Token-level and sequence-level loss smoothing for RNN language models

Maha Elbayad\textsuperscript{1,2}, Laurent Besacier\textsuperscript{1}, and Jakob Verbeek\textsuperscript{2}

\textsuperscript{1}LIG, \textsuperscript{2}INRIA, Grenoble, France

ACL 2018 Melbourne, Australia
Ground truth sequences lie in a union of low-dimensional subspaces where sequences convey the same message.

- France won the world cup for the second time.
- France captured its second world cup title.

Some words in the vocabulary share the same meaning.

- Capture, conquer, win, gain, achieve, accomplish, ...
Contributions

Take into consideration the nature of the target language space with:

- A token-level smoothing for a “robust” multi-class classification.
- A sequence-level smoothing to explore relevant alternative sequences.
Maximum likelihood estimation (MLE)

For a pair \((x, y)\), we model the conditional distribution:

\[
p_\theta(y|x) = \prod_{t} p_\theta(y_t|y_{<t}, x)
\] (1)

Given the ground truth target sequence \(y^*\):

\[
\ell_{\text{MLE}}(y^*, x) = -\ln p_\theta(y^*|x) = D_{\text{KL}}(\delta(y|y^*)\|p_\theta(y|x))
\] (2)

\[
= \sum_{t=1}^{\mid y^* \mid} D_{\text{KL}}(\delta(y_t|y_t^*)\|p_\theta(y_t|y_{<t}^*, x))
\] (3)
Maximum likelihood estimation (ML)

\[ \ell_{\text{MLE}}(y^*, x) = -\ln p_\theta(y^* | x) \]
\[ = D_{\text{KL}}(\delta(y | y^*)\|p_\theta(y | x)) \quad (2) \]
\[ = \sum_{t=1}^{T} D_{\text{KL}}(\delta(y_t | y_t^*)\|p_\theta(y_t | h_t)) \quad (3) \]

**Issues:**

- Zero-one loss, all the outputs \( y \neq y^* \) are treated equally.
- Discrepancy at the sentence level between the training (1-gram) and evaluation metric (4-gram).
Loss smoothing

\[ \delta(y^*) \]

\[ D_{KL}(\delta(y|y^*)\|p_\theta(y|x)) \]

\[ r_\tau(y|y^*) \]

\[ \ell_{SEQ}^{RAML}(y^*, x) = D_{KL}(r_\tau(y|y^*)\|p_\theta(y|x)) \] (Norouzi et al, 2016)
Loss smoothing

\[ \delta(y^*) \text{ (resp. } \delta(y^*_t)) \]

\[ D_{KL}(\delta(y|y^*)\|p_\theta(y|x)) \]

\[ \sum_{t=1}^{T} D_{KL}(\delta(y_t|y^*_t)\|p_\theta(y_t|h_t)) \]

\[ r_\tau(y|y^*) \text{ (resp. } r_\tau(y_t|y^*_t)) \]

\[ \ell_{seq}^{RAML}(y^*, x) = D_{KL}(r_\tau(y|y^*)\|p_\theta(y|x)) \text{ (Norouzi et al, 2016)} \]

\[ \ell_{tok}^{RAML}(y^*, x) = \sum_{t=1}^{T} D_{KL}(r_\tau(y_t|y^*_t)\|p_\theta(y_t|h_t)) \]

ACL 2018, Melbourne
M. Elbayad || Token-level and Sequence-level Loss Smoothing
Token-level smoothing
Loss smoothing | Token-level

\[
\ell_{RAML}^{\text{tok}}(y^*, x) = \sum_{t=1}^{T} D_{\text{KL}}(r_\tau(y_t|y^*_t)||p_\theta(y_t|h_t))
\]  

- Uniform label smoothing over all words in the vocabulary:

\[
r_\tau(y_t|y^*_t) = \delta(y_t|y^*_t) + \tau.u(V)
\]  

(Szegedy et al. 2016)

- We can leverage word co-occurrence statistics to build a non-uniform and “meaningful” distribution.
Loss smoothing

\[
\ell_{RAML}^{tok}(y^*, x) = \sum_{t=1}^{T} D_{KL}(r_\tau(y_t|y_t^*)\|p_\theta(y_t|h_t))
\]  

Prerequisite: A word embedding \( w \) (e.g. Glove) in the target space and a distance \( d \).

\[
r_\tau(y_t|y_t^*) = \frac{1}{Z} \exp \left( -\frac{d(w(y_t), w(y_t^*))}{\tau} \right),
\]

with a temperature \( \tau \) st. \( r_\tau \xrightarrow{\tau \to 0} \delta \).

\[
Z \text{ st. } \sum_{y_t \in V} r_\tau(y_t|y_t^*) = 1
\]
Loss smoothing | Token-level

$\tau = 0.12$

$\tau = 0.70$
Loss smoothing | Token-level

\[
\mathcal{L}_{\text{RAML}}^{\text{tok}}(y^*, x) = \sum_{t=1}^{T} D_{\text{KL}}(r_{\tau}(y_t|y^*_t) \| p_\theta(y_t|h_t)) = \sum_{t=1}^{T} \sum_{y_t \in V} r_{\tau}(y_t|y^*_t) \log \left( \frac{r_{\tau}(y_t|y^*_t)}{p_\theta(y_t|h_t)} \right)
\]

(4)

(5)

We can estimate the exact KL divergence for every target token. No approximation needed.
Sequence-level smoothing
\[
\ell_{\text{RAML}}^{\text{seq}}(y^*, x) = D_{\text{KL}}(r_\tau(y|y^*) \| p_\theta(y|x))
\]

**Prerequisite:** A distance \( d \) in the sequences space \( \mathcal{V}^n, n \in \mathbb{N} \).

\[
r_\tau(y|y^*) = \frac{1}{Z} \exp \left( -\frac{d(y, y^*)}{\tau} \right),
\]

\( Z \) st. \[ \sum_{y \in \mathcal{V}^n, n \in \mathbb{N}} r_\tau(y|y^*) = 1 \]

Possible (pseudo-)distances:

- Hamming
- Edit
- 1–BLEU
- 1–CIDEr
Can we evaluate the partition function $Z$ for a given reward?

$$r_\tau(y_t|y^*_t) = \frac{1}{Z} \exp \left( \frac{-d(y, y^*)}{\tau} \right),$$

$$Z = \sum_{y \in \mathcal{V}^n, n \in \mathbb{N}} \exp \left( \frac{-d(y, y^*)}{\tau} \right)$$

We can approximate $Z$ for Hamming distance.
Assumption:
consider only sequences of the same length as $y^\star$ ($d(y, y') = 0$ if $|y| \neq |y'|$).
We partition the set of sequences $\mathcal{V}^T$ w.r.t. their distance to the ground truth $y^\star$:

$$
\begin{align*}
S_d &= \{y \in \mathcal{V}^T_{sub} | d(y, y^\star) = d\}, \\
\mathcal{V}^T = \bigcup_d S_d, \\
\forall d, d' : S_d \cap S_{d'} = \emptyset.
\end{align*}
$$

- The reward in each subset is a constant.
- The cardinality of each subset is known.

$$
Z = \sum_d |S_d| \exp \left( -\frac{d}{\tau} \right)
$$
We can easily draw from $r_T$ with Hamming distance:

1. Sample a distance $d$ from $\{0, \ldots, T\}$.
2. Pick $d$ positions in the sequence to be changed among $\{1, \ldots, T\}$.
3. Sample substitutions from $\mathcal{V}$ of the vocabulary.
We can easily draw from $r_T$ with Hamming distance:

1. Sample a distance $d$ from \{0, \ldots, T\}.
2. Pick $d$ positions in the sequence to be changed among \{1, \ldots, T\}.
3. Sample substitutions from $\mathcal{V}$ of the vocabulary.

**Monte Carlo estimation:**

$$\ell^{\text{seq}}_{\text{RAML}}(y^*, x) = D_{\text{KL}}(r_T(y|y^*)\| p_\theta(y|x))$$

$$= -\mathbb{E}_{r_T}[\log p_\theta(.|x)] + \text{cst}$$

$$(y' \sim r_T) \approx -\frac{1}{L} \sum_{l=1}^L \log p_\theta(y'|x)$$
We cannot “easily” sample from more complicated rewards such as BLEU or CIDEr. 

**Importance sampling:**

\[
\ell_{\text{seq}}^{\text{RAML}}(y^*, x) = -\mathbb{E}_{r^\tau}[\log p_\theta(.|x)] \\
= -\mathbb{E}_q[r^\tau \log p_\theta] \\
(y' \sim q) \approx -\frac{1}{L} \sum_{l=1}^L \omega_l \log p_\theta(y'|x) \\
\omega_l \approx \frac{r^\tau(y'|y^*)/q(y'|y^*)}{\sum_{k=1}^L r^\tau(y^k|y^*)/q(y^k|y^*)},
\]

Choose \( q \) the reward distribution relative to Hamming distance.
Loss smoothing | Sequence-level | Support reduction

\[ \ell_{\text{RAML}}^{\text{seq}}(y^*, x) = D_{\text{KL}}(r_\tau(y|y^*)\|p_\theta(y|x)) \]  

Can we reduce the support of \( r_\tau \)?

\[ r_\tau(y|y^*) = \frac{1}{Z} \exp \left( -\frac{d(y, y^*)}{\tau} \right), \quad Z = \sum_{y \in \mathcal{Y}^\tau} \exp \left( -\frac{d(y, y^*)}{\tau} \right) \]

Reduce the support from \( \mathcal{Y}|y^* | \) to \( \mathcal{Y}_{\text{sub}}|y^* | \) where \( \mathcal{Y}_{\text{sub}} \subset \mathcal{Y} \).

- \( \mathcal{Y}_{\text{sub}} = \mathcal{Y}_{\text{batch}} \): tokens occurring in the SGD mini-batch.
- \( \mathcal{Y}_{\text{sub}} = \mathcal{Y}_{\text{refs}} \): tokens occurring in the available references.
Loss smoothing  |  Sequence-level  |  Lazy training

Default training

\[
\ell_{\text{seq}}^{\text{RAML}}(y^*, x) = -\mathbb{E}_{r}[\log p_{\theta}(\cdot|x)] \\
\approx -\frac{1}{L} \sum_{l=1}^{L} \log p_{\theta}(y'_l|x)
\]

\forall l, y'_l is:
1. forwarded in the RNN.
2. used as target.

\[
\log p_{\theta}(y'_l|y'_l, x)
\]

Lazy training

\[
\ell_{\text{seq}}^{\text{RAML}}(y^*, x) = -\mathbb{E}_{r}[\log p_{\theta}(\cdot|x)] \\
\approx -\frac{1}{L} \sum_{l=1}^{L} \log p_{\theta}(y'_l|x)
\]

\forall l, y'_l is:
1. not forwarded in the RNN.
2. used as target.

\[
\log p_{\theta}(y'_l|y^*, x)
\]
Loss smoothing | Sequence-level | Lazy training

**Default training**

\[
\ell_{\text{RAAML}}^{\text{seq}}(y^*, x) = -\mathbb{E}_{r_T} [\log p_\theta(.|x)] \\
\approx -\frac{1}{L} \sum_{l=1}^{L} \log p_\theta(y^l|x)
\]

\(\forall l, y^l\) is:
1. **forwarded** in the RNN.
2. used as target.

\[\log p_\theta(y_l|y_l^l, x)\]

Complexity : \(O(2L\lambda)\)

\[\lambda = |y||\theta_{\text{cell}}|, \text{ where } \theta_{\text{cell}} \text{ are the cell parameters.}\]

**Lazy training**

\[
\ell_{\text{RAAML}}^{\text{seq}}(y^*, x) = -\mathbb{E}_{r_T} [\log p_\theta(.|x)] \\
\approx -\frac{1}{L} \sum_{l=1}^{L} \log p_\theta(y^l|x)
\]

\(\forall l, y^l\) is:
1. **not forwarded** in the RNN.
2. used as target.

\[\log p_\theta(y_l|y^*, x)\]

Complexity: \(O((L+1)\lambda)\)
Experiments
**Image captioning on MS-COCO | Setup**

**Ground truth:**
- two soccer players pushing against each other as they try to get to the ball
- a man standing next to another man while kicking a soccer ball
- two men in a soccer field chasing a ball
- two soccer players pushing each other for the ball
- two soccer players appear to be pushing each other

**Generated:**
a couple of men playing a game of soccer

**Ground truth:**
- a small blue plane sitting on top of a field
- an e2 airplane painted blue with black and white stripes
- model airplane with an american insignia and stripes on wings
- an old warplane is on display in a field.
- a blue small plane standing at the airstri

**Generated:**
a small plane is sitting on the grass

- **5 captions for every image.**
- \(|\mathcal{V}| \approx 10k\) words (freq \(\geq 5\))

<table>
<thead>
<tr>
<th>images</th>
</tr>
</thead>
<tbody>
<tr>
<td>Train</td>
</tr>
<tr>
<td>Dev</td>
</tr>
<tr>
<td>Test</td>
</tr>
</tbody>
</table>

(Lin et al. 2014, Karpathy et al. 2015)

- **Architecture:**
  - Top-down attention
  
  (Anderson et al. 2017)
<table>
<thead>
<tr>
<th>Loss</th>
<th>Reward</th>
<th>$\nu_{sub}$</th>
<th>BLEU-1</th>
<th>BLEU-4</th>
<th>CIDEr</th>
</tr>
</thead>
<tbody>
<tr>
<td>MLE</td>
<td></td>
<td></td>
<td>73.40</td>
<td>33.11</td>
<td>101.63</td>
</tr>
<tr>
<td>Tok</td>
<td>Glove, cosine</td>
<td></td>
<td><strong>74.01</strong></td>
<td><strong>33.25</strong></td>
<td><strong>102.81</strong></td>
</tr>
</tbody>
</table>
### Results

<table>
<thead>
<tr>
<th>Loss</th>
<th>Reward</th>
<th>$\mathcal{V}_{sub}$</th>
<th>BLEU-1</th>
<th>BLEU-4</th>
<th>CIDEr</th>
</tr>
</thead>
<tbody>
<tr>
<td>MLE</td>
<td></td>
<td></td>
<td>73.40</td>
<td>33.11</td>
<td>101.63</td>
</tr>
<tr>
<td>Tok</td>
<td>Glove, cosine</td>
<td></td>
<td>74.01</td>
<td>33.25</td>
<td>102.81</td>
</tr>
<tr>
<td>Seq</td>
<td>Hamming</td>
<td>$\mathcal{V}$</td>
<td>73.12</td>
<td>32.71</td>
<td>101.25</td>
</tr>
<tr>
<td>Seq</td>
<td>Hamming</td>
<td>$\mathcal{V}_{batch}$</td>
<td>73.26</td>
<td>32.73</td>
<td>101.90</td>
</tr>
<tr>
<td>Seq, lazy</td>
<td>Hamming</td>
<td>$\mathcal{V}_{batch}$</td>
<td>73.43</td>
<td>32.95</td>
<td>102.03</td>
</tr>
</tbody>
</table>

(Norouzi et al. 2016)
<table>
<thead>
<tr>
<th>Loss</th>
<th>Reward</th>
<th>$\mathcal{V}_{sub}$</th>
<th>BLEU-1</th>
<th>BLEU-4</th>
<th>CIDEr</th>
</tr>
</thead>
<tbody>
<tr>
<td>MLE</td>
<td></td>
<td></td>
<td>73.40</td>
<td>33.11</td>
<td>101.63</td>
</tr>
<tr>
<td>Tok</td>
<td>Glove, cosine</td>
<td></td>
<td>74.01</td>
<td>33.25</td>
<td>102.81</td>
</tr>
<tr>
<td>Seq</td>
<td>Hamming</td>
<td>$\mathcal{V}$</td>
<td>73.12</td>
<td>32.71</td>
<td>101.25</td>
</tr>
<tr>
<td>Seq</td>
<td>Hamming</td>
<td>$\mathcal{V}_{batch}$</td>
<td>73.26</td>
<td>32.73</td>
<td>101.90</td>
</tr>
<tr>
<td>Seq, lazy</td>
<td>Hamming</td>
<td>$\mathcal{V}_{batch}$</td>
<td>73.43</td>
<td>32.95</td>
<td>102.03</td>
</tr>
<tr>
<td>Seq</td>
<td>CIDEr</td>
<td>$\mathcal{V}_{batch}$</td>
<td>73.50</td>
<td>33.04</td>
<td>102.98</td>
</tr>
<tr>
<td>Seq</td>
<td>CIDEr</td>
<td>$\mathcal{V}_{refs}$</td>
<td>73.42</td>
<td>32.91</td>
<td>102.23</td>
</tr>
<tr>
<td>Seq, lazy</td>
<td>CIDEr</td>
<td>$\mathcal{V}_{refs}$</td>
<td>73.92</td>
<td>33.10</td>
<td>102.64</td>
</tr>
</tbody>
</table>
## Image captioning on MS-COCO | Results

<table>
<thead>
<tr>
<th>Loss</th>
<th>Reward</th>
<th>( V_{sub} )</th>
<th>BLEU-1</th>
<th>BLEU-4</th>
<th>CIDEr</th>
</tr>
</thead>
<tbody>
<tr>
<td>MLE</td>
<td></td>
<td>73.40</td>
<td></td>
<td>33.11</td>
<td>101.63</td>
</tr>
<tr>
<td>Tok</td>
<td>Glove, cosine</td>
<td>74.01</td>
<td></td>
<td>33.25</td>
<td>102.81</td>
</tr>
<tr>
<td>Seq</td>
<td>Hamming</td>
<td>( V )</td>
<td>73.12</td>
<td>32.71</td>
<td>101.25</td>
</tr>
<tr>
<td>Seq</td>
<td>Hamming</td>
<td>( V_{\text{batch}} )</td>
<td>73.26</td>
<td>32.73</td>
<td>101.90</td>
</tr>
<tr>
<td>Seq, lazy</td>
<td>Hamming</td>
<td>( V_{\text{batch}} )</td>
<td>73.43</td>
<td>32.95</td>
<td>102.03</td>
</tr>
<tr>
<td>Seq</td>
<td>CIDEr</td>
<td>( V_{\text{batch}} )</td>
<td>73.50</td>
<td>33.04</td>
<td>102.98</td>
</tr>
<tr>
<td>Seq</td>
<td>CIDEr</td>
<td>( V_{\text{refs}} )</td>
<td>73.42</td>
<td>32.91</td>
<td>102.23</td>
</tr>
<tr>
<td>Seq, lazy</td>
<td>CIDEr</td>
<td>( V_{\text{refs}} )</td>
<td>73.92</td>
<td>33.10</td>
<td>102.64</td>
</tr>
<tr>
<td>Tok-Seq</td>
<td>CIDEr</td>
<td>( V_{\text{refs}} )</td>
<td><strong>74.28</strong></td>
<td><strong>33.34</strong></td>
<td><strong>103.81</strong></td>
</tr>
</tbody>
</table>
**Machine translation | Setup**

- **Architecture:** Bi-LSTM encoder-decoder with attention (Bahdanau et al. 2015)
- **Corpora:**

  **IWSLT’14 DE→EN**
  
<table>
<thead>
<tr>
<th>Pairs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Train</td>
</tr>
<tr>
<td>Dev</td>
</tr>
<tr>
<td>Test</td>
</tr>
</tbody>
</table>

  - $|\mathcal{V}| = 22k$ words.

  **WMT’14 EN→FR**
  
<table>
<thead>
<tr>
<th>Pairs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Train</td>
</tr>
<tr>
<td>Dev</td>
</tr>
<tr>
<td>Test</td>
</tr>
</tbody>
</table>

  - $|\mathcal{V}| = 30k$ words.
## Machine translation | Results

<table>
<thead>
<tr>
<th>Loss</th>
<th>Reward</th>
<th>$\mathcal{V}_{sub}$</th>
<th>WMT’14 En→Fr</th>
<th>IWSLT’14 De→En</th>
</tr>
</thead>
<tbody>
<tr>
<td>MLE</td>
<td></td>
<td></td>
<td>30.03</td>
<td>27.55</td>
</tr>
<tr>
<td>tok</td>
<td>Glove, cosine</td>
<td></td>
<td><strong>30.19</strong></td>
<td><strong>27.83</strong></td>
</tr>
</tbody>
</table>
### Machine translation Results

<table>
<thead>
<tr>
<th>Loss</th>
<th>Reward</th>
<th>$\mathcal{V}_{sub}$</th>
<th>WMT’14 En→Fr</th>
<th>IWSLT’14 De→En</th>
</tr>
</thead>
<tbody>
<tr>
<td>MLE</td>
<td></td>
<td></td>
<td>30.03</td>
<td>27.55</td>
</tr>
<tr>
<td>tok</td>
<td>Glove, cosine</td>
<td>$\mathcal{V}$</td>
<td>30.19</td>
<td>27.83</td>
</tr>
<tr>
<td>Seq</td>
<td>Hamming</td>
<td>$\mathcal{V}_{batch}$</td>
<td>30.85</td>
<td>27.98</td>
</tr>
<tr>
<td>Seq</td>
<td>Hamming</td>
<td>$\mathcal{V}_{batch}$</td>
<td>31.18</td>
<td><strong>28.54</strong></td>
</tr>
<tr>
<td>Seq</td>
<td>BLEU-4</td>
<td>$\mathcal{V}_{batch}$</td>
<td><strong>31.29</strong></td>
<td>28.53</td>
</tr>
</tbody>
</table>

(Norouzi et al. 2016)
<table>
<thead>
<tr>
<th>Loss</th>
<th>Reward</th>
<th>$\mathcal{V}_{sub}$</th>
<th>WMT’14 En→Fr</th>
<th>IWSLT’14 De→En</th>
</tr>
</thead>
<tbody>
<tr>
<td>MLE</td>
<td></td>
<td>$\mathcal{V}$</td>
<td>30.03</td>
<td>27.55</td>
</tr>
<tr>
<td>tok</td>
<td>Glove, cosine</td>
<td>$\mathcal{V}$</td>
<td>30.19</td>
<td>27.83</td>
</tr>
<tr>
<td>Seq</td>
<td>Hamming</td>
<td>$\mathcal{V}$</td>
<td>30.85</td>
<td>27.98</td>
</tr>
<tr>
<td>Seq</td>
<td>Hamming</td>
<td>$\mathcal{V}_{batch}$</td>
<td>31.18</td>
<td>28.54</td>
</tr>
<tr>
<td>Seq</td>
<td>BLEU-4</td>
<td>$\mathcal{V}_{batch}$</td>
<td>31.29</td>
<td>28.53</td>
</tr>
<tr>
<td>Tok-Seq</td>
<td>Hamming</td>
<td>$\mathcal{V}_{batch}$</td>
<td>31.36</td>
<td>28.70</td>
</tr>
<tr>
<td>Tok-Seq</td>
<td>BLEU-4</td>
<td>$\mathcal{V}_{batch}$</td>
<td><strong>31.39</strong></td>
<td><strong>28.74</strong></td>
</tr>
</tbody>
</table>
Conclusion
Takeaways

Improving over MLE with:

- **Sequence-level smoothing**: an extension of RAML (Norouzi et al. 2016)
  - Reduced support of the reward distribution.
  - Importance sampling.
  - Lazy training.

- **Token-level smoothing**: smoothing across semantically similar tokens instead of the usual uniform noise.

Both schemes can be combined for better results.
Takeaways

Improving over MLE with:

- **Sequence-level smoothing**: an extension of RAML (Norouzi et al. 2016)
  - Reduced support of the reward distribution.
  - Importance sampling.
  - Lazy training.

- **Token-level smoothing**: smoothing across semantically similar tokens instead of the usual uniform noise.

Both schemes can be combined for better results.
Future work

- Validate on other seq2seq models besides LSTM encoder-decoders.
- Validate on models with BPE instead of words.
- **Sequence-level smoothing:**
  - Experiment with other distributions for sampling other than the Hamming distance.
- **Token-level smoothing:**
  - Sparsify the reward distribution for scalability.
Thank you!
Appendices
Hyper-parameters: $\alpha, \alpha_1, \alpha_2 \in (0, 1)$ ($\forall \alpha, \bar{\alpha} = 1 - \alpha$).

**Combining ML and RAML:**

$$
\ell_{\text{seq},\alpha}(y^*, x) = \alpha \ell_{\text{seq}}^{\text{RAML}}(y^*, x) + \bar{\alpha} \ell_{\text{MLE}}(y^*, x) \tag{12}
$$

$$
\ell_{\text{tok},\alpha}(y^*, x) = \alpha \ell_{\text{tok}}^{\text{RAML}}(y^*, x) + \bar{\alpha} \ell_{\text{MLE}}(y^*, x) \tag{13}
$$

**Combining the smoothing schemes:**

$$
\ell_{\text{seq, tok}}^{\text{RAML},\alpha_1,\alpha_2}(y^*, x) = \alpha_1 \mathbb{E}_r[\ell_{\text{tok}}^{\text{RAML}}(y, x)] + \bar{\alpha}_1 \ell_{\text{tok}}^{\text{RAML}}(y^*, x)
$$

$$
= \alpha_1 \mathbb{E}_r[\alpha_2 \ell_{\text{tok}}^{\text{RAML}}(y, x) + \bar{\alpha}_2 \ell_{\text{MLE}}(y, x)]
$$

$$
+ \bar{\alpha}_1(\alpha_2 \ell_{\text{tok}}^{\text{RAML}}(y^*, x) + \bar{\alpha}_2 \ell_{\text{MLE}}(y^*, x)) \tag{14}
$$
### Training time

Average wall time to process a single batch (10 images 50 captions) when training the RNN language model with fixed CNN (without attention) on a Titan X GPU.

<table>
<thead>
<tr>
<th>Loss</th>
<th>MLE</th>
<th>Tok</th>
<th>Seq</th>
<th>Seq lazy</th>
<th>Seq</th>
<th>Seq lazy</th>
<th>Seq</th>
<th>Seq lazy</th>
<th>Tok-Seq</th>
<th>Tok-Seq</th>
<th>Tok-Seq</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reward</td>
<td>Glove sim</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\mathcal{V}_{\text{sub}}$ ms/batch</td>
<td>347</td>
<td>359</td>
<td>390</td>
<td>349</td>
<td>395</td>
<td>337</td>
<td>401</td>
<td>336</td>
<td>445</td>
<td>446</td>
<td>453</td>
</tr>
</tbody>
</table>
Generated captions

**Ground truth:**
two zebra’s standing in a grassy field and one is eating grass
a zebra looking up as another grazes in a field
the zebras are grazing out in the field of grass.
a group of zebras stand together in a field
several zebras eating grass in a wildlife park

**Generated:**
Baseline: a couple of zebra standing on top of a grass covered field
Seq: a couple of zebra standing on top of a grass covered field
Tok: a couple of zebra standing next to each other on a field
Tok-Seq: a couple of zebras are standing in a field

**Ground truth:**
a bunch of bananas and a orange sitting in a pile
bananas and an orange are sitting together on the cplot
five yellow bananas and one orange orange togethe
a tangering sitting on top of some bananas
there is one orange laying among five banana

**Generated:**
Baseline: a bunch of bananas sitting on a table
Seq: a close up of a bunch of bananas
Tok: a bunch of bananas that are on a table
Tok-Seq: a bunch of bananas sitting next to a banana
Generated captions

**Ground truth:**
- a living room with a sectional couch, easy chair and glass desk covered in paper
- home living room with brown walls with white trim, fireplace, and tan furnishings
- a living room includes a beige sofa and a black fireplace
- a couch and a chair in a small living room
- a living area with sofa, chair and a fireplace

**Generated:**
- Baseline: a living room filled with furniture and a large window
- Seq: a living room filled with furniture and a tv
- Tok: a living room with a couch and a desk
- Tok-Seq: a living room filled with furniture and a fireplace

**Ground truth:**
- an antique pickup truck restored and displayed at a fair or car show
- a vintage truck is parked at an outdoor event
- antique blue truck displayed to crowd at outdoor event
- an old blue truck is on a grassy area
- a car at a show with people in background

**Generated:**
- Baseline: a red truck is parked in a field
- Seq: an old truck is parked in a field
- Tok: a blue truck parked in a grassy field
- Tok-Seq: an old blue truck parked in a field
### Generated translations En→Fr

<table>
<thead>
<tr>
<th>Source (en)</th>
<th>I think it’s conceivable that these data are used for mutual benefit.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Target (fr)</td>
<td>J’estime qu’il est concevable que ces données soient utilisées dans leur intérêt mutuel.</td>
</tr>
<tr>
<td>MLE</td>
<td>Je pense qu’il est possible que ces données soient utilisées à des fins réciproques.</td>
</tr>
<tr>
<td>Tok-Seq</td>
<td>Je pense qu’il est possible que ces données soient utilisées pour le bénéfice mutuel.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Source (en)</th>
<th>The public will be able to enjoy the technical prowess of young skaters, some of whom, like Hyeres’ young star, Lorenzo Palumbo, have already taken part in top-notch competitions.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Target (fr)</td>
<td>Le public pourra admirer les prouesses techniques de jeunes qui, pour certains, fréquentent déjà les compétitions au plus haut niveau, à l’instar du jeune prodige hyérois Lorenzo Palumbo.</td>
</tr>
<tr>
<td>MLE</td>
<td>Le public sera en mesure de profiter des connaissances techniques des jeunes garçons, dont certains, à l’instar de la jeune star américaine, Lorenzo, ont déjà participé à des compétitions de gymnastique.</td>
</tr>
<tr>
<td>Tok-Seq</td>
<td>Le public sera en mesure de profiter de la finesse technique des jeunes musiciens, dont certains, comme la jeune star de l’entreprise, Lorenzo, ont déjà pris part à des compétitions de gymnastique.</td>
</tr>
</tbody>
</table>
# MS-COCO server results

<table>
<thead>
<tr>
<th></th>
<th>BLEU-1</th>
<th>BLEU-2</th>
<th>BLEU-3</th>
<th>BLEU-4</th>
<th>METEOR</th>
<th>ROUGE-L</th>
<th>CIDEr</th>
<th>SPICE</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>c5</td>
<td>c40</td>
<td>c5</td>
<td>c40</td>
<td>c5</td>
<td>c40</td>
<td>c5</td>
<td>c40</td>
</tr>
<tr>
<td>Google NIC+ (Vinyals et al., 2015)</td>
<td>71.3</td>
<td>89.5</td>
<td>54.2</td>
<td>80.2</td>
<td>40.7</td>
<td>69.4</td>
<td>30.9</td>
<td>58.7</td>
</tr>
<tr>
<td>Hard-Attention (Xu et al., 2015)</td>
<td>70.5</td>
<td>88.1</td>
<td>52.8</td>
<td>77.9</td>
<td>38.3</td>
<td>65.8</td>
<td>27.7</td>
<td>53.7</td>
</tr>
<tr>
<td>ATT-FCN+ (You et al., 2016)</td>
<td>73.1</td>
<td>90.0</td>
<td>56.5</td>
<td>81.5</td>
<td>42.4</td>
<td>70.9</td>
<td>31.6</td>
<td>59.9</td>
</tr>
<tr>
<td>Review Net+ (Yang et al., 2016)</td>
<td>72.0</td>
<td>90.0</td>
<td>55.0</td>
<td>81.2</td>
<td>41.4</td>
<td>70.5</td>
<td>31.3</td>
<td>59.7</td>
</tr>
<tr>
<td>Adaptive+ (Lu et al., 2017)</td>
<td>74.8</td>
<td>92.0</td>
<td>58.4</td>
<td>84.5</td>
<td>44.4</td>
<td>74.4</td>
<td>33.6</td>
<td>63.7</td>
</tr>
<tr>
<td>SCST:Att2all++ (Rennie et al., 2017)</td>
<td>78.1</td>
<td>93.7</td>
<td>61.9</td>
<td>86.0</td>
<td>47.0</td>
<td>75.9</td>
<td>35.2</td>
<td>64.5</td>
</tr>
<tr>
<td>LSTM-A3+++ (Yao et al., 2017)</td>
<td>78.7</td>
<td>93.7</td>
<td>62.7</td>
<td>86.7</td>
<td>47.6</td>
<td>76.5</td>
<td>35.6</td>
<td>65.2</td>
</tr>
<tr>
<td>Up-Down+++ (Anderson et al., 2017)</td>
<td>80.2</td>
<td>95.2</td>
<td>64.1</td>
<td>88.8</td>
<td>49.1</td>
<td>79.4</td>
<td>36.9</td>
<td>68.5</td>
</tr>
<tr>
<td>Ours: Tok-Seq CIDEr</td>
<td>72.6</td>
<td>89.7</td>
<td>55.7</td>
<td>80.9</td>
<td>41.2</td>
<td>69.8</td>
<td>30.2</td>
<td>58.3</td>
</tr>
<tr>
<td>Ours: Tok-Seq CIDEr +</td>
<td>74.9</td>
<td>92.4</td>
<td>58.5</td>
<td>84.9</td>
<td>44.8</td>
<td>75.1</td>
<td>34.3</td>
<td>64.7</td>
</tr>
</tbody>
</table>

Table: MS-COCO ’s server evaluation. (+) for ensemble submissions, (†) for submissions with CIDEr optimization and (◦) for models using additional data.


K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhutdinov, R. Zemel, and Y. Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *ICML*.
