Specializing Word Embeddings (for Parsing) by Information Bottleneck

Best Paper Award in EMNLP 2019

Li and Eisner, 2019

Gangwoo Kim

Data Mining & Information Systems Lab.
Department of Computer Science and Engineering,
College of Informatics, Korea University

0 Introduction

Abstract

a variational information bottleneck(VIB) method to nonlinearly compress contextual embedding, keeping **only** information that helps down stream tasks

compress each embeddings to either **discrete** tag or a **continuous** vector

Properties of Variational Information Bottleneck (VIB)

1. exploiting the *existing* information
   <-> Fine-tuning, introducing new info.
   → less danger of overfitting / very fast

2. Stochasticity
   <-> Dimension reduction techniques, deterministic
   → blurring unneeded capacity via randomness

0 Introduction

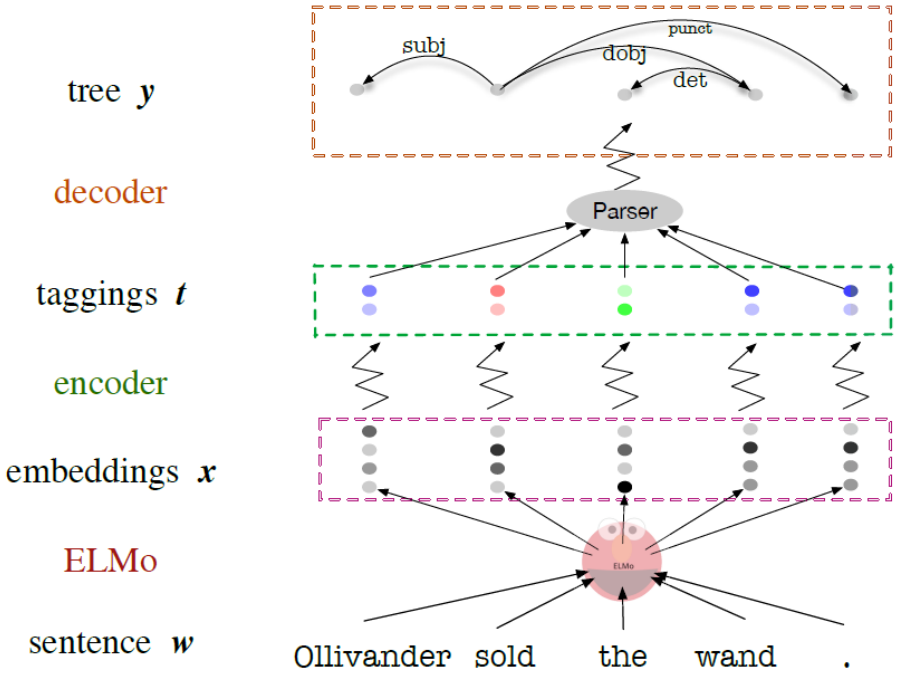


Figure 1: Our instantiation of the information bottleneck, with bottleneck variable T. A jagged arrow indicates a stochastic mapping, i.e. the jagged arrow points from the parameters of a distribution to a sample drawn from that distribution.

1 Background: Information Bottleneck

Usage

The information bottleneck (IB) method originated in information theory and has been adopted in macine learning

as a training objective

N. Tishby et al, 1999, “The information bottleneck method”.

and a framework for analyzing deep neural network

N. Tishby et al, 2015, “Deep Learning and the Information Bottleneck Principle”

Our goal is to learn a stochastic map from X to some compressed representation T with

1) leave us only the parts of X relevant to Y

2) squeeze out any information in X irrelevant to Y

both are measured by the mutual information

$$\min_{p(t|x):I(T;Y)\geq I^*} I(X;T)$$

2 Method

$$\min_{p(t|x):I(T;Y)\geq I^*} I(X;T)$$

$$\mathcal{L}_{IB} = -\mathrm{I}(Y;T) + \beta\,\mathrm{I}(X;T) + \gamma \sum_{i=1}^{n} \mathrm{I}(T_i; X|\hat{X}_i) \quad (2)$$

, where T_i is the tag associated with the i-th word, X_i is the ELMo embedding and $\hat{X}_i$ is the type embedding (typical embedding)

We extend the original IB objective (1) and add terms to control the context-sensitivity of the extracted tags.

We instantiate the variational IB estimation method for each of

the Decoder / the Token Encoder / the Type Encoder

2 Method

$I(X;T)$ – the Token Encoder $p_\theta(t|x)$

$$\mathbb{E}_{x,t}\left[\log \frac{p_\theta(t|x)}{p_\theta(t)}\right] = \mathbb{E}_x\left[\mathbb{E}_{t\sim p_\theta(t|x)}\left[\log \frac{p_\theta(t|x)}{p_\theta(t)}\right]\right].$$

The outer expectation is over the true distribution of sentences x
→ an emprical estimate

To estimate the inner expectation, we could sample, drawing taggings t

The troublesome term is $p_\theta(t) = \mathbb{E}_{x'}\left[p_\theta(t \mid x')\right],$

$$\overbrace{\mathop{\mathbb{E}}_{x}\left[\mathop{\mathbb{E}}_{t\sim p_\theta(t|x)}\left[\log \frac{p_\theta(t|x)}{r_\psi(t)}\right]\right]}^{\text{upper bound}} - \overbrace{\mathop{\mathbb{E}}_{x}\left[\mathop{\mathbb{E}}_{t\sim p_\theta(t|x)}\left[\log \frac{p_\theta(t \mid x)}{p_\theta(t)}\right]\right]}^{\text{I}(X;T)}$$
$$= \mathop{\mathbb{E}}_{x}\left[\mathrm{KL}(p_\theta(t) \,||\, r_\psi(t))\right] \geq 0$$

2 Method

$I(T;X)$ – Two Token Encoder Architectures

To obtain continuous tags, define p_θ such that t_i is Gaussian-distributed

To obtain discrete tags, define p_θ such that t_i follows a softmax distribution

2 Method

$I(T; X \mid \hat{X})$ – **the Type Encoder** $s_\xi(t_i|\hat{x}_i)$

We were concerned that it might not be interpretable as a tag of word I specifically
→ no guarantee that contextual info. came from word i not its neighbors

Also we do want tag to consider context
→ use a word type embedding that does not depend on context – ELMo's level-0 embedding of word

$$\overbrace{\mathop{\mathbb{E}}_{x}\Big[\mathop{\mathbb{E}}_{t_i\sim p_\theta(t_i|x)}\Big[\log\frac{p_\theta(t_i|x)}{s_\xi(t_i|\hat{x}_i)}\Big]\Big]}^{\text{upper bound}} - \overbrace{\mathop{\mathbb{E}}_{x}\Big[\mathop{\mathbb{E}}_{t_i\sim p_\theta(t_i|x)}\Big[\log\frac{p_\theta(t_i|x)}{p_\theta(t_i|\hat{x}_i)}\Big]\Big]}^{\mathrm{I}(T_i;X|\hat{X}_i)}$$
$$= \mathop{\mathbb{E}}_{x}\left[\mathrm{KL}(p_\theta(t_i \mid \hat{x}_i) \,||\, s_\xi(t_i \mid \hat{x}_i))\right] \geq 0$$

2 Method

$I(Y;T)$ – the Decoder $q_\varphi(y|t)$

$$\overbrace{\mathop{\mathbb{E}}_{y,t\sim p_\theta}\left[\log \frac{p_\theta(y|t)}{p(y)}\right]}^{\mathrm{I}(Y;T)} - \overbrace{\mathop{\mathbb{E}}_{y,t\sim p_\theta}\left[\log \frac{q_\phi(y|t)}{p(y)}\right]}^{\text{lower bound}}$$
$$= \mathop{\mathbb{E}}_{t\sim p_\theta}\left[\mathrm{KL}(p_\theta(y \mid t) \,||\, q_\phi(y \mid t))\right] \geq 0$$

Architecture

This parser uses a Bi-LSTM to extract features from compressed tags or vectors and assign scores to each tree edge

During training, the decoder compute only an approximation to $q_\varphi(y|t)$ for the golden tree y

3 Training and Inference

Objective Function

$$\mathcal{L}_{IB} = -\mathrm{I}(Y;T) + \beta\,\mathrm{I}(X;T) + \gamma \sum_{i=1}^{n} \mathrm{I}(T_i; X|\hat{X}_i) \quad (2)$$

$$\mathop{\mathbb{E}}_{x,y}\Big[\mathop{\mathbb{E}}_{t\sim p_\theta(t|x)}[-\log q_\phi(y|t)] + \beta\,\mathrm{KL}(p_\theta(t|x)||r_\psi(t)) + \gamma \sum_{i=1}^{n} \mathrm{KL}(p_\theta(t_i \mid x) \,||\, s_\xi(t_i \mid \hat{x}_i))\Big] \quad (3)$$

Details

We must apply the re-parametrization trick to backpropagate
We use the Gumbel-softmax variant for discrete t

4 Experimental Setup

Objective Function

We will examine our compressed tags on a subset of **Universal Dependencies(UD),** a collection of dependency treebanks across 76 languages using the same POS tags and dependency labels

We alternate between improving the model on even epochs and the variational distributions on odd epochs

5 Scientific Evaluation

1. We show the relationship between I(Y;T) and I(X;T) on English

2. Across 9 languages, we study how our automatic tags correlates with gold POS tags

3. We also show how decreasing β gradually refines the automatic discrete tag set

5 Scientific Evaluation

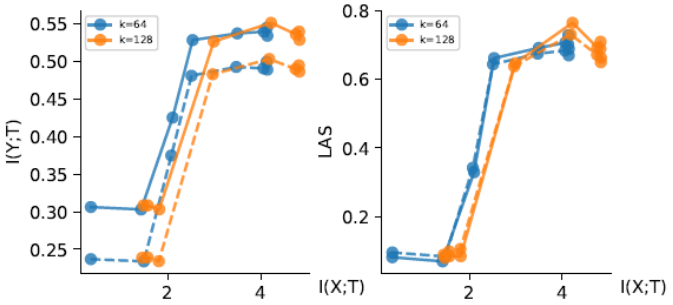
1. **We show the relationship between I(Y;T) and I(X;T) on English**

2. Across 9 languages, we study how our automatic tags correlates with gold POS tags

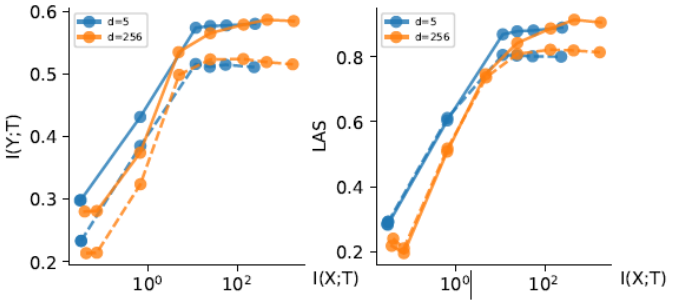
3. We also show how decreasing β gradually refines the automatic discrete tag set

5 Scientific Evaluation

1. **We show the relationship between I(Y;T) and I(X;T) on English**



(a) Discrete Version



(b) Continuous Version

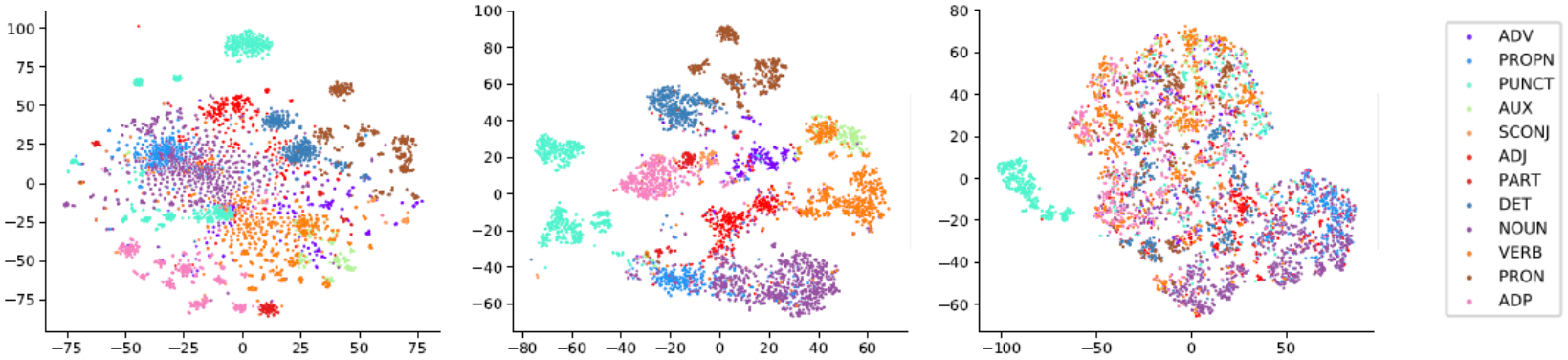
Diminishing returns
after some point the additional info. does not contribute much to predicting Y

The simpler, the better
at each level of I(X, T), low-dimensional one perform on par with hig-dimensional one

5 Scientific Evaluation

2. Across 9 languages, we study how our automatic tags correlates with gold POS tags

Continuous Version



(a) ELMo, $I(X;T) = H(X) \approx 400.6$ (b) $I(X;T) \approx 24.3$ (c) $I(X;T) \approx 0.069$

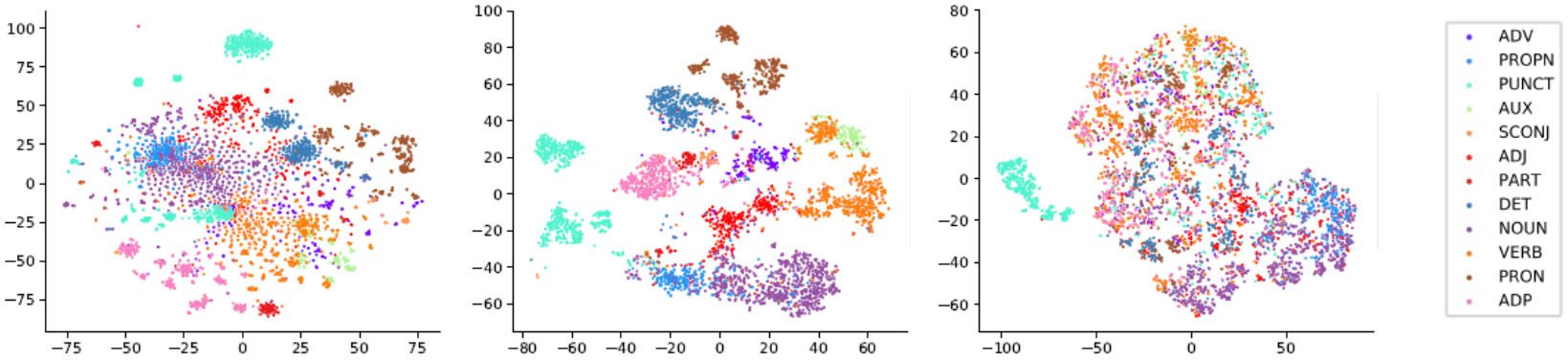
Figure 3: t-SNE visualization of VIB model ($d = 256$) on the projected space of the continuous tags. Each marker in the figure represents a word token, colored by its gold POS tag. This series of figures (from left to right) shows a progression from no compression to moderate compression and to too-much compression.

- Note that the gold POS tags were not used in training
- An interesting observation is that NOUN and PROPN
- heavier compression squeezes out information

5 Scientific Evaluation

2. Across 9 languages, we study how our automatic tags correlates with gold POS tags

Continuous Version



(a) ELMo, $I(X;T) = H(X) \approx 400.6$ (b) $I(X;T) \approx 24.3$ (c) $I(X;T) \approx 0.069$

Figure 3: t-SNE visualization of VIB model ($d = 256$) on the projected space of the continuous tags. Each marker in the figure represents a word token, colored by its gold POS tag. This series of figures (from left to right) shows a progression from no compression to moderate compression and to too-much compression.

- Note that the gold POS tags were not used in training
- An interesting observation is that NOUN and PROPN
- heavier compression squeezes out information

5 Scientific Evaluation

3. We also show how decreasing β gradually refines the automatic discrete tag set
Discrete Version

Deterministic Annealing, a method that gradually decrease β during training

→ results hierarchical structure refelects properties of English syntax

i.e. POS → (anaphors / Possesive pronouns)

6 Engineering Evaluation

Learing how to compress ELMo's tags for a given task is a fast alternative to fine-tuning all the ELMo parameters.

→ With on a single GPU, it is able to train on 10K sents. in 100 sec per epoch

Training our compression method does improve our generalization performance

Models	Arabic	Hindi	English	French	Spanish	Portuguese	Russian	Chinese	Italian
Iden	0.751	**0.870**	0.824	0.784	0.808	0.813	0.783	0.709	**0.863**
PCA	0.743	**0.866**	0.823	0.749	0.802	0.808	0.777	0.697	0.857
MLP	0.759	**0.871**	0.839	0.816	**0.835**	0.821	0.800	0.734	**0.867**
VIBc	**0.779**	**0.866**	**0.851**	**0.828**	**0.837**	**0.836**	**0.814**	**0.754**	**0.867**
POS	0.652	0.713	0.712	0.718	**0.739**	**0.743**	**0.662**	0.510	0.779
VIBd	**0.672**	**0.736**	**0.742**	**0.723**	**0.725**	0.710	**0.651**	**0.591**	**0.781**

Table 2: Parsing accuracy of 9 languages (LAS). Black rows use continuous tags; gray rows use discrete tags (which does worse). In each column, the best score for each color is boldfaced, along with all results of that color that are not significantly worse (paired permutation test, $p < 0.05$). These results use only ELMo layer 1; results from all layers are shown in Table 3 in the appendix, for both LAS and UAS metrics.

1 Learning Setup

task labels

a protected attribute

classifier

attacker

a representation

document

1 Learning Setup

Goal : We want decision y = f(x) to be
oblivious to z

Attacker
After the classifier c(h(x)) is fully trained,
we use the encoder h → z

Definition
A protected attribute has *leaked*
if we can train a classifier *h* → *z*

A protected attribute has *guarded*
if we cannot train it

2 Data, Tasks and Protected Attr.

Corpus - Twitter messages

1. **DIAL**
   - Task : binary emoji-based **sentiment** and binary **tweet-mention** prediction
     Sentiment : Positive v Negative
     Mention : conversational v non-conversational
   - Protected : race of authors
     Race : Light(Standard American English) v Dark (Else)

2. PAN 16
   - Task : to classify a given **dependency relation** between two tokens
   - Protected : Age and gender
     Age : (18-34) v (35+)
     gender : male v female

We collected 160K for training and 10K for development from each.
Each split is balanced with respect to both the main and the protected labels.