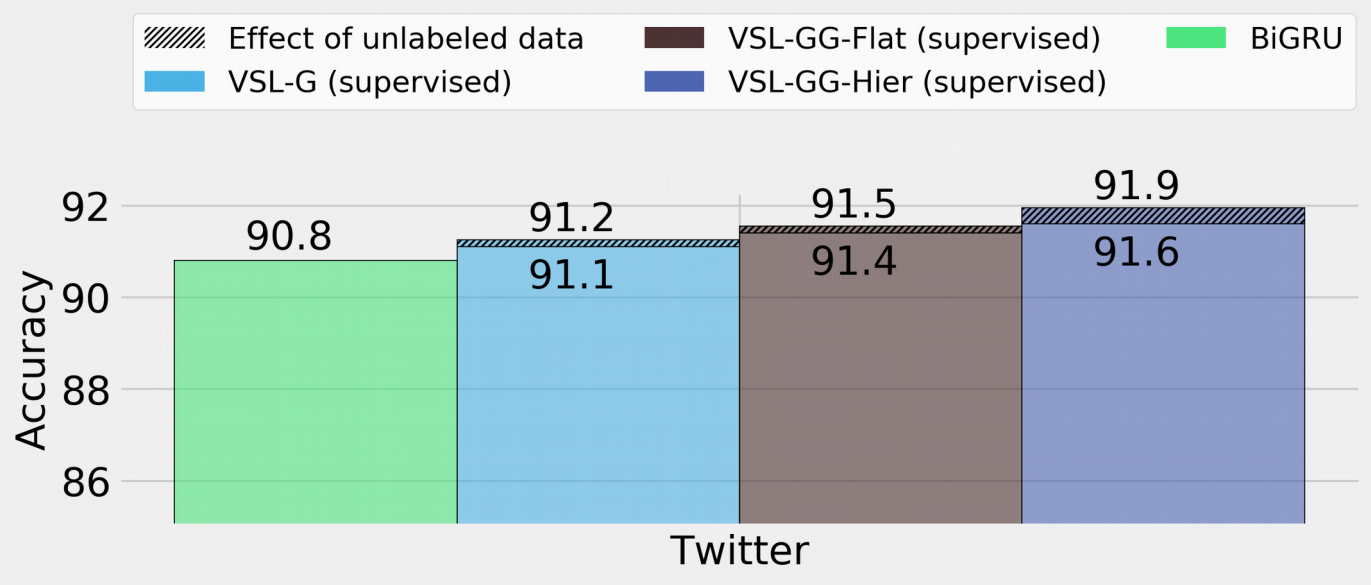
VSL-GG-Hier

Stands for “Gaussian”

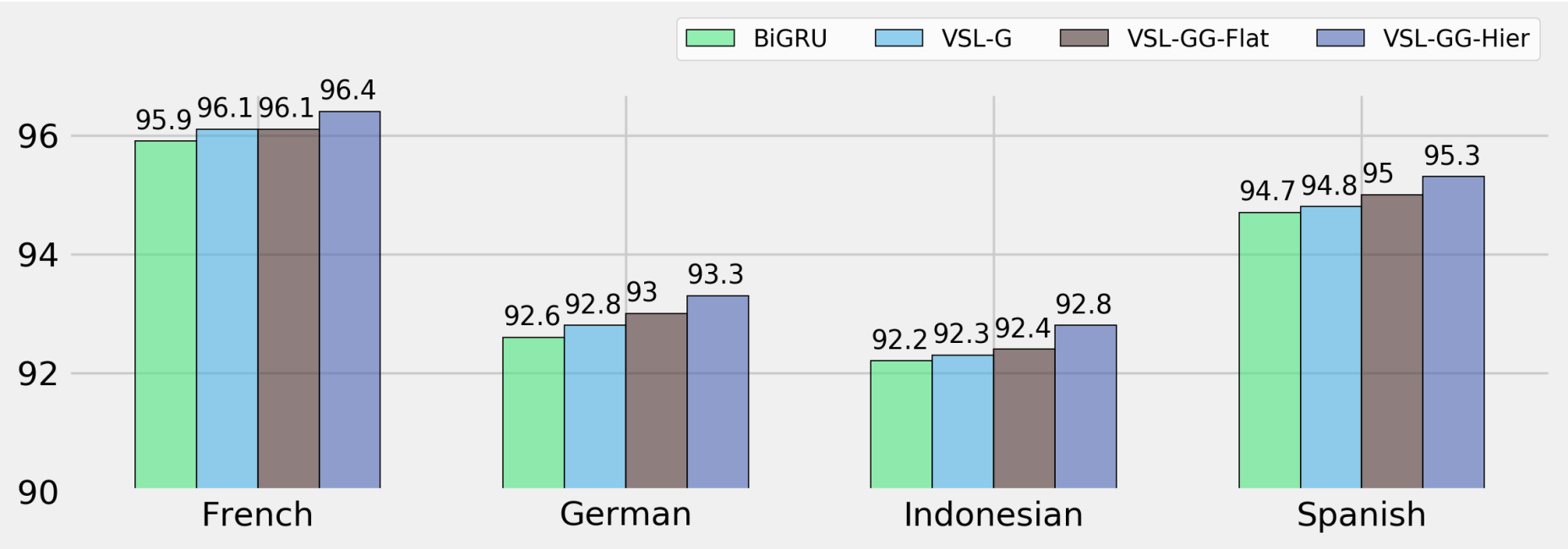
Experiments

- Twitter POS Dataset
  - Subset of 56 million English tweets as unlabeled data
  - 25 tags
- Universal Dependencies POS Datasets
  - 20% of original training set as labeled data
  - 50% of original training set as unlabeled data
  - 6 languages
  - 17 tags
- CoNLL 2003 English NER Dataset
  - 10% of original training set as labeled data
  - 50% of original training set as unlabeled data
  - BIOES labeling scheme

Results

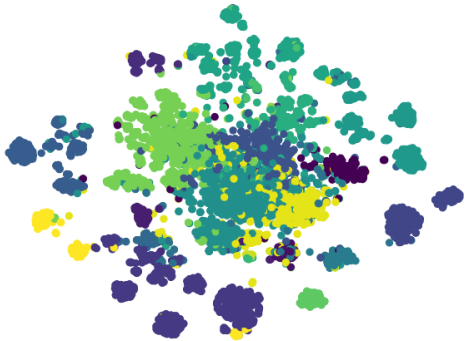

Effect of unlabeled data
VSL-G (supervised)
VSL-GG-Flat (supervised)
VSL-GG-Hier (supervised)
BiGRU
Accuracy
92
90
88
86
90.8
91.2
91.1
91.5
91.4
91.9
91.6
Twitter
F1
88
87
86
85
87.6
87.9
87.8
88.1
88.0
88.6
88.4
NER

Universal Dependencies POS



t-SNE Visualization

- Each point represents a word token
- Color indicates gold standard POS tag in Twitter dev set

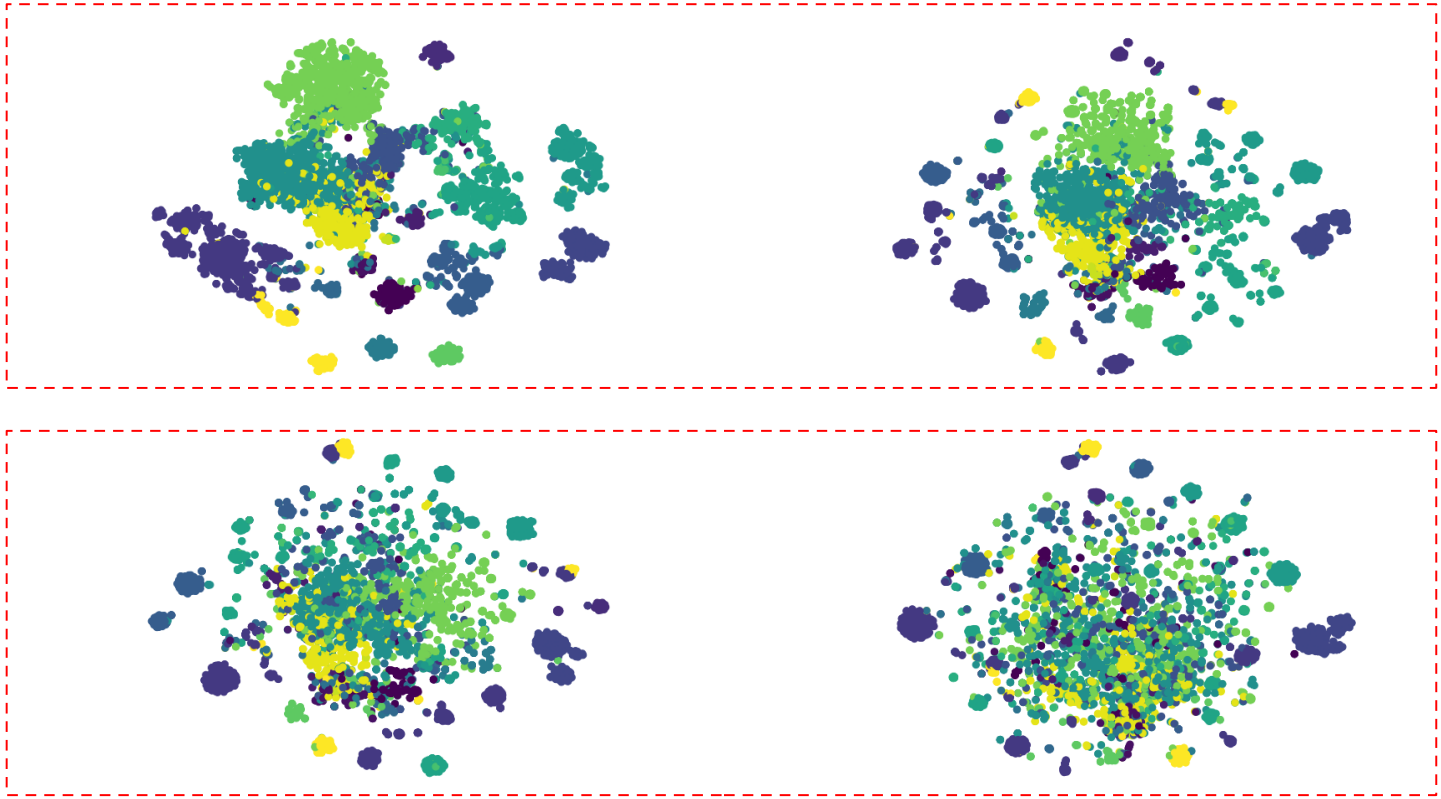


BiGRU baseline

t-SNE Visualization

y (label) variable

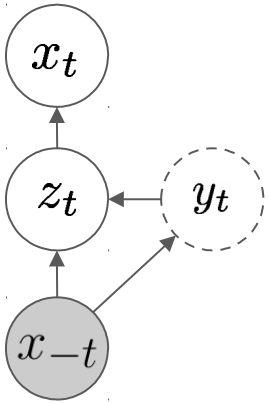
z variable



VSL-GG-Hier

VSL-GG-Flat

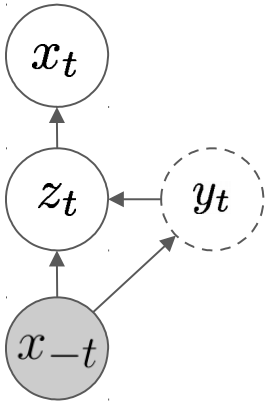
Effect of Position of Classification Loss



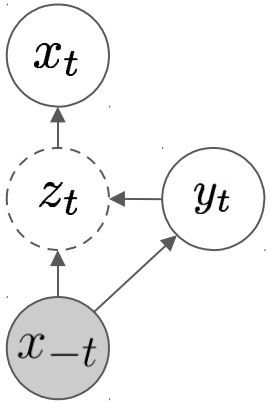
VSL-GG-Hier

Position of classifier

Effect of Position of Classification Loss



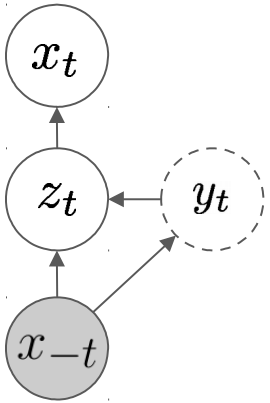
VSL-GG-Hier



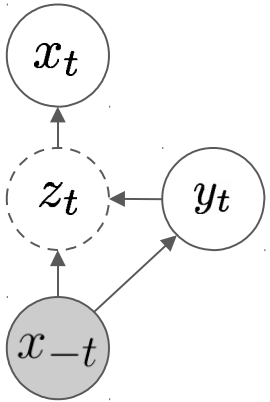
VSL-GG-Hier with classifier on z_t

Position of classifier

Effect of Position of Classification Loss



VSL-GG-Hier

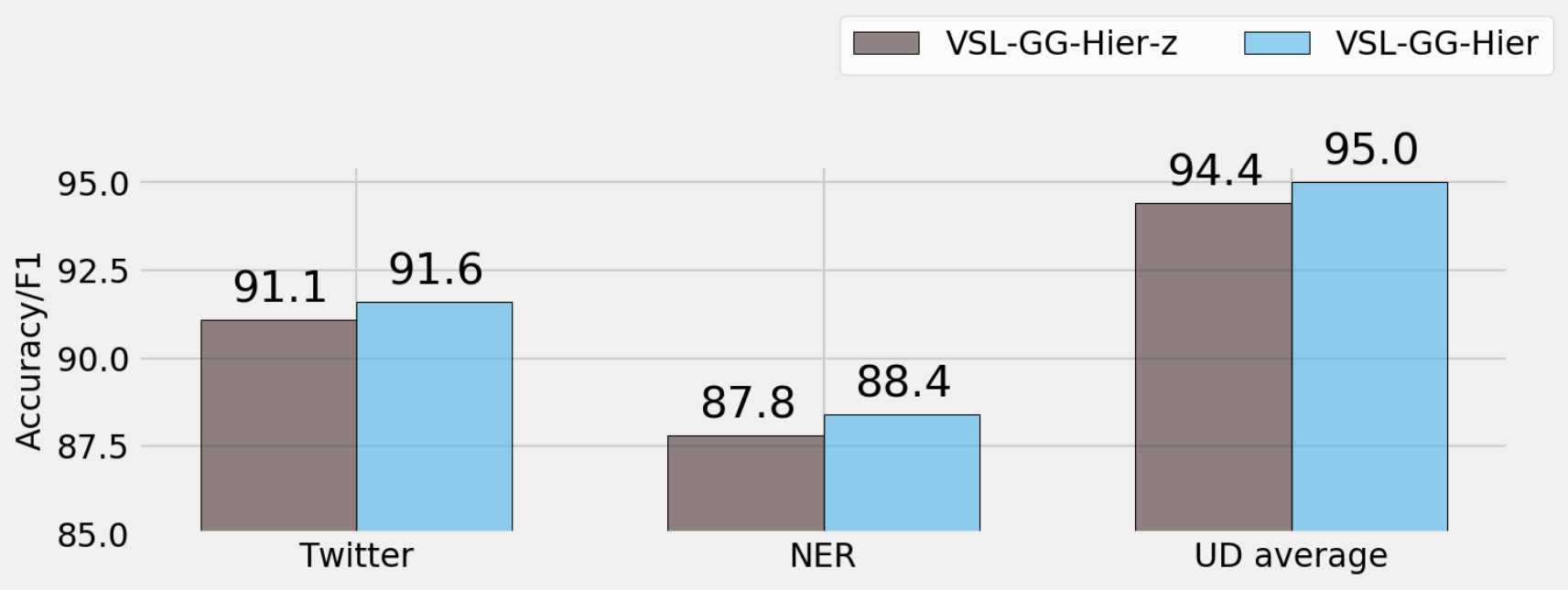


VSL-GG-Hier with classifier on z_t

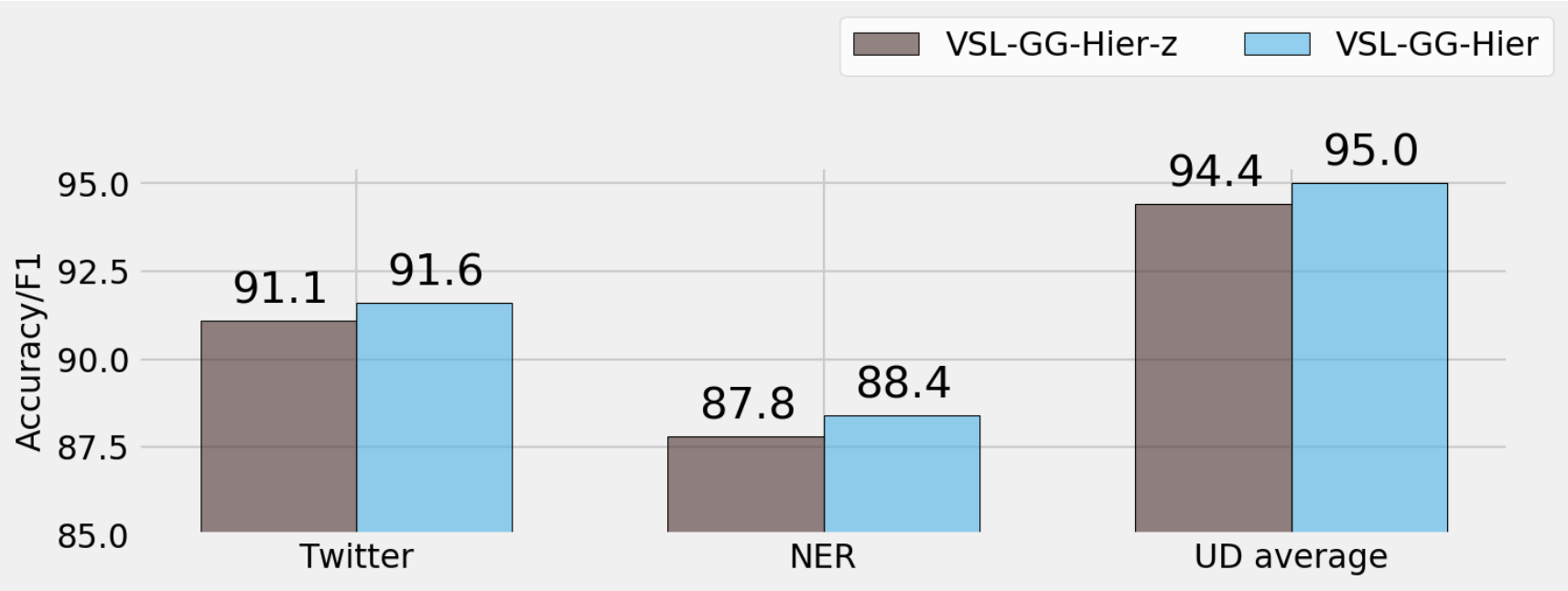
VSL-GG-Hier-z

Position of classifier

Effect of Position of Classification Loss



Effect of Position of Classification Loss



Hierarchical structure is only helpful when classification loss and reconstruction loss are attached to different latent variables

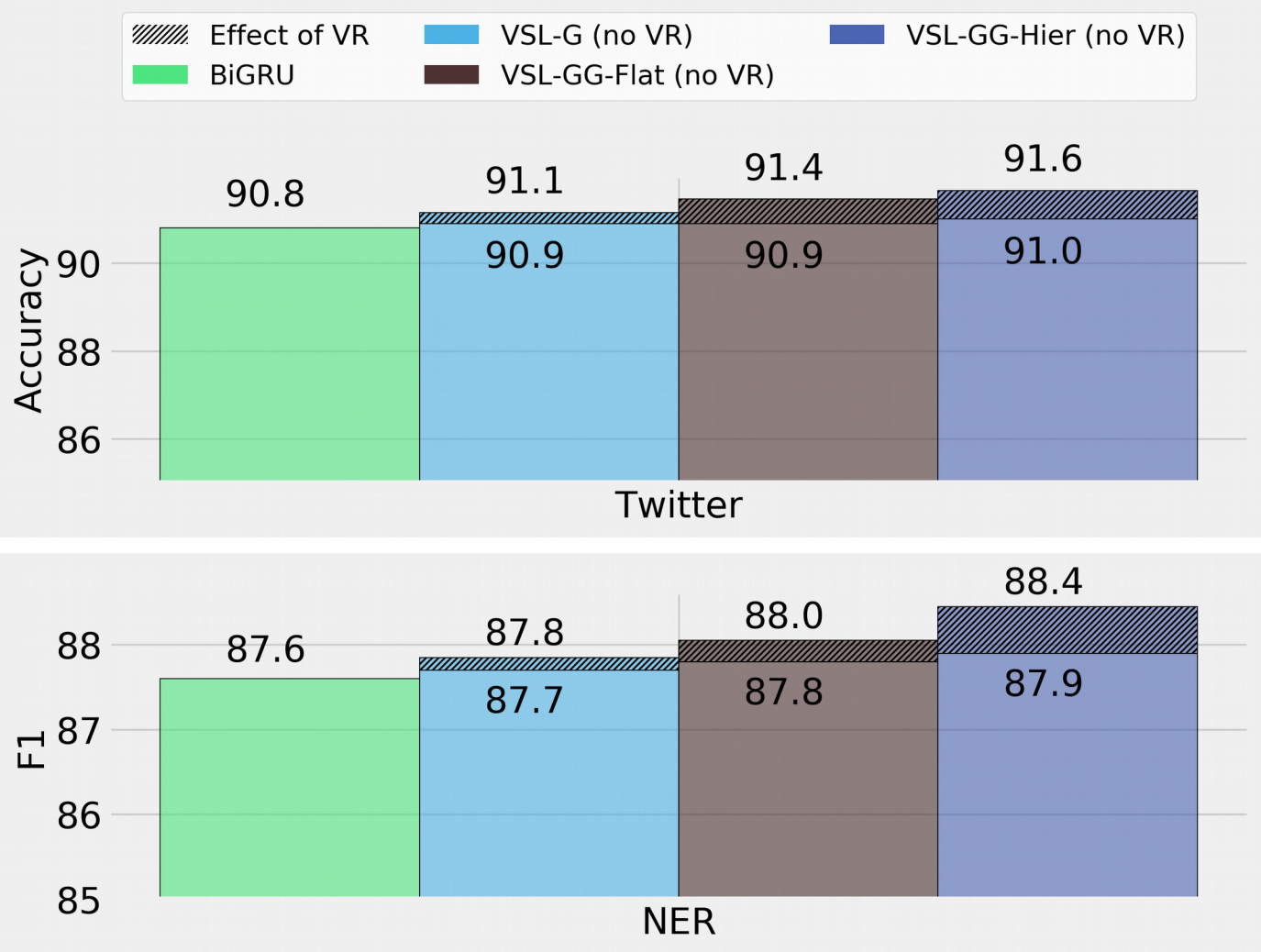
Effect of Variational Regularization (VR)

VR

KL divergence between approximated posterior and prior

Randomness in the latent space

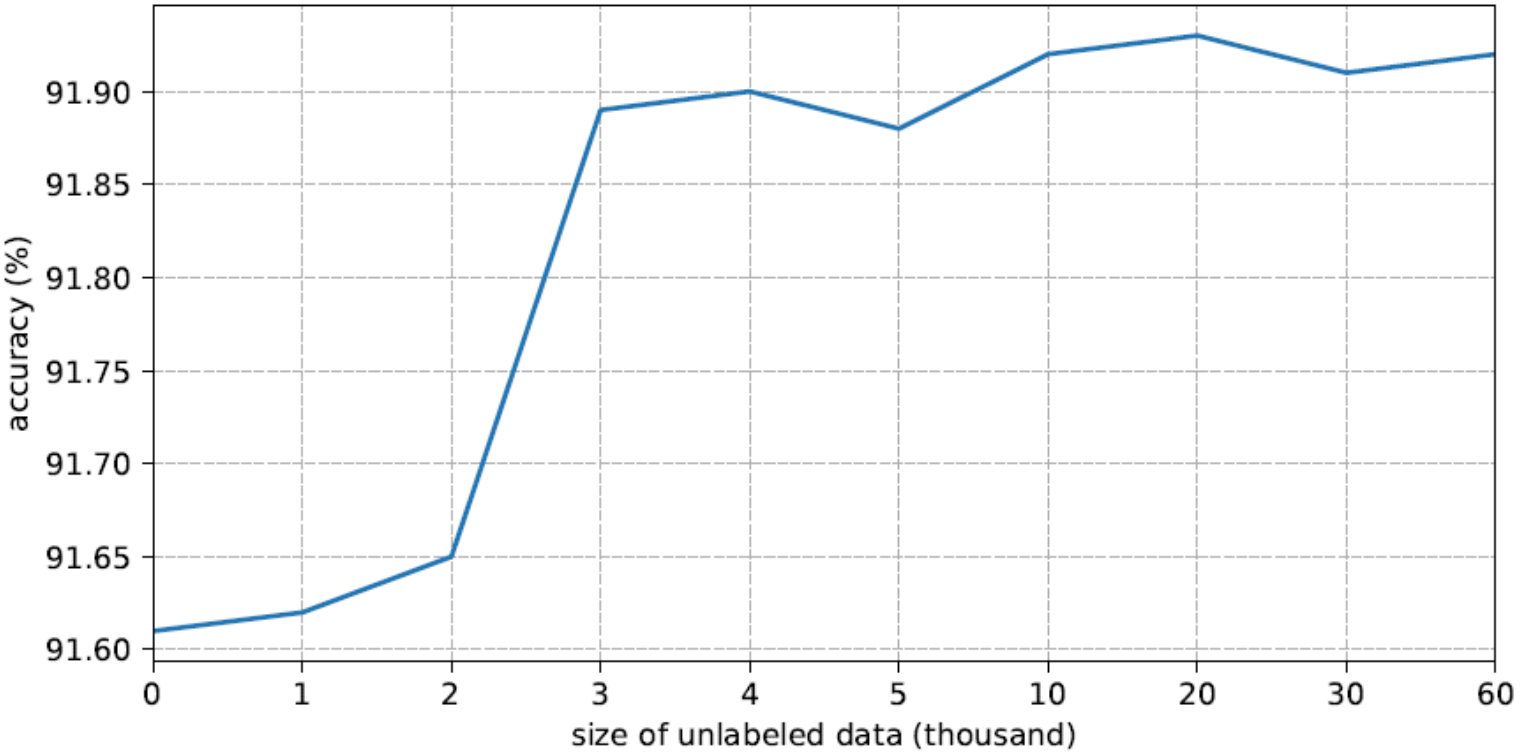
Effect of VR


Effect of VR
BiGRU
VSL-G (no VR)
VSL-GG-Flat (no VR)
VSL-GG-Hier (no VR)
Accuracy
90
88
86
90.8
91.1
90.9
91.4
90.9
91.6
91.0
Twitter
F1
88
87
86
85
87.6
87.8
87.7
88.0
87.8
88.4
87.9
NER

Effect of Unlabeled data

- Evaluate VSL-GG-Hier on Twitter dataset
- Subsample unlabeled data from 56 million tweets
- Vary the number of unlabeled data

Effect of Unlabeled data



Summary

- We introduced VSLs for semi-supervised learning
- Best VSL uses multiple latent variable and arranged in hierarchical structure
- Hierarchical structure is only helpful when classification loss and reconstruction loss are attached to different latent variables
- VSLs show consistent improvements across 8 datasets over a strong baseline

Thank you!