Machine Translation
06: Attention Models
Analysis and Variants

Rico Sennrich

University of Edinburgh
1 Refresher

2 Problems with Attention

3 Attention Model Variants
Encoder-Decoder with Attention

Decoder

Encoder

natürlich  hat  john  spaß

of  course  john  has  fun

0.7  0.1  0.1  0.1  0.1

R. Sennrich  MT – 2018 – 06
Encoder-Decoder with Attention

Decoder

```
\begin{align*}
  & h_1 \rightarrow h_2 \rightarrow h_3 \rightarrow h_4 \\
  & \text{Decoder} \quad \text{of} \quad \text{course} \quad \text{john} \quad \text{has} \quad \text{fun} \\
  & s_1 \rightarrow s_2 \rightarrow s_3 \rightarrow s_4 \rightarrow s_5 \\
  & x_1 \rightarrow x_2 \rightarrow x_3 \rightarrow x_4 \\
  & \text{Encoder} \\
  & \text{natürlich} \quad \text{hat} \quad \text{john} \quad \text{spaß}
\end{align*}
```

R. Sennrich

MT – 2018 – 06
Encoder-Decoder with Attention

Decoder

\[ y_1 \rightarrow s_1 \rightarrow y_2 \rightarrow s_2 \rightarrow y_3 \rightarrow s_3 \rightarrow y_4 \rightarrow s_4 \rightarrow y_5 \rightarrow s_5 \]

Encoder

\[ x_1 \rightarrow h_1 \rightarrow x_2 \rightarrow h_2 \rightarrow x_3 \rightarrow h_3 \rightarrow x_4 \rightarrow h_4 \]

words:
- natürlich
- hat
- john
- Spaß
- of
- course
- has
- fun

R. Sennrich
MT – 2018 – 06
Encoder-Decoder with Attention

Decoder

\[
\begin{align*}
&y_1 \quad s_1 \\
&y_2 \quad s_2 \\
&y_3 \quad s_3 \\
&y_4 \quad s_4 \\
&y_5 \quad s_5
\end{align*}
\]

Encoder

\[
\begin{align*}
&h_1 \\
&h_2 \\
&h_3 \\
&h_4
\end{align*}
\]

0.1 0.1 0.7 0.1

\[
\begin{align*}
&x_1 \\
&x_2 \\
&x_3 \\
&x_4
\end{align*}
\]

natürlich  hat  john  spaß
Encoder-Decoder with Attention

Decoder

Decoder

Encoder

Naturlich

Hat

John

Spaß

Of course John has fun.
(one type of) attention model

\[ e_{ij} = v_a^\top \tanh(W_a s_{i-1} + U_a h_j) \]
\[ \alpha_{ij} = \text{softmax}(e_{ij}) \]
\[ c_i = \sum_{j=1}^{T_x} \alpha_{ij} h_j \]
Fig. 5. Examples of the attention-based model attending to the correct object (*white* indicates the attended regions, *underlines* indicated the corresponding word) [22]
Equation 11 suggests a Monte Carlo based sampling approximation of the gradient with respect to the model parameters. This can be done by sampling the location $t$ from a multinouilli distribution defined by Equation 8.

$$\tilde{s}_t \sim \text{Multinouilli} \left( \{\alpha_i\} \right)$$

$$\frac{\partial L}{\partial W} \approx \frac{1}{N} \sum_{n=1}^{N} \left[ \frac{\partial \log p(y|\tilde{s}_n, a)}{\partial W} + \log p(y|\tilde{s}_n, a) \frac{\partial \log p(\tilde{s}_n|a)}{\partial W} \right]$$

A moving average baseline is used to reduce the variance in the Monte Carlo estimator of the gradient, following Weaver & Tao (2001). Similar, but more complicated variance reduction techniques have previously been used by Mnih et al. (2014) and Ba et al. (2014). Upon seeing the $k$th mini-batch, the moving average baseline is estimated as an accumulated sum of the previous log likelihoods with exponential decay:

$$b_k = 0.9 \times b_{k-1} + 0.1 \times \log p(y|\tilde{s}_k, a)$$

To further reduce the estimator variance, an entropy term on the multinouilli distribution $H[s]$ is added. Also, with probability 0.5 for a given image, we set the sampled attention location $\tilde{s}$ to its expected value $\alpha$. Both techniques improve the robustness of the stochastic attention learning algorithm. The final learning rule for the model is then the following:

$$\frac{\partial L}{\partial W} \approx \frac{1}{N} \sum_{n=1}^{N} \left[ \frac{\partial \log p(y|\tilde{s}_n, a)}{\partial W} + \lambda_r \left( \log p(y|\tilde{s}_n, a) - b \right) \frac{\partial \log p(\tilde{s}_n|a)}{\partial W} + \lambda_e \frac{\partial H(\tilde{s}_n)}{\partial W} \right]$$

where, $\lambda_r$ and $\lambda_e$ are two hyper-parameters set by cross-validation. As pointed out and used in Ba et al. (2014) and Mnih et al. (2014), this formulation is equivalent to the REINFORCE learning rule (Williams, 1992), where the reward for the attention choosing a sequence of actions is a real value proportional to the log likelihood of the target sentence under the sampled attention trajectory.

In making a hard choice at every point, $\phi(\{a_i\}, \{\alpha_i\})$ from Equation 6 is a function that returns a sampled $a_i$ at every point in time based upon a multinouilli distribution parameterized by $\alpha$.

### 4.2. Deterministic "Soft" Attention

Learning stochastic attention requires sampling the attention location $s_t$ each time, instead we can take the expectation of the context vector $\hat{z}_t$ directly,

$$E_{p(s_t|a)}[\hat{z}_t] = \sum_{i=1}^{L} \alpha_{t,i} a_i$$

and formulate a deterministic attention model by computing a soft attention weighted annotation vector $\phi(\{a_i\}, \{\alpha_i\}) = \sum_{i=1}^{L} \alpha_i a_i$ as introduced by Bahdanau et al. (2014). This corresponds to feeding in a soft $\alpha$.

---

**Figure 5.** Examples of mistakes where we can use attention to gain intuition into what the model saw.

- A large white bird standing in a forest.
- A woman holding a clock in her hand.
- A man wearing a hat and a hat on a skateboard.
- A person is standing on a beach with a surfboard.
- A woman is sitting at a table with a large pizza.
- A man is talking on his cell phone while another man watches.
word-alignment between source and target words is used for various applications

translate rare/unknown words with back-off dictionary:

<table>
<thead>
<tr>
<th>source</th>
<th>The indoor temperature is very pleasant.</th>
</tr>
</thead>
<tbody>
<tr>
<td>reference</td>
<td>Das Raumklima ist sehr angenehm.</td>
</tr>
<tr>
<td>[Bahdanau et al., 2015]</td>
<td>Die UNK ist sehr angenehm</td>
</tr>
<tr>
<td>[Jean et al., 2015]</td>
<td>Die Temperatur ist sehr angenehm.</td>
</tr>
</tbody>
</table>

(more on open-vocabulary MT in future lecture)

attention has been used to obtain alignments. However, ...
1. Refresher

2. Problems with Attention

3. Attention Model Variants
Figure 8: Word alignment for English–German: comparing the attention model states (green boxes with probability in percent if over 10) with alignments obtained from fast-align (blue outlines).

[Koehn and Knowles, 2017]
Attention is not alignment

Figure 9: Mismatch between attention states and desired word alignments (German–English).

[Koehn and Knowles, 2017]
Attention is not alignment

discuss in pairs
how can NMT model translate text, even if attention is off?
Attention is not alignment
1 Refresher

2 Problems with Attention

3 Attention Model Variants
Obtaining Attention Scores

\[
\text{score}(h_t, \bar{h}_s) = \begin{cases} 
   h_t^\top h_s & \text{dot} \\
   h_t^\top W_a \bar{h}_s & \text{general} \\
   \nu_a^\top \tanh (W_a[h_t; \bar{h}_s]) & \text{concat}
\end{cases}
\]

Attention variants from [Luong et al., 2015]

- many ways to score encoder states:
- \textit{concat}: attention as introduced by [Bahdanau et al., 2015]
- \textit{dot}: more attention on similar vectors
Conditioning Attention on Past Decisions

attention in dl4mt-tutorial (and Nematus):

\[ s'_i = GRU_1(s_{i-1}, y_{i-1}) \]
\[ c_i = ATT(C, s'_i) \]
\[ s_i = GRU_2(c_i, s'_i) \]

**motivation**

- (simple) attention model from lecture 4 is only conditioned on \( s_{i-1} \)...
- ...but it also matters which word we predicted last (\( y_{i-1} \))
- more transitions per timestep → more depth
  [Miceli Barone et al., 2017]
Guided Alignment Training [Chen et al., 2016]

**core idea**

1. compute alignment with external tool (IBM models; discussed in later lecture)
2. if multiple source words align to same target words, normalize so that \( \sum_j A_{ij} = 1 \)
3. modify objective function of NMT training:
   - minimize target sentence cross-entropy (as before)
   - minimize divergence between model attention \( \alpha \) and external alignment \( A \):

\[
H(A, \alpha) = -\frac{1}{T_y} \sum_{i=1}^{T_y} \sum_{j=1}^{T_x} A_{ij} \log \alpha_{ij}
\]
Incorporating Structural Alignment Biases

core idea [Cohn et al., 2016]

we know that alignment has some biases, which are exploited in statistical word alignment algorithms [Brown et al., 1990, Koehn et al., 2003]:

- position bias: relative position is highly informative for alignment
- fertility/coverage: some words produce multiple words in target language
  all source words should be covered (respecting fertility)
- bilingual symmetry: $\alpha^s\leftarrow t$ and $\alpha^s\rightarrow t$ are symmetrical
Incorporating Structural Alignment Biases

position bias

- provide attention model with positional information
- found to be especially helpful with non-recurrent architectures
- different choices for positional encoding:
  - [Cohn et al., 2016]: $\log(1 + i)$
  - [Gehring et al., 2017]: positional embedding: $E(i)$
  - [Vaswani et al., 2017]: sine/cosine function
Incorporating Structural Alignment Biases

coverage without fertility

reminder:

\[ \sum_j \alpha_{ij} = 1 \] (softmax)

idea: model should attend to each source word exactly once:

\[ \sum_i \alpha_{ij} \approx 1 \] (our goal)

we can bias model towards this goal with regularisation term:

\[ \sum_j (1 - \sum_i \alpha_{ij})^2 \] (to be minimized)

discuss in pairs

is this the right goal? why / why not?
Incorporating Structural Alignment Biases

coverage with fertility [Cohn et al., 2016, Tu et al., 2016]

idea: learn fertility of words with neural network:

\[ f_j = N \sigma(W_j h_j) \]

coverage objective that takes fertility into account:

\[ \sum_{j} (f_j - \sum_i \alpha_{i,j})^2 \]  

(to be minimized)
Incorporating Structural Alignment Biases

**bilingual symmetry**

joint training objective with *trace bonus* $B$, which rewards symmetric attention:

$$B(\alpha^{s\leftarrow t}, \alpha^{s\rightarrow t}) = \sum_{i=1}^{T_y} \sum_{j=1}^{T_x} \alpha_{i,j}^{s\rightarrow t} \alpha_{j,i}^{s\leftarrow t}$$
Further Reading

Coursework

- available at the end of this week
- deadline: March 15, 3pm
- you are encouraged to work in pairs. More details to follow
- training models takes hours or days, so **start early**
- I will have no sympathy if you don’t realize you can’t do this coursework last minute

Lab Sessions

- two lab sessions will provide support getting started (installation of tools and virtual environment)
  - Tuesday, February 6, 15.10-16.00
    Room 4.12, Appleton Tower
  - Wednesday, February 7, 15.10-16.00
    Room 5.08, North Lab, Appleton Tower
- attendance **not** mandatory
Neural Machine Translation by Jointly Learning to Align and Translate.

Computational Linguistics, 16(2):79–85.

Guided Alignment Training for Topic-Aware Neural Machine Translation.
CoRR, abs/1607.01628.

Incorporating Structural Alignment Biases into an Attentional Neural Translation Model.

Convolutional Sequence to Sequence Learning.
CoRR, abs/1705.03122.

On Using Very Large Target Vocabulary for Neural Machine Translation.
In Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing, pages 1–10, Beijing, China. Association for Computational Linguistics.
Six Challenges for Neural Machine Translation.

Statistical Phrase-based Translation.

Effective Approaches to Attention-based Neural Machine Translation.

Deep Architectures for Neural Machine Translation.

Modeling Coverage for Neural Machine Translation.

Attention Is All You Need.
CoRR, abs/1706.03762.