CS230: Lecture 10
Sequence models II

Kian Katanforoosh, Andrew Ng, Younes Bensouda Mourri
Today’s outline

We will learn how to:

- Automatically **score an NLP model**
  I. BLEU score

- Improve **Machine Translation** results with Beam search
  II. Beam Search

- Build a **speech recognition application**
  III. Speech Recognition
Neural Machine Translation

Encoder

Decoder

Translation:

```
“Brian” “is” “in” “the” “kitchen” <eos>
```

```
“Brian” “est” “dans” “la” “cuisine” <eos>
```

Parameters:

```
argmax softmax
```

BLEU score
Motivation: Human Evaluation of MT are extensive but expensive.

Goal: Construct a quick, inexpensive, language independent (and correlates highly with human evaluation) method to automatically evaluate Machine Translation models.

Central idea: ”The closer a machine translation is to a professional human translation, the better it is.”
**BLEU score**

**Needs two ingredients:**
- a numerical “translation closeness” metric
- a corpus of good quality human reference translations

**In speech recognition:**
a successful metric is *word error rate*. BLEU’s closeness metric was built after it.
### Machine Translations

| Candidate 1 | It is a guide to action which ensures that the military always obeys the commands of the party |
| Candidate 2 | It is to insure the troops forever hearing the activity guidebook that party directs. |

### Human Translations

| Reference 1 | It is a guide to action that ensures that the military will forever heed Party commands |
| Reference 2 | It is the guiding principle which guarantees the military forces always being under the command of the Party |
| Reference 3 | It is the practical guide for the army always to heed the directions of the party |
### BLEU score

#### Unigram count as precision metric:

<table>
<thead>
<tr>
<th>Machine Translation</th>
<th>Human Translations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Candidate the the the the the the the</td>
<td>Reference 1 The cat is on the mat.</td>
</tr>
<tr>
<td></td>
<td>Reference 2 There is a cat on the mat.</td>
</tr>
</tbody>
</table>

**Standard Unigram Precision**

\[
\text{Standard Unigram Precision} = \frac{\text{# MT words occurring in any reference HT}}{\text{# MT words}} = 100\%
\]

**Modified Unigram Precision**

\[
\text{Modified Unigram Precision} = \frac{\text{# MT words occurring in any reference HT (clipped)}}{\text{# MT words}} = \frac{2}{6} = 33.3\%
\]

---

## BLEU score

### Machine Translations

| Candidate 1 | It is a guide to action which ensures that the military always obeys the commands of the party |
| Candidate 2 | It is to insure the troops forever hearing the activity guidebook that party directs. |

### Human Translations

| Reference 1 | It is a guide to action that ensures that the military will forever heed Party commands |
| Reference 2 | It is the guiding principle which guarantees the military forces always being under the command of the Party |
| Reference 3 | It is the practical guide for the army always to heed the directions of the party |

**Modified Unigram Precision (1)**

\[
\text{Modified Unigram Precision (1)} = \frac{\text{# MT words occurring in any reference HT (clipped)}}{\text{# MT words}} = \frac{17}{18} = 94\%
\]

**Modified Unigram Precision (2)**

\[
\text{Modified Unigram Precision (2)} = \frac{\text{# MT words occurring in any reference HT (clipped)}}{\text{# MT words}} = \frac{8}{14} = 57\%
\]
## BLEU score

**Generalizing to modified n-gram precision metric:**

Modified n-gram precision = \[
\frac{\text{# MT n-grams occurring in any reference HT (clipped)}}{\text{# MT n-grams}}
\]

<table>
<thead>
<tr>
<th>Machine Translations</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Candidate 1</strong></td>
<td>It is a guide to action which ensures that the military always obeys the commands of the party</td>
</tr>
<tr>
<td><strong>Candidate 2</strong></td>
<td>It is to insure the troops forever hearing the activity guidebook that party directs.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Human Translations</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Reference 1</strong></td>
</tr>
<tr>
<td><strong>Reference 2</strong></td>
</tr>
<tr>
<td><strong>Reference 3</strong></td>
</tr>
</tbody>
</table>

**Modified bi-gram precision (Candidate 1) =** 10/17

**Modified bi-gram precision (Candidate 2) =** 1/13
Generalizing from a sentence precision, to a corpus precision:

$$p_n = \frac{\sum_{C \in \{Candidates\}} \sum_{ngram \in C} \text{Count}_{clip}(ngram)}{\sum_{C' \in \{Candidates\}} \sum_{ngram' \in C} \text{Count}(ngram')}$$
Geometric average of modified n-grams precision measures = \[ \exp\left(\sum_{n=1}^{N} w_n \log(p_n)\right) = p_1^{w_1} p_2^{w_2} \ldots p_N^{w_N} \]
## BLEU Score

### Sentence length

<table>
<thead>
<tr>
<th>too short translation</th>
<th>too long translation</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Machine Translations</strong></td>
<td><strong>Machine Translations</strong></td>
</tr>
<tr>
<td>Candidate 1</td>
<td>I always invariably perpetually do</td>
</tr>
<tr>
<td>Candidate 2</td>
<td>I always do</td>
</tr>
<tr>
<td><strong>Human Translations</strong></td>
<td><strong>Human Translations</strong></td>
</tr>
<tr>
<td>Reference 1</td>
<td>It is a guide to action that ensures that the military will forever heed Party commands</td>
</tr>
<tr>
<td>Reference 2</td>
<td>It is the guiding principle which guarantees the military forces always being under the command of the Party</td>
</tr>
<tr>
<td>Reference 3</td>
<td>It is the practical guide for the army always to heed the directions of the party</td>
</tr>
</tbody>
</table>

| Reference 1 | I always do |
| Reference 2 | I invariably do |
| Reference 3 | I perpetually do |

**Conclusion:** *longer translations are already penalized by the modified n-gram precision measure, not the too short translations*

\[
BP = \text{Brevity penalty} = \frac{\text{decaying exponential}}{\text{Total length of the candidate translation corpus}} 
\]

BLEU Score

**BLEU definition:**

\[
BLEU = \left( p_1^{w_1} p_2^{w_2} \ldots p_N^{w_N} \right) \cdot BP
\]

where

\[
BP = \begin{cases} 
1 & \text{if } c > r \\
\exp^{1-\frac{r}{c}} & \text{if } c \leq r
\end{cases}
\]

**log(BLEU) definition:**

\[
\log(BLEU) = \min(1 - \frac{r}{c}, 0) + \sum_{n=1}^{N} w_n \log(p_n)
\]

Beam search

Questions

• What is the main advantage of Beam search compared to other search algorithms?

*It is fast, and requires less computations.*

• What is the main disadvantage of Beam search compared to other search algorithms?

*It may not result in the optimal solution in terms of probability.*

• What is the time and memory complexity of Beam search?

*It is $O(b^*T_x)$ in memory and $O(b^*T_x)$ in time.*
Speech Recognition Pipeline

Audio Data:

X? → model? → Y?

“Hello”
This never happens in practice because:

input length ≠ output length

Speech Recognition

cross-entropy

H
E
L
L
O

softmax
softmax
softmax
softmax
softmax

argmax
argmax
argmax
argmax
argmax

probability
distribution
shape = (28, 1)

0.32
0.21
0.33
0.12
0.22

... 
...
...
...
...

0.64
0.43
0.15
0.14
0.03

Spectrogram

Raw Audio

LSTM → LSTM → LSTM → LSTM → LSTM

softmax
softmax
softmax
softmax
softmax

argmax
argmax
argmax
argmax
argmax

This never happens in practice because:

input length ≠ output length

Kian Katanforoosh, Andrew Ng, Younes Bensouda Mourri
$\beta(\cdot) \rightarrow \hat{y} = "HELLO"$

Graves et al., 2006, Connectionist Temporal Classification: Labelling Unsegmented Sequence Data with Recurrent Neural Networks

https://distill.pub/2017/ctc/
Examples

$$\beta(HH \_EEE \_LL \_LOO) = \text{"HELLO"}$$

$$\beta(H \_E \_L \_LOO) = \text{"HELLO"}$$

$$\beta(H \_LLL \_OO) = \text{"HELO"}$$

$$\beta(BBAA \_NA \_NN \_AA \_A) = \text{"BANANAA"}$$
Speech Recognition

Independence assumption

\[
P(c|x) = \begin{bmatrix}
0.13 & HH\_E\_L\_LL\_OO \\
0.04 & H\_EE\_L\_LL\_HO \\
0.03 & HH\_E\_L\_LL\_OO \\
0.01 & H\_EE\_L\_LL\_OO \\
0.001 & H\_I\_ILL\_L\_OOOO \\
\vdots & \vdots
\end{bmatrix}
\]

\[
P(y|x) = \sum_{c: \beta(c) = y} P(c|x)
\]

\[
P(c_1|x) = \prod_{t=1}^{T_x} P(c_1^{(t)} | x)
\]

\[
P("HELLO") =
\]
Speech Recognition

Loss?

\[ \hat{y} = "HELLO" \]

\[ L = -\sum_{t} \log P(\hat{y}^{(t)} | x^{(t)}) = -\sum_{t} \log \left( \sum_{c \in \mathcal{C}} P(c | x^{(t)}) \right) \]

\[
\begin{bmatrix}
\text{softmax} \\
0 \\
0
\end{bmatrix}
\]

\[
\begin{bmatrix}
\text{softmax} \\
0.31 \\
0.64
\end{bmatrix}
\]

\[ h_1 \]

\[ h_2 \]

\[ h_3 \]

\[ h_4 \]

\[ h_5 \]

\[ h_6 \]

\[ h_7 \]

\[ h_8 \]

\[ h_9 \]

\[ h_{10} \]

\[ h_{11} \]

probability distribution shape = (28,1)

Raw Audio

Graves et al., 2006, Connectionist Temporal Classification: Labelling Unsegmented Sequence Data with Recurrent Neural Networks

Kian Katanforoosh, Andrew Ng, Younes Bensouda Mourri
Speech Recognition

Inference?

BEAM SEARCH > MAX DECODING :)

Graves et al., 2006, Connectionist Temporal Classification: Labelling Unsegmented Sequence Data with Recurrent Neural Networks

Kian Katanforoosh, Andrew Ng, Younes Bensouda Mourri
Speech Recognition

\[ L = - \sum_i \log P(\hat{y}^{(i)}|x^{(i)}) = - \sum_i \log \left( \sum_{c: \beta(c) = \hat{y}^{(i)}} P(c|x^{(i)}) \right) \]

Implementations of CTC loss

`tf.nn.ctc_loss(…)`

Keras -> Custom loss
What is the problem with our output $\hat{y}$?

CTC model makes a lot of spelling and linguistic mistakes because $P(y|x)$ directly models audio data. Some words are hard to spell based on their audios.

What’s the consequence of $P(c_i|x) = \prod_{t=1}^{T} P(c_i^{(t)}|x)$ (conditional independence)?

A model like CTC may have trouble producing such diverse transcripts for the same utterance because of conditional independence assumptions between frames.

But, on the other hand, it makes the model more robust to a change of settings.

How to incorporate information about the future?

A good way to efficiently incorporate future information in speech recognition is still an open problem.

What is the problem with our output $\hat{y}$?

Lipreading.