Recent Advances in Transferable Representation Learning

The Basics of Embeddings

Muhao Chen, Kai-Wei Chang, Dan Roth

AAAI 2020 Tutorial
Machine Learning Pipelines

Traditional ML framework

Input → Feature Extractor → ML Model → Inference Alg. → Evaluation

Features

Sub-scores (e.g., CRF potentials)
Machine Learning Pipelines

Traditional ML framework

Input ➔ Feature Extractor ➔ ML Model ➔ Inference Alg. ➔ Evaluation

Feature Engineering
Machine Learning Pipelines

Traditional ML Framework

Input → Feature Extractor → ML Model → Inference Alg. → Evaluation

Deep Learning Framework

Input → DNN → Representation → Inference → Evaluation

Representation Learning

Kai-Wei Chang (http://kwchang.net/talks/programming/ssl/)
Transferable Representation

Deep Learning Framework

Input → DNN → Inference → Evaluation

Transferable Representation

Learn representation from large unannotated data

SQuAD2.0 (Rajpurkar & Jia et al. '18)

Packet switching contrasts with another principal networking paradigm, circuit switching, a method which pre-allocates dedicated network bandwidth specifically for each communication session, each having a constant bit rate and latency between nodes. In cases of billable services, such as cellular communication services, circuit switching is characterized by a fee per unit of connection time, even when no data is transferred, while packet switching may be characterized by a fee per unit of information transmitted, such as characters, packets, or messages.

Q: Packet Switching contrast with what other principal
A: circuit switching
A history of Word Representation
How do we present a word?

Credit: https://www.flickr.com/photos/182229932@N07/48688109908
Idea 1: Representing words as discrete symbols

- Represent word as a “one-hot” vector, e.g.,
  \[
  \text{happy} = \begin{bmatrix}
  0 & 0 & 0 & 1 & 0 & \ldots & 0
\end{bmatrix}
  \]
  \[
  \text{glad} = \begin{bmatrix}
  0 & 0 & 1 & 0 & 0 & \ldots & 0
\end{bmatrix}
  \]

- How large is this vector?
  - Vector dimension = number of words in vocabulary
    - PTB data: \(~50k\)
    - Google 1T data: \(13M\)

- Issue: no notion of similarity. \text{happy} and \text{glad} are orthogonal.
Idea 2: Similarity = Clustering

Idea 3: Vector representations

- Discrete ⇒ continuous: A dense vector for each word
- Words with similar meaning are closer in the embedding space
- Word meanings are vector of “basic concepts”
  - The “basic concepts” might not be explicit
  
  \[
  \begin{align*}
  v_{\text{king}} & = [ 0.8 & 0.9 & 0.1 & 0 & \ldots ] \\
  v_{\text{queen}} & = [ 0.8 & 0.1 & 0.8 & 0 & \ldots ] \\
  v_{\text{apply}} & = [ 0.1 & 0.2 & 0.1 & 0.8 & \ldots ] 
  \end{align*}
  \]

- Difference between word vectors captures their relations

[Pennington et al., EMNLP 2014]
How to learn word vectors

- **Distributional** hypothesis:
  “You shall know a word by the company it keeps” (J. R. Firth 1957: 11)
  *linguistic items with similar distributions have similar meanings.*

He curtains open and the stars and the cold, close through the night with the made in the light of the surely under the bright sun, the seasons of the m is dazzling snow, the un and the temple of the in the dark and now the bird on the shape of the But I could n’t see the they love the sun, the r the light of the shiny stars. The splash of flowing w man ’s first look at the rief information on both shining in on the barely stars " . And neither of the w shining so brightly , it stars. It all boils down , wr stars, thrilled by ice-white stars? Home , alone, Jay pla stars have risen full and cold stars, driving out of the hug stars rise, full and amber a stars over the trees in front stars or the moon , only the stars and the stars. None of stars. The various exhibits, aer stars and constellations, inc
How to learn word vectors

- Learn word representations based on co-occurrences
- E.g., Word2vec (Mikolov et al. 2013)

\[
P(o|c) = \frac{\exp(u_o^T v_c)}{\sum_{w \in V} \exp(u_w^T v_c)}
\]

Slide credit: Stanford cs224n
How to learn word vectors

- Learn word representations based on co-occurrences
- E.g., Word2vec (Mikolov et al. 2013)

\[ P(o|c) = \frac{\exp(u_0^T v_c)}{\sum_{w \in V} \exp(u_w^T v_c)} \]

Slide credit: Stanford cs224n
Continuous representations for entities

- Republic Party
- Democratic Party
- George W Bush
- Laura Bush
- Michelle Obama
- ?

Michelle Obama
Democratic Party
Laura Bush
George W Bush
?
Continuous representations for entities

- Useful resources for NLP applications
  - Semantic Parsing & Question Answering
  - Information Extraction

Freebase
DBpedia
YAGO
NELL
OpenIE/ReVerb
Contextualized Word Representations

- Most words have multiple meanings
- Can we encode word also based on the surrounding contexts?

He taught himself to **play** the violin.

Do you enjoy the **play**?

Embedding visualization from context\textsuperscript{1}

Word Embedding

Contextualized Word Embedding

from context\textsuperscript{2}
Embeddings from Language Models (ELMo)

Deep contextualized word representations (Peters+2018)

Train Separate Left-to-Right and Right-to-Left LMs

Apply as “Pre-trained Embeddings”
Embeddings from Language Models (ELMo)

... games or play online via ...

ELMo embeddings from language models
Bidirectional Encoder Representations from Transformers (BERT)
Mulit-Lingual Representations
Cross-lingual Representations

Continuous approximation of discrete translation dictionaries in a vector space.
Traditional Approach

- Does not exploit relationship between the languages. No parameter/resource sharing.
- Requires annotated data – cumbersome for low-resource languages.
Direct Model Transfer through Shared Cross-lingual Space

Training Data in English

\((x, y)\)

\(\Phi\)

\((\Phi(x), y)\)

Test Data in French

\((x, y)\)

Shared Cross-lingual Space

Learning Algorithm

\(\mathcal{A} : (\Phi(x), y) \rightarrow M\)

\(\hat{y}\)

Learnt Model

\(M : \Phi(x) \rightarrow y\)
Joint Training through Shared Cross-lingual Space

Training Data in English: \((x, y)\)

Training Data in French: \((x, y)\)

Shared Cross-lingual Space

Learning Algorithm: \(A : (\Phi(x), y) \rightarrow M\)

Learnt Model: \(M : \Phi(x) \rightarrow y\)
Approaches to Learning Cross-lingual Representations

Based on
Cross-Lingual Models of Word Embeddings: An Empirical Comparison
ACL 2016 (long)
Approaches to learn Cross-lingual Representations

- Can we identify a general recipe to learn shared cross-lingual representations?

- What kind of cross-lingual signal is best suited for a given task?
  - Tasks considered – word similarity, dictionary induction, document classification and dependency parsing
  - What nature of supervision needed for a task (word-level, sentence-level etc.)

- Can we get good performance with cheap signals?
Forms of Cross-lingual Supervision

Decreasing Cost

Nature:
- word + sentence
- sentence
- word
- document

BiSkip
Luong et al. 15

BiCVM
Hermann et al. 14

BiCCA
Faruqui et al. 14

BiVCD
Vulic et al. 15

Je t’aime
(Love, aime)
(I, je)

Bonjour! Je t’aime
(Hello! How are you? I love you)

Je t’aime
(I love you)

Je
(I)
t’
(l’)
aime
(love)
You

Nature:
word + sentence
sentence
word
document
General Algorithm for Learning Shared Representations

\[(W^*, \ V^*) \leftarrow \text{argmin } \alpha A(W) + \beta B(V) + C(W, V)\]

Mono. Obj. for English

Mono. Obj. for French

Cross-lingual Obj.

\[E^*(\bar{v}, \bar{w}) = \sum \sum f(\bar{v}) g(\bar{w}) \log P(n_c \mid w)\]

\[E(\bar{v}, \bar{B}(\bar{v}^n)) = \max \left( \sum + \sum \Delta E(\bar{v}, \bar{w}, \bar{v}^n) \cdot 0 \right)\]

\[\Delta E(\bar{v}, \bar{w}, \bar{v}^n) = \sum \sum \sum \sum f(\bar{v}) g(\bar{w}) P(n_s \mid w)\]

\[C(W, V) = \sum \sum \sum \sum E(\bar{v}, \bar{w}, \bar{v}^n)\]

\[C(W, V) = D_{21}(W, V) + D_{22}(W, V)\]
Performance improves with cost of supervision, with gaps > 10 pts b/w some models.

No clear trend observed in word similarity tasks, did not correlate with other tasks.
Extrinsic Evaluation: Cross-lingual Dependency Parsing

Aujourd'hui j'ai rencontré un accident. J'ai besoin de prendre le vol. Je ne pouvais pas déjeuner aujourd'hui à cause d'une réunion.
When transferring semantic knowledge across languages, sentence + word alignment information is superior to sentence or word alignment alone.

When transferring syntactic knowledge across languages, using word alignments for training embeddings is crucial.
Case Study: Cross-lingual Transfer for Dependency Parsing

Based on
Cross-lingual Dependency Parsing

Train

English Treebank

Parse

French Corpus

Aujourd'hui j'ai rencontré un accident
J'ai besoin de prendre le vol
Je ne pouvais pas déjeuner aujourd'hui à cause d'une réunion
Background: Deep Biaffine Parser

- **Graph-based** parser
- Encoder: RNN (Order-sensitive); Decoder: Graph (Order-free)

\[ H^{(arc-dep)} \oplus 1 \quad U^{(arc)} \quad H^{(arc-head)} \quad S^{(arc)} \]

MLP: \( h_i^{(arc-dep)}, h_i^{(arc-head)} \)

BiLSTM: \( r_i \)

Embeddings: \( x_i \)

Dozat and Manning (ICLR2017)
Cross-lingual Transfer for Dependency Parsing

- Research question: how to transfer a model across languages
  - Remove language-specific knowledge (e.g., word order) from encoder
  - Add language-specific knowledge to decoder

- Examine and verify our hypothesis on cross-lingual dependency parsing
  - UD annotation for over 70 languages
  - Parser is a low-level task that reflects the problems
Multi-Head Self-Attention with Relative Position

In the original paper:

\[
\begin{align*}
PE_{(pos,2i)} &= \sin(pos/10000^{2i/d_{model}}) \\
PE_{(pos,2i+1)} &= \cos(pos/10000^{2i/d_{model}})
\end{align*}
\]

Vaswani et. al. (NIPS 2017)

Encoder absolute distance

Flexible positional encoding (order-free)

Shaw et. al. (NAACL2018)
Architectures for Cross-lingual Parser

- **Embedding**
  - Facebook
  - MUSE

- **Encoders**
  - BiLSTMs *(order-sensitive)* v.s.
  - Multi-Head Self-Attention with Absolute Relative Positional Encoding *(order-free)*

- **Decoders**
  - Pointer Network *(order-sensitive)* v.s.
  - BiAffine Attention *(order-free)*
Experiments

- **Datasets:**
  - UD (V2.2)
  - 31 languages, 12 families

- **Setting:**
  - Train/Dev on English
  - Directly predict on the rest 30 languages (zero-shot)

<table>
<thead>
<tr>
<th>Language Families</th>
<th>Languages</th>
</tr>
</thead>
<tbody>
<tr>
<td>Afro-Asiatic</td>
<td>Arabic (ar), Hebrew (he)</td>
</tr>
<tr>
<td>Austronesian</td>
<td>Indonesian (id)</td>
</tr>
<tr>
<td>IE.Baltic</td>
<td>Latvian (lv)</td>
</tr>
<tr>
<td>IE.Germanic</td>
<td>Danish (da), Dutch (nl), English (en), German (de), Norwegian (no), Swedish (sv)</td>
</tr>
<tr>
<td>IE.Indic</td>
<td>Hindi (hi)</td>
</tr>
<tr>
<td>IE.Latin</td>
<td>Latin (la)</td>
</tr>
<tr>
<td>IE.Romance</td>
<td>Catalan (ca), French (fr), Italian (it), Portuguese (pt), Romanian (ro), Spanish (es)</td>
</tr>
<tr>
<td>IE.Slavic</td>
<td>Bulgarian (bg), Croatian (hr), Czech (cs), Polish (pl), Russian (ru), Slovak (sk), Slovenian (sl), Ukrainian (uk)</td>
</tr>
<tr>
<td>Japanese</td>
<td>Japanese (ja)</td>
</tr>
<tr>
<td>Korean</td>
<td>Korean (ko)</td>
</tr>
<tr>
<td>Sino-Tibetan</td>
<td>Chinese (zh)</td>
</tr>
<tr>
<td>Uralic</td>
<td>Estonian (et), Finnish (fi)</td>
</tr>
</tbody>
</table>
Selected Transfer Results of Different Architectures

Zero-shot Transfer UAS Results (Except for English)

Distances to English increase, Transfer performances decrease.

- English
- Swedish
- Spanish
- Croatian
- Hindi

- RNN-Stack
- SelfAtt-Stack
- RNN-Graph
- SelfAtt-Graph

order-sensitive  order-free
The languages (x-axis) are sorted by this relative frequency from high to low.
Adversarial Learning for Retrofitting Text Representation

Based on
Main Idea

Aujourd'hui j'ai rencontré un accident.
J'ai besoin de prendre le vol.
Je ne pouvais pas déjeuner aujourd'hui à cause d'une réunion.

У меня сильная головная боль.
Они прекрасно проводят время вместе.
Они скоро поженятся.

English Treebank

Train

Parse

French Corpus

Russian Corpus

Source Language

Auxiliary Language

Target Language
Main Idea

- Use unlabeled corpora of auxiliary languages
- Adversarial training to learn language-agnostic representation
  - **Discriminator**: predicts language label
Training Procedure

Step 1. Warm-start the parser

- Mini-batch training using source language treebank
- For $k$ iterations

I prefer the morning flight through Denver

United canceled the morning flights to Houston

JetBlue canceled our flight this morning which was already late
Training Procedure

Step 2. Jointly train using auxiliary languages

- Train the parser on source language
- Adversarial training on both source and auxiliary languages

JetBlue canceled our flight this morning which was already late

Je ne pouvais pas déjeuner aujourd'hui à cause d'une réunion
Experiment Setup

Embedding

- Token embeddings
  - Multilingual Embeddings (MUSE) [Smith et al., 2017, Bojanowski et al., 2017]
  - Multilingual BERT (M-BERT) [Devlin et al., 2017]

- Part-of-speech embeddings

Parsers [Ahmad et al., 2019]

- Graph-based: Self-attentive-Graph
  - Multi-Head Self-Attention (order-free)

- Transition-based: RNN-StackPtr
  - BiLSTMs (order-dependent)
Experiment Setup

Single Source Transfer Parsing
- Train parser on one language
- Source language: English

Adversarial training
- Using a language pair (one source and one auxiliary language)
- Auxiliary languages are selected based on -
  - Covering different language families
  - Average distance between auxiliary language and all target languages
Impact of Adversarial Training

Self-attentive-Graph Parser

Multilingual Word Embeddings

- - - Baseline French Russian

Multilingual BERT

- - - Baseline French Russian

x-axis = language labels
y-axis = performance_diff (model_trained_with_aux_lang, model_trained_on_src_lang)
Impact on Language Families

Self-attentive-Graph Parser

IE.Slavic
IE.Romance
IE.Germanic
Decoding with Language-Specific Knowledge

- Leveraging linguistics knowledge (WALS features) in decoding

Constraint: In an ADP-NOUN arc in Hindi, ADP is more likely to be on the right.

Kai-Wei Chang (http://kwchang.net/talks/genderbias/)
Corpus-Statistics Constraints

- Consider constraints in two forms:
  - the ratio $r$ of POS1 being on the left in all POS1-POS2 arcs
  - specifies the ratio $r$ of the heads of a particular POS appears on the left of that POS

- Compiling from WALS features:
  - Dominant order: e.g. 85A – Binary constraint (ADP, NN)
    - Prepositions: $r \in (0,0.25)$
    - No dominant order: $r \in (0.25,0.75)$
    - Postpositions: $r \in (0.75,1)$

Kai-Wei Chang (http://kwchang.net/talks/genderbias/)
Constrained Inference

- **Lagrangian Relaxation**
  - Introduce Lagrangian multipliers for each constraints.
  - Apply sub-gradient descent method to solve the dual form.

- **Posterior Regularization**
  - Use constraints to define a feasible distribution set $Q$.
  - Find the closest distribution $q \in Q$ from $p_\theta$, and do MAP inference on $q$. 

\[
KL(Q || p_\theta)
\]
## Experiments Results

<table>
<thead>
<tr>
<th>Lang.</th>
<th>Baseline</th>
<th>Lagrangian Relaxation</th>
<th>Posterior Regularization</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Oracle WALS</td>
<td>WALS</td>
</tr>
<tr>
<td>en</td>
<td>90.5</td>
<td>90.3 90.4 -0.1</td>
<td></td>
</tr>
<tr>
<td>IE.Indic</td>
<td>18.3</td>
<td>35.2 34.0 +15.7</td>
<td></td>
</tr>
<tr>
<td>Dravidian</td>
<td>36.1</td>
<td>42.8 43.4 +7.3</td>
<td></td>
</tr>
<tr>
<td>Turkic</td>
<td>31.2</td>
<td>35.2 37.1 +5.9</td>
<td></td>
</tr>
<tr>
<td>Austronesian</td>
<td>49.3</td>
<td>53.1 52.3 +3.0</td>
<td></td>
</tr>
<tr>
<td><strong>Average Performance</strong></td>
<td>54.3</td>
<td>58.4 57.8 +3.5</td>
<td></td>
</tr>
</tbody>
</table>

- LR, PR get improvements in 15,17 out of 19 target languages from variant of language families, respectively
- The improvements are closely related to the ratio gap in constraints
- LR has greater average improvement, while PR is a more robust inference algorithm

Kai-Wei Chang (http://kwchang.net/talks/genderbias/)
Multi-modal Contextualized Language Representation
Transferable Representations

Several people walking on a sidewalk in the rain with umbrellas. Main training objective is to predict missing words.

VisualBERT

The model projects words and image regions into the same vector space and uses multiple Transformer layers to build joint representations.

Several people [MASK] on a [MASK] in the [MASK] with [MASK].

Input consists of an image and a caption with some masked words. Such data is easy to obtain from the internet.

Unsupervised pre-training on vision and language

Is it raining outside?

a) Yes, it is snowing.
b) Yes, [person8] and [person10] are outside.
c) No, it looks to be fall.
d) Yes, it is raining heavily.

An example from the VCR dataset

Transfer to answering commonsense questions
Goal: Joint Embedding Space

man wearing white shirt is walking on sidewalk alongside other pedestrians

PEOPLE

isA

MAN

SIDEWALK

WALK_ALONGSIDE

PEDESTRIAN

WALK_ON

WALK_ON
Challenge 1: Grounding Language in X

man wearing white shirt is walking on sidewalk alongside other pedestrians
Challenge 2: (Contextualized) Commonsense Embedding

Based on BERT (NAACL 19)

12 layers of self-attention captures association between text-text, text-image

A natural way to bridge multi-modalities
- May combine this with GATs (graph attention network)

Related Work: VideoBERT
# A (potentially non-exhaustive) list of BERT with Vision

<table>
<thead>
<tr>
<th>Architecture</th>
<th>Architecture</th>
<th>Pre-training resource*</th>
<th>Pre-training Tasks**</th>
</tr>
</thead>
<tbody>
<tr>
<td>ViLBERT (Lu et al., 2019)</td>
<td>Two-stream</td>
<td>CC</td>
<td>1, 2, 3</td>
</tr>
<tr>
<td>B2T2 (Alberti et al., 2019)</td>
<td>Single-stream</td>
<td>CC</td>
<td>1, 2</td>
</tr>
<tr>
<td>LXMERT (Tan &amp; Bansal, 2019)</td>
<td>Two-stream</td>
<td>COCO, VG, VQA, GQA</td>
<td>1, 2, 3, 4, 5</td>
</tr>
<tr>
<td>VisualBERT (Li et al., 2019a)</td>
<td>Single-stream</td>
<td>COCO</td>
<td>1, 2</td>
</tr>
<tr>
<td>Unicoder-VL (Li et al., 2019b)</td>
<td>Single-stream</td>
<td>CC</td>
<td>1, 2, 3</td>
</tr>
<tr>
<td>VL-BERT (Su et al., 2019)</td>
<td>Single-stream</td>
<td>CC</td>
<td>2, 3</td>
</tr>
<tr>
<td>UNITER (Chen et al., 2019)</td>
<td>Single-stream</td>
<td>COCO, VG, CC, SBU</td>
<td>1, 2, 3, 4</td>
</tr>
</tbody>
</table>

* CC stands for Conceptual Captions, VG stands for Visual Genome

** 1 means cross modality alignment; 2 means grounded masked LM; 3 means masked visual classification; 4 means visual regression; 5 means cross modality QA
An example of single-stream architecture: VisualBERT
Architectural Difference

Separate Transformers for text and vision at first and then a joint Transformer

An example of two-stream architecture: LXMERT
Pre-training Objectives

- Visual Regression
- Visual Classification
- Cross Modality Matching
- Cross Modality QA
- Grounded Masked LM

Pre-training objectives of LXMERT
Performance Improvement

Image caption data (MSCOCO):
~300,000 images, 5 captions per image

- **VCR** 71.6 (best single model 72.6)
- **VQA** 70.8 (baseline: 68.5, best ~75)
- **NLVR²** 67.4 (best on leaderboard: 54.1)
- **Flickr30k** R@10: 86.61 (Best: 86.35)
Implicit Grounding (1)
Implicit Grounding (2)
Entity Grounding

- man wearing white shirt
- is walking on sidewalk along side other pedestrians

Layer 3

Layer 4

Layer 6

Layer 10

- Man
- Shirt
- Sidewalk
- Pedestrians
- Sidewalk*
Analyze the Attention Weights

**Syntactic Grounding**

- Man
- Shirt
- Sidewalk
- Pedestrians
- Sidewalk*
What does BERT with Vision Look At?

- Entity grounding
  1. Certain heads are accurate
  2. Accuracy peaks at higher layers
What does BERT with Vision Look At?

- **Syntactic grounding**
  1) Certain heads are accurate
  2) Accuracy peaks at higher layers

<table>
<thead>
<tr>
<th>Type</th>
<th>Baseline</th>
<th>Acc</th>
<th>Head</th>
</tr>
</thead>
<tbody>
<tr>
<td>det</td>
<td>19.59</td>
<td>54.01</td>
<td>10-1</td>
</tr>
<tr>
<td>pobj</td>
<td>17.34</td>
<td>32.82</td>
<td>11-11</td>
</tr>
<tr>
<td>amod</td>
<td>18.67</td>
<td>45.96</td>
<td>10-9</td>
</tr>
<tr>
<td>nsubj</td>
<td>23.19</td>
<td>44.64</td>
<td>5-1</td>
</tr>
<tr>
<td>prep</td>
<td>20.61</td>
<td>49.27</td>
<td>9-11</td>
</tr>
<tr>
<td>dobj</td>
<td>9.82</td>
<td>30.24</td>
<td>11-11</td>
</tr>
<tr>
<td>punct</td>
<td>23.32</td>
<td>48.80</td>
<td>3-6</td>
</tr>
<tr>
<td>partmod</td>
<td>21.41</td>
<td>38.15</td>
<td>4-9</td>
</tr>
<tr>
<td>nn</td>
<td>16.33</td>
<td>34.06</td>
<td>10-9</td>
</tr>
<tr>
<td>num</td>
<td>23.15</td>
<td>67.44</td>
<td>9-11</td>
</tr>
</tbody>
</table>
What does BERT with Vision Look At?

- **Syntactic grounding**

  1) Certain heads are accurate
  2) Accuracy peaks at higher layers