Probing pretrained models

CS 685, Fall 2020
Introduction to Natural Language Processing
http://people.cs.umass.edu/~miyyer/cs685/

Mohit Iyyer
College of Information and Computer Sciences
University of Massachusetts Amherst

most slides from Tu Vu
Logistics stuff

• Reminder: don’t discuss exam, people are still doing makeups!

• Exam grading: to be done by end of the month

• Remaining todos for you: final project report, a couple quizzes, and (optionally) the extra credit HW
BERTology
studying the inner working of large-scale Transformer language models like BERT

• what are captured in different model components, e.g., attention / hidden states?
tools & examples

BERTology - HuggingFace’s Transformers
https://huggingface.co/transformers/bertology.html

• accessing all the hidden states of BERT

• accessing all the attention weights for each head of BERT

• retrieving heads output values and gradients
Are Sixteen Heads Really Better than One? Michel et al., NeurIPS 2019

large percentage of attention heads can be removed at test time without significantly impacting performance

What Does BERT Look At? An Analysis of BERT’s Attention, Clark et al., BlackBoxNLP 2019

substantial syntactic information is captured in BERT’s attention
AllenNLP Interpret
https://allennlp.org/interpret

Tools & Examples

Simple Gradients Visualization
See saliency map interpretations generated by visualizing the gradient.

Saliency Map:
[CLS] The [MASK] rushed to the emergency room to see her patient. [SEP]

Mask 1 Predictions:
- nurse: 47.1%
- woman: 16.4%
- doctor: 10.0%
- mother: 3.4%
- girl: 3.0%
understanding contextualized representations

two most prominent methods

• visualization

• linguistic probe tasks
Sentiment neuron

While training the linear model with L1 regularization, we noticed it used surprisingly few of the learned units. Digging in, we realized there actually existed a single “sentiment neuron” that’s highly predictive of the sentiment value.

The sentiment neuron within our model can classify reviews as negative or positive, even though the model is trained only to predict the next character in the text.
what is a linguistic probe task?

given an encoder model (e.g., BERT) pre-trained on a certain task, we use the representations it produces to train a classifier (without further fine-tuning the model) to predict a linguistic property of the input text
predict the length (number of tokens)
of the input sentence $s$

(Adi et al., 2017)
Sentence length

Predict the length (number of tokens) of the input sentence $s$

Probe network

Classifier

Sent. repr.

BERT [CLS] representation, kept frozen

(Adi et al., 2017)
predict the length (number of tokens) of the input sentence $s$

sentence length

probe network

classifier

sent. repr.

Feed-forward NN trained from scratch

BERT [CLS] representation, kept frozen

(Adi et al., 2017)
**sentence length**

predict the length (number of tokens) of the input sentence \( s \)

probe network

classifier

sent. repr.

**word content**

predict the word \( w \) appears in the sentence \( s \)

classifier

sent. repr.

word repr.

(Adi et al., 2017)
predict the length (number of tokens) of the input sentence $s$

predict the word $w$ appears in the sentence $s$

BERT [CLS] representation, kept frozen

Possibly BERT subword embedding

(Adi et al., 2017)
sentence length
predict the length (number of tokens) of the input sentence \( s \)

word content
predict the word \( w \) appears in the sentence \( s \)

word order
predict whether \( w_1 \) appears before or after \( w_2 \) in the sentence \( s \)

(Adi et al., 2017)
token labeling: POS tagging
predict a POS tag for each token

segmentation: NER
predict the entity type of the input token

pairwise relations: syntactic dep. arc
predict if there is a syntactic dependency arc between tok₁ and tok₂

(Liu et al., 2019)
predict whether two spans of tokens ("mentions") refer to the same entity (or event)

(Tenney et al., 2019)
motivation of probe tasks

• if we can train a classifier to predict a property of the input text based on its representation, it means the property is encoded somewhere in the representation

• if we cannot train a classifier to predict a property of the input text based on its representation, it means the property is not encoded in the representation or not encoded in a useful way, considering how the representation is likely to be used
characteristics of probe tasks

• usually classification problems that focus on simple linguistic properties

• ask simple questions, minimizing interpretability problems

• because of their simplicity, it is easier to control for biases in probing tasks than in downstream tasks

• the probing task methodology is agnostic with respect to the encoder architecture, as long as it produces a vector representation of input text

• does not necessarily correlate with downstream performance (Conneau et al., 2018)
probe approach

- predict a linguistic property of the input
- the classifier's weights are updated
- train the classifier only
- no further fine-tuning
- the encoder's weights are fixed

input text
lowest layers focus on local syntax, while upper layers focus more semantic content

(Peters et al., 2018)
BERT represents the steps of the traditional NLP pipeline: POS tagging $\rightarrow$ parsing $\rightarrow$ NER $\rightarrow$ semantic roles $\rightarrow$ coreference

![Expected layer & center-of-gravity graph](image)

- **POS:** 3.39, 11.68
- **Consts.:** 3.79, 13.06
- **Dep.:** 5.69, 13.75
- **Entities:** 4.64, 13.16
- **SRL:** 6.54, 13.63
- **Coref.:** 9.47, 15.80
- **SPR:** 9.93, 12.72
- **Relations:** 9.40, 12.83

The expected layer at which the probing model correctly labels an example.

A higher center-of-gravity means that the information needed for that task is captured by higher layers.

(Tenney et al., 2019)
does BERT encode syntactic structure?

The chef who ran to the store was out of food

(Hewitt and Manning et al., 2019)
understanding the syntax of the language may be useful in language modeling

The chef who ran to the store was out of food.

1. Because there was no food to be found, the chef went to the next store.

2. After stocking up on ingredients, the chef returned to the restaurant.

(Hewitt and Manning et al., 2019)
how to probe for trees?

trees as distances and norms

the distance metric—the path length between each pair of words—recovers the tree $T$ simply by identifying that nodes $u, v$ with distance $d_T(u, v) = 1$ are neighbors

the node with greater norm—depth in the tree—is the child

(Hewitt and Manning et al., 2019)
a structural probe

- probe task 1 — distance: predict the path length between each given pair of words
- probe task 2 — depth/norm: predict the depth of a given word in the parse tree

(Hewitt and Manning et al., 2019)
Yes, BERT knows the structure of syntax trees

<table>
<thead>
<tr>
<th>Method</th>
<th>Distance</th>
<th>Depth</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>UUAS</td>
<td>D Spr.</td>
</tr>
<tr>
<td>ELMo1</td>
<td>77.0</td>
<td>0.83</td>
</tr>
<tr>
<td>BERTBase7</td>
<td>79.8</td>
<td>0.85</td>
</tr>
<tr>
<td>BERTLarge15</td>
<td>82.5</td>
<td>0.86</td>
</tr>
<tr>
<td>BERTLarge16</td>
<td>81.7</td>
<td>0.87</td>
</tr>
</tbody>
</table>

(Hewitt and Manning et al., 2019)
does BERT know numbers?

what is the sum of eleven and fourteen?

25

what is the sum of eleven and fourteen?
probing for numeracy

(Wallace et al., 2019)
ELMo is actually better than BERT at this!

<table>
<thead>
<tr>
<th>Interpolation</th>
<th>List Maximum (5-classes)</th>
<th>Decoding (RMSE)</th>
<th>Addition (RMSE)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>[0.99] [0.999] [0.9999]</td>
<td>[0.99] [0.999] [0.9999]</td>
<td>[0.99] [0.999] [0.9999]</td>
</tr>
<tr>
<td>Random Vectors</td>
<td>0.16 0.23 0.21</td>
<td>29.86 292.88 2882.62</td>
<td>42.03 410.33 4389.39</td>
</tr>
<tr>
<td>Untrained CNN</td>
<td>0.97 0.87 0.84</td>
<td>2.64 9.67 44.40</td>
<td>1.41 14.43 69.14</td>
</tr>
<tr>
<td>Untrained LSTM</td>
<td>0.70 0.66 0.55</td>
<td>7.61 46.5 210.34</td>
<td>5.11 45.69 510.19</td>
</tr>
<tr>
<td>Pre-trained</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Word2Vec</td>
<td>0.90 0.78 0.71</td>
<td>2.34 18.77 333.47</td>
<td>0.75 21.23 210.07</td>
</tr>
<tr>
<td>GloVe</td>
<td>0.90 0.78 0.72</td>
<td>2.23 13.77 174.21</td>
<td>0.80 16.51 180.31</td>
</tr>
<tr>
<td>ELMo</td>
<td>0.98 0.88 0.76</td>
<td>2.35 13.48 62.20</td>
<td>0.94 15.50 45.71</td>
</tr>
<tr>
<td>BERT</td>
<td>0.95 0.62 0.52</td>
<td>3.21 29.00 431.78</td>
<td>4.56 67.81 454.78</td>
</tr>
</tbody>
</table>

(Wallace et al., 2019)
Why?

character-level CNNs are the best architecture for capturing numeracy

subword encoding is a poor method to encode digits, e.g., two numbers which are similar in value can have very different sub-word divisions

(Wallace et al., 2019)
Can BERT serve as a structured knowledge base?

Query: (Dante, born-in, X)

Florence
LAMA (LAnguage Model Analysis) probe

(Petroni et al., 2019)
• manually define templates for considered relations, e.g., “[S] was born in [O]” for “place of birth”

• find sentences that contain both the subject and the object, then mask the object within the sentences and use them as templates for querying

• create cloze-style questions, e.g., rewriting “Who developed the theory of relativity?” as “The theory of relativity was developed by [MASK]”

(Petroni et al., 2019)
<table>
<thead>
<tr>
<th>Relation</th>
<th>Query</th>
<th>Answer</th>
<th>Generation</th>
</tr>
</thead>
<tbody>
<tr>
<td>P54</td>
<td>Dani Alves plays with ____</td>
<td>Barcelona</td>
<td>Santos [-2.4], Porto [-2.5], Sporting [-3.1], Brazil [-3.3], Portugal [-3.7]</td>
</tr>
<tr>
<td>P106</td>
<td>Paul Tougui is a ____ by profession.</td>
<td>politician</td>
<td>lawyer [-1.1], journalist [-2.4], teacher [-2.7], doctor [-3.0], physician [-3.7]</td>
</tr>
<tr>
<td>P527</td>
<td>Sodium sulfide consists of ____</td>
<td>sodium</td>
<td>water [-1.2], sulfur [-1.7], sodium [-2.5], zinc [-2.8], salt [-2.9]</td>
</tr>
<tr>
<td>P102</td>
<td>Gordon Scholes is a member of the ____ political party.</td>
<td>Labor</td>
<td>Labour [-1.3], Conservative [-1.6], Green [-2.4], Liberal [-2.9], Labor [-2.9]</td>
</tr>
<tr>
<td>P530</td>
<td>Kenya maintains diplomatic relations with ____</td>
<td>Uganda</td>
<td>India [-3.0], Uganda [-3.2], Tanzania [-3.5], China [-3.6], Pakistan [-3.6]</td>
</tr>
<tr>
<td>P176</td>
<td>iPod Touch is produced by ____</td>
<td>Apple</td>
<td>Apple [-1.6], Nokia [-1.7], Sony [-2.0], Samsung [-2.6], Intel [-3.1]</td>
</tr>
<tr>
<td>P30</td>
<td>Bailey Peninsula is located in ____</td>
<td>Antarctica</td>
<td>Antarctica [-1.4], Bermuda [-2.2], Newfoundland [-2.5], Alaska [-2.7], Canada [-3.1]</td>
</tr>
<tr>
<td>P178</td>
<td>JDK is developed by ____</td>
<td>Oracle</td>
<td>IBM [-2.0], Intel [-2.3], Microsoft [-2.5], HP [-3.4], Nokia [-3.5]</td>
</tr>
<tr>
<td>P1412</td>
<td>Carl III used to communicate in ____</td>
<td>Swedish</td>
<td>German [-1.6], Latin [-1.9], French [-2.4], English [-3.0], Spanish [-3.0]</td>
</tr>
<tr>
<td>P17</td>
<td>Sunshine Coast, British Columbia is located in ____</td>
<td>Canada</td>
<td>Canada [-1.2], Alberta [-2.8], Yukon [-2.9], Labrador [-3.4], Victoria [-3.4]</td>
</tr>
<tr>
<td>AtLocation</td>
<td>You are likely to find a overflow in a ____</td>
<td>drain</td>
<td>sewer [-3.1], canal [-3.2], toilet [-3.3], stream [-3.6], drain [-3.6]</td>
</tr>
<tr>
<td>CapableOf</td>
<td>Ravens can ____</td>
<td>fly</td>
<td>fly [-1.5], fight [-1.8], kill [-2.2], die [-3.2], hunt [-3.4]</td>
</tr>
<tr>
<td>CausesDesire</td>
<td>Joke would make you want to ____</td>
<td>laugh</td>
<td>cry [-1.7], die [-1.7], laugh [-2.0], vomit [-2.6], scream [-2.6]</td>
</tr>
<tr>
<td>Causes</td>
<td>Sometimes virus causes ____</td>
<td>infection</td>
<td>disease [-1.2], cancer [-2.0], infection [-2.6], plague [-3.3], fever [-3.4]</td>
</tr>
<tr>
<td>HasA</td>
<td>Birds have ____</td>
<td>feathers</td>
<td>wings [-1.8], nests [-3.1], feathers [-3.2], died [-3.7], eggs [-3.9]</td>
</tr>
<tr>
<td>HasPrerequisite</td>
<td>Typing requires ____</td>
<td>speed</td>
<td>patience [-3.5], precision [-3.6], registration [-3.8], accuracy [-4.0], speed [-4.1]</td>
</tr>
<tr>
<td>HasProperty</td>
<td>Time is ____</td>
<td>finite</td>
<td>short [-1.7], passing [-1.8], precious [-2.9], irrelevant [-3.2], gone [-4.0]</td>
</tr>
<tr>
<td>MotivatedByGoal</td>
<td>You would celebrate because you are ____</td>
<td>alive</td>
<td>happy [-2.4], human [-3.3], alive [-3.3], young [-3.6], free [-3.9]</td>
</tr>
<tr>
<td>ReceivesAction</td>
<td>Skills can be ____</td>
<td>taught</td>
<td>acquired [-2.5], useful [-2.5], learned [-2.8], combined [-3.9], varied [-3.9]</td>
</tr>
<tr>
<td>UsedFor</td>
<td>A pond is for ____</td>
<td>fish</td>
<td>swimming [-1.3], fishing [-1.4], bathing [-2.0], fish [-2.8], recreation [-3.1]</td>
</tr>
</tbody>
</table>

(Petroni et al., 2019)
BERT contains relational knowledge competitive with symbolic knowledge bases and excels on open-domain QA

(Petroni et al., 2019)

<table>
<thead>
<tr>
<th>Corpus</th>
<th>Relation</th>
<th>Statistics #Facts</th>
<th>#Rel</th>
<th>Baselines Freq</th>
<th>DrQA</th>
<th>KB RE$_n$</th>
<th>RE$_o$</th>
<th>LM Fs</th>
<th>Txl</th>
<th>Eb</th>
<th>E5B</th>
<th>Bb</th>
<th>Bl</th>
</tr>
</thead>
<tbody>
<tr>
<td>Google-RE</td>
<td>birth-place</td>
<td>2937</td>
<td>1</td>
<td>4.6</td>
<td>-</td>
<td>3.5</td>
<td>13.8</td>
<td>4.4</td>
<td>2.7</td>
<td>5.5</td>
<td>7.5</td>
<td>14.9</td>
<td>16.1</td>
</tr>
<tr>
<td></td>
<td>birth-date</td>
<td>1825</td>
<td>1</td>
<td>1.9</td>
<td>-</td>
<td>0.0</td>
<td>1.9</td>
<td>0.3</td>
<td>1.1</td>
<td>0.1</td>
<td>0.1</td>
<td>1.5</td>
<td>1.4</td>
</tr>
<tr>
<td></td>
<td>death-place</td>
<td>765</td>
<td>1</td>
<td>6.8</td>
<td>-</td>
<td>0.1</td>
<td>7.2</td>
<td>3.0</td>
<td>0.9</td>
<td>0.3</td>
<td>1.3</td>
<td>13.1</td>
<td>14.0</td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>5527</td>
<td>3</td>
<td>4.4</td>
<td>-</td>
<td>1.2</td>
<td>7.6</td>
<td>2.6</td>
<td>1.6</td>
<td>2.0</td>
<td>3.0</td>
<td>9.8</td>
<td>10.5</td>
</tr>
<tr>
<td>T-REx</td>
<td>1-1</td>
<td>937</td>
<td>2</td>
<td>1.78</td>
<td>-</td>
<td>0.6</td>
<td>10.0</td>
<td>17.0</td>
<td>36.5</td>
<td>10.1</td>
<td>13.1</td>
<td>68.0</td>
<td>74.5</td>
</tr>
<tr>
<td></td>
<td>N-1</td>
<td>20006</td>
<td>23</td>
<td>23.85</td>
<td>-</td>
<td>5.4</td>
<td>33.8</td>
<td>6.1</td>
<td>18.0</td>
<td>3.6</td>
<td>6.5</td>
<td>32.4</td>
<td>34.2</td>
</tr>
<tr>
<td></td>
<td>N-M</td>
<td>13096</td>
<td>16</td>
<td>21.95</td>
<td>-</td>
<td>7.7</td>
<td>36.7</td>
<td>12.0</td>
<td>16.5</td>
<td>5.7</td>
<td>7.4</td>
<td>24.7</td>
<td>24.3</td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>34039</td>
<td>41</td>
<td>22.03</td>
<td>-</td>
<td>6.1</td>
<td>33.8</td>
<td>8.9</td>
<td>18.3</td>
<td>4.7</td>
<td>7.1</td>
<td>31.1</td>
<td>32.3</td>
</tr>
<tr>
<td>ConceptNet</td>
<td>Total</td>
<td>11458</td>
<td>16</td>
<td>4.8</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>3.6</td>
<td>5.7</td>
<td>6.1</td>
<td>6.2</td>
</tr>
<tr>
<td>SQuAD</td>
<td>Total</td>
<td>305</td>
<td>-</td>
<td>-</td>
<td>37.5</td>
<td>-</td>
<td>-</td>
<td>3.6</td>
<td>3.9</td>
<td>1.6</td>
<td>4.3</td>
<td>14.1</td>
<td>17.4</td>
</tr>
</tbody>
</table>
probe complexity

arguments for “simple” probes

we want to find easily accessible information in a representation

arguments for “complex” probes

useful properties might be encoded non-linearly

(Hewitt et al., 2019)
## control tasks

![Diagram showing control tasks](image)

<table>
<thead>
<tr>
<th>Control Task Vocab</th>
<th>!</th>
<th>after</th>
<th>ran</th>
<th>The</th>
<th>cat</th>
<th>quickly</th>
<th>dog</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>3</td>
<td>42</td>
<td>10</td>
<td>37</td>
<td>15</td>
<td>15</td>
<td>3</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Sentence 1</th>
<th>The</th>
<th>cat</th>
<th>ran</th>
<th>quickly</th>
<th>.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Part-of-speech</td>
<td>DT</td>
<td>NN</td>
<td>VBD</td>
<td>RB</td>
<td>3</td>
</tr>
<tr>
<td>Control task</td>
<td>10</td>
<td>37</td>
<td>10</td>
<td>15</td>
<td>3</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Sentence 2</th>
<th>The</th>
<th>dog</th>
<th>ran</th>
<th>after</th>
<th>!</th>
</tr>
</thead>
<tbody>
<tr>
<td>Part-of-speech</td>
<td>DT</td>
<td>NN</td>
<td>VBD</td>
<td>IN</td>
<td>.</td>
</tr>
<tr>
<td>Control task</td>
<td>10</td>
<td>15</td>
<td>10</td>
<td>42</td>
<td>42</td>
</tr>
</tbody>
</table>

(Hewitt et al., 2019)
designing control tasks

- independently sample a control behavior $C(v)$ for each word type $v$ in the vocabulary
- specifies how to define $y_i \in Y$ for a word token $x_i$ with word type $v$
- control task is a function that maps each token $x_i$ to the label specified by the behavior $C(x_i)$

$$f_{\text{control}}(x_{1:T}) = f(C(x_1), C(x_2), \ldots C(x_T))$$

(Hewitt et al., 2019)
selectivity: high linguistic task accuracy + low control task accuracy

measures the probe model’s ability to make output decisions independently of linguistic properties of the representation

(Hewitt et al., 2019)
be careful about probe accuracies

| Model  | Linear  |  |  | MLP-1  |  |  |
|--------|---------|  |  |        |  |  |
|        | Accuracy| Selectivity |        | Accuracy| Selectivity |        |
| Proj0  | 96.3    | 20.6     |        | 97.1    | 1.6        |        |
| ELMo1  | 97.2    | 26.0     |        | 97.3    | 4.5        |        |
| ELMo2  | 96.6    | 31.4     |        | 97.0    | 8.8        |        |
how to use probe tasks to improve downstream task performance?

• what kinds of linguistic knowledge are important for your task?

• probe BERT for them

• if BERT struggles then fine-tune it with additional probe objectives

\[ \mathcal{L}_{new} = \mathcal{L}_{BERT} + \alpha \mathcal{L}_{probe} \]
example: KnowBERT

(Peters et al., 2019)