Token-level and sequence-level loss smoothing for RNN language models
Maha Elbayad, Laurent Besacier, Jakob Verbeek

To cite this version:

HAL Id: hal-01790879
https://inria.hal.science/hal-01790879
Submitted on 14 May 2018

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
Abstract

Despite the effectiveness of recurrent neural network language models, their maximum likelihood estimation suffers from two limitations. It treats all sentences that do not match the ground truth as equally poor, ignoring the structure of the output space. Second, it suffers from “exposure bias”: during training tokens are predicted given ground-truth sequences, while at test time prediction is conditioned on generated output sequences. To overcome these limitations we build upon the recent reward augmented maximum likelihood approach i.e. sequence-level smoothing that encourages the model to predict sentences close to the ground truth according to a given performance metric. We extend this approach to token-level loss smoothing, and propose improvements to the sequence-level smoothing approach. Our experiments on two different tasks, image captioning and machine translation, show that token-level and sequence-level loss smoothing are complementary, and significantly improve results.

1 Introduction

Recurrent neural networks (RNNs) have recently proven to be very effective sequence modeling tools, and are now state of the art for tasks such as machine translation (Cho et al., 2014; Sutskever et al., 2014; Bahdanau et al., 2015), image captioning (Kiros et al., 2014; Vinyals et al., 2015; Anderson et al., 2017) and automatic speech recognition (Chorowski et al., 2015; Chiu et al., 2017).

The basic principle of RNNs is to iteratively compute a vectorial sequence representation, by applying at each time-step the same trainable function to compute the new network state from the previous state and the last symbol in the sequence. These models are typically trained by maximizing the likelihood of the target sentence given an encoded source (text, image, speech).

Maximum likelihood estimation (MLE), however, has two main limitations. First, the training signal only differentiates the ground-truth target output from all other outputs. It treats all other output sequences as equally incorrect, regardless of their semantic proximity from the ground-truth target. While such a “zero-one” loss is probably acceptable for coarse grained classification of images, e.g. across a limited number of basic object categories (Everingham et al., 2010) it becomes problematic as the output space becomes larger and some of its elements become semantically similar to each other. This is in particular the case for tasks that involve natural language generation (captioning, translation, speech recognition) where the number of possible outputs is practically unbounded. For natural language generation tasks, evaluation measures typically do take into account structural similarity, e.g. based on n-grams, but such structural information is not reflected in the MLE criterion. The second limitation of MLE is that training is based on predicting the next token given the input and preceding ground-truth output tokens, while at test time the model predicts conditioned on the input and the so-far generated output sequence. Given the exponentially large output space of natural language sentences, it is not obvious that the learned RNNs generalize well beyond the relatively sparse distribution of ground-truth sequences used during MLE optimization. This phenomenon is known as “exposure bias” (Ranzato et al., 2016; Bengio et al., 2015).

MLE minimizes the KL divergence between a target Dirac distribution on the ground-truth sentence(s) and the model’s distribution. In this pa-
per, we build upon the “loss smoothing” approach by Norouzi et al. (2016), which smooths the Dirac target distribution over similar sentences, increasing the support of the training data in the output space. We make the following main contributions:

- We propose a token-level loss smoothing approach, using word-embeddings, to achieve smoothing among semantically similar terms, and we introduce a special procedure to promote rare tokens.
- For sequence-level smoothing, we propose to use restricted token replacement vocabularies, and a “lazy evaluation” method that significantly speeds up training.
- We experimentally validate our approach on the MSCOCO image captioning task and the WMT’14 English to French machine translation task, showing that on both tasks combining token-level and sequence-level loss smoothing improves results significantly over maximum likelihood baselines.

In the remainder of the paper, we review the existing methods to improve RNN training in Section 2. Then, we present our token-level and sequence-level approaches in Section 3. Experimental evaluation results based on image captioning and machine translation tasks are laid out in Section 4.

2 Related work

Previous work aiming to improve the generalization performance of RNNs can be roughly divided into three categories: those based on regularization, data augmentation, and alternatives to maximum likelihood estimation.

Regularization techniques are used to increase the smoothness of the function learned by the network, e.g. by imposing an $\ell_2$ penalty on the network weights, also known as “weight decay”. More recent approaches mask network activations during training, as in dropout (Srivastava et al., 2014) and its variants adapted to recurrent models (Pham et al., 2014; Krueger et al., 2017). Instead of masking, batch-normalization (Ioffe and Szegedy, 2015) rescales the network activations to avoid saturating the network’s non-linearities. Instead of regularizing the network parameters or activations, it is also possible to directly regularize based on the entropy of the output distribution (Pereyra et al., 2017).

Data augmentation techniques improve the robustness of the learned models by applying transformations that might be encountered at test time to the training data. In computer vision, this is common practice, and implemented by, e.g., scaling, cropping, and rotating training images (LeCun et al., 1998; Krizhevsky et al., 2012; Paulin et al., 2014). In natural language processing, examples of data augmentation include input noising by randomly dropping some input tokens (Iyyer et al., 2015; Bowman et al., 2015; Kumar et al., 2016), and randomly replacing words with substitutes sampled from the model (Bengio et al., 2015). Xie et al. (2017) introduced data augmentation schemes for RNN language models that leverage n-gram statistics in order to mimic Kneser-Ney smoothing of n-grams models. In the context of machine translation, Fadaee et al. (2017) modify sentences by replacing words with rare ones when this is plausible according to a pre-trained language model, and substitutes its equivalent in the target sentence using automatic word alignments. This approach, however, relies on the availability of additional monolingual data for language model training.

The de facto standard way to train RNN language models is maximum likelihood estimation (MLE) (Cho et al., 2014; Sutskever et al., 2014; Bahdanau et al., 2015). The sequential factorization of the sequence likelihood generates an additive structure in the loss, with one term corresponding to the prediction of each output token given the input and the preceding ground-truth output tokens. In order to directly optimize for sequence-level structured loss functions, such as measures based on n-grams like BLEU or CIDER, Ranzato et al. (2016) use reinforcement learning techniques that optimize the expectation of a sequence-level reward. In order to avoid early convergence to poor local optima, they pre-train the model using MLE.

Leblond et al. (2018) build on the learning to search approach to structured prediction (Daumé III et al., 2009; Chang et al., 2015) and adapts it to RNN training. The model generates candidate sequences at each time-step using all possible tokens, and scores these at sequence-level to derive a training signal for each time step. This leads to an approach that is structurally close to MLE, but computationally expensive. Norouzi et al. (2016) introduce a reward augmented maximum likelihood (RAML) approach, that incorpo-
rates a notion of sequence-level reward without facing the difficulties of reinforcement learning. They define a target distribution over output sentences using a soft-max over the reward over all possible outputs. Then, they minimize the KL divergence between the target distribution and the model's output distribution. Training with a general reward distribution is similar to MLE training, except that we use multiple sentences sampled from the target distribution instead of only the ground-truth sentences.

In our work, we build upon the work of Norouzi et al. (2016) by proposing improvements to sequence-level smoothing, and extending it to token-level smoothing. Our token-level smoothing approach is related to the label smoothing approach of Szegedy et al. (2016) for image classification. Instead of maximizing the probability of the correct class, they train the model to predict the correct class with a large probability and all other classes with a small uniform probability. This regularizes the model by preventing overconfident predictions. In natural language generation with large vocabularies, preventing such “narrow” over-confident distributions is imperative, since for many tokens there are nearly interchangeable alternatives.

3 Loss smoothing for RNN training

We briefly recall standard recurrent neural network training, before presenting sequence-level and token-level loss smoothing below.

3.1 Maximum likelihood RNN training

We are interested in modeling the conditional probability of a sequence \( y = (y_1, \ldots, y_T) \) given a conditioning observation \( x \),

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

where \( y_{<t} = (y_1, \ldots, y_{t-1}) \), the model parameters are given by \( \theta \), and \( x \) is a source sentence or an image in the contexts of machine translation and image captioning, respectively.

In a recurrent neural network, the sequence \( y \) is predicted based on a sequence of states \( h_t \),

\[
p_\theta(y_t|x, y_{<t}) = p_\theta(y_t|h_t), \tag{2}
\]

where the RNN state is computed recursively as

\[
h_t = \begin{cases} f_\theta(h_{t-1}, y_{t-1}, x) & \text{for } t \in \{1, \ldots, T\}, \\ g_\theta(x) & \text{for } t = 0. \end{cases} \tag{3}
\]

The input is encoded by \( g_\theta \) and used to initialize the state sequence, and \( f_\theta \) is a non-linear function that updates the state given the previous state \( h_{t-1} \), the last output token \( y_{t-1} \), and possibly the input \( x \). The state update function can take different forms, the ones including gating mechanisms such as LSTMs (Hochreiter and Schmidhuber, 1997) and GRUs (Chung et al., 2014) are particularly effective to model long sequences.

In standard teacher-forced training, the hidden states will be computed by forwarding the ground truth sequence \( y^* \) i.e. in Eq. (3), the RNN has access to the true previous token \( y_{t-1}^* \). In this case we will note the hidden states \( h_t^* \).

Given a ground-truth target sequence \( y^* \), maximum likelihood estimation (MLE) of the network parameters \( \theta \) amounts to minimizing the loss

\[
\ell_{\text{MLE}}(y^*, x) = -\ln p_\theta(y^*|x) \tag{4}
\]

\[
= -\sum_{t=1}^{T} \ln p_\theta(y_t^*|h_t^*). \tag{5}
\]

The loss can equivalently be expressed as the KL-divergence between a Dirac centered on the target output (with \( \delta_\alpha(x) = 1 \) at \( x = \alpha \) and 0 otherwise) and the model distribution, either at the sequence-level or at the token-level:

\[
\ell_{\text{MLE}}(y^*, x) = D_{\text{KL}}(\delta_{y^*} || p_\theta(y|x)) \tag{6}
\]

\[
= \sum_{t=1}^{T} D_{\text{KL}}(\delta_{y_t^*} || p_\theta(y_t|h_t^*)). \tag{7}
\]

Loss smoothing approaches considered in this paper consist in replacing the Dirac on the ground-truth sequence with distributions with larger support. These distributions can be designed in such a manner that they reflect which deviations from ground-truth predictions are preferred over others.

3.2 Sequence-level loss smoothing

The reward augmented maximum likelihood approach of Norouzi et al. (2016) consists in replacing the sequence-level Dirac \( \delta_{y^*} \) in Eq. (6) with a distribution

\[
r(y|y^*) \propto \exp r(y, y^*)/\tau, \tag{8}
\]

where \( r(y, y^*) \) is a “reward” function that measures the quality of sequence \( y \) w.r.t. \( y^* \), e.g. metrics used for evaluation of natural language processing tasks can be used, such as BLEU (Papineni et al., 2002) or CIDER (Vedantam et al., 2015).

The input is encoded by \( g_\theta \) and used to initialize the state sequence, and \( f_\theta \) is a non-linear function that updates the state given the previous state \( h_{t-1} \), the last output token \( y_{t-1} \), and possibly the input \( x \). The state update function can take different forms, the ones including gating mechanisms such as LSTMs (Hochreiter and Schmidhuber, 1997) and GRUs (Chung et al., 2014) are particularly effective to model long sequences.

In standard teacher-forced training, the hidden states will be computed by forwarding the ground truth sequence \( y^* \) i.e. in Eq. (3), the RNN has access to the true previous token \( y_{t-1}^* \). In this case we will note the hidden states \( h_t^* \).

Given a ground-truth target sequence \( y^* \), maximum likelihood estimation (MLE) of the network parameters \( \theta \) amounts to minimizing the loss

\[
\ell_{\text{MLE}}(y^*, x) = -\ln p_\theta(y^*|x) \tag{4}
\]

\[
= -\sum_{t=1}^{T} \ln p_\theta(y_t^*|h_t^*). \tag{5}
\]

The loss can equivalently be expressed as the KL-divergence between a Dirac centered on the target output (with \( \delta_\alpha(x) = 1 \) at \( x = \alpha \) and 0 otherwise) and the model distribution, either at the sequence-level or at the token-level:

\[
\ell_{\text{MLE}}(y^*, x) = D_{\text{KL}}(\delta_{y^*} || p_\theta(y|x)) \tag{6}
\]

\[
= \sum_{t=1}^{T} D_{\text{KL}}(\delta_{y_t^*} || p_\theta(y_t|h_t^*)). \tag{7}
\]

Loss smoothing approaches considered in this paper consist in replacing the Dirac on the ground-truth sequence with distributions with larger support. These distributions can be designed in such a manner that they reflect which deviations from ground-truth predictions are preferred over others.

3.2 Sequence-level loss smoothing

The reward augmented maximum likelihood approach of Norouzi et al. (2016) consists in replacing the sequence-level Dirac \( \delta_{y^*} \) in Eq. (6) with a distribution

\[
r(y|y^*) \propto \exp r(y, y^*)/\tau, \tag{8}
\]

where \( r(y, y^*) \) is a “reward” function that measures the quality of sequence \( y \) w.r.t. \( y^* \), e.g. metrics used for evaluation of natural language processing tasks can be used, such as BLEU (Papineni et al., 2002) or CIDER (Vedantam et al., 2015).
The temperature parameter $\tau$ controls the concentration of the distribution around $y^*$. When $m > 1$ ground-truth sequences are paired with the same input $x$, the reward function can be adapted to fit this setting and be defined as $r(y, \{y^{(1)}, \ldots, y^{(m)}\})$. The sequence-level smoothed loss function is then given by

$$\ell_{\text{seq}}(y^*, x) = D_{\text{KL}}(f(y|y^*)||p_{\theta}(y|x)) = H(r(y|y^*)) - E_r[\ln p_{\theta}(y|x)],$$

where the entropy term $H(r(y|y^*))$ does not depend on the model parameters $\theta$.

In general, expectation in Eq. (9) is intractable due to the exponentially large output space, and replaced with a Monte-Carlo approximation:

$$E_r[-\ln p_{\theta}(y|x)] \approx -\sum_{l=1}^{L} \ln p_{\theta}(y^l|x).$$

### Stratified sampling
Norouzi et al. (2016) show that when using the Hamming or edit distance as a reward, we can sample directly from $r(y|y^*)$ using a stratified sampling approach. In this case sampling proceeds in three stages. (i) Sample a distance $d$ from $\{0, \ldots, T\}$ from a prior distribution on $d$. (ii) Uniformly select $d$ positions in the sequence to be modified. (iii) Sample the $d$ substitutions uniformly from the token vocabulary.

Details on the construction of the prior distribution on $d$ for a reward based on the Hamming distance can be found in Appendix A.

### Importance sampling
For a reward based on BLEU or CIDEr, we cannot directly sample from $r(y|y^*)$ since the normalizing constant, or “partition function”, of the distribution is intractable to compute. In this case we can resort to importance sampling. We first sample $L$ sequences $y^l$ from a tractable proposal distribution $q(y|y^*)$. We then compute the importance weights

$$\omega_l \approx \frac{r(y^l|y^*)/q(y^l|y^*)}{\sum_{k=1}^{L} r(y^k|y^*)/q(y^k|y^*)},$$

where $r(y^k|y^*)$ is the un-normalized reward distribution in Eq. (8). We finally approximate the expectation by reweighing the samples in the Monte Carlo approximation as

$$E_r[-\ln p_{\theta}(y|x)] \approx -\sum_{l=1}^{L} \omega_l \ln p_{\theta}(y^l|x).$$

In our experiments we use a proposal distribution based on the Hamming distance, which allows for tractable stratified sampling, and generates sentences that do not stray away from the ground truth.

We propose two modifications to the sequence-level loss smoothing of Norouzi et al. (2016): sampling to a restricted vocabulary (described in the following paragraph) and lazy sequence-level smoothing (described in section 3.4).

**Restricted vocabulary sampling.** In the stratified sampling method for Hamming and edit distance rewards, instead of drawing from the large vocabulary $V$, containing typically in the order of $10^4$ words or more, we can restrict ourselves to a smaller subset $V_{\text{sub}}$ more adapted to our task. We considered three different possibilities for $V_{\text{sub}}$:

- $V$: the full vocabulary from which we sample uniformly (default), or draw from our token-level smoothing distribution defined below in Eq. (13).
- $V_{\text{refs}}$: uniformly sample from the set of tokens that appear in the ground-truth sentence(s) associated with the current input.
- $V_{\text{batch}}$: uniformly sample from the tokens that appear in the ground-truth sentences across all inputs that appear in a given training mini-batch.

Uniformly sampling from $V_{\text{batch}}$ has the effect of boosting the frequencies of words that appear in many reference sentences, and thus approximates to some extent sampling substitutions from the uni-gram statistics of the training set.

### 3.3 Token-level loss smoothing
While the sequence-level smoothing can be directly based on performance measures of interest such as BLEU or CIDEr, the support of the smoothed distribution is limited to the number of samples drawn during training. We propose smoothing the token-level Diracs $\delta_{y_t^*}$ in Eq. (7) to increase its support to similar tokens. Since we apply smoothing to each of the tokens independently, this approach implicitly increases the support to an exponential number of sequences, unlike the sequence-level smoothing approach. This comes at the price, however, of a naive token-level independence assumption in the smoothing.

We define the smoothed token-level distribution, similar as the sequence-level one, as a softmax over a token-level “reward” function,

$$r(y_t|y_t^*) \propto \exp r(y_t, y_t^*)/\tau,$$
where $\tau$ is again a temperature parameter. As a token-level reward $r(y_t, y_t^*)$ we use the cosine similarity between $y_t$ and $y_t^*$ in a semantic word-embedding space. In our experiments we use GloVe (Pennington et al., 2014); preliminary experiments with word2vec (Mikolov et al., 2013) yielded somewhat worse results.

### Promoting rare tokens

We can further improve the token-level smoothing by promoting rare tokens. To do so, we penalize frequent tokens when smoothing over the vocabulary, by subtracting $\beta \text{freq}(y_t)$ from the reward, where $\text{freq}()$ denotes the term frequency and $\beta$ is a non-negative weight. This modification encourages frequent tokens into considering the rare ones. We experimentally found that it is also beneficial for rare tokens to boost frequent ones, as they tend to have mostly rare tokens as neighbors in the word-embedding space. With this in mind, we define a new token-level reward as:

$$
r^\text{freq}(y_t, y_t^*) = r(y_t, y_t^*) - \beta \min \left( \frac{\text{freq}(y_t)}{\text{freq}(y_t^*)}, \frac{\text{freq}(y_t^*)}{\text{freq}(y_t)} \right), \quad (14)$$

where the penalty term is strongest if both tokens have similar frequencies.

### 3.4 Combining losses

In both loss smoothing methods presented above, the temperature parameter $\tau$ controls the concentration of the distribution. As $\tau$ gets smaller the distribution peaks around the ground-truth, while for large $\tau$ the uniform distribution is approached. We can, however, not separately control the spread of the distribution and the mass reserved for the ground-truth output. We therefore introduce a second parameter $\alpha \in [0, 1]$ to interpolate between the Dirac on the ground-truth and the smooth distribution. Using $\bar{\alpha} = 1 - \alpha$, the sequence-level and token-level loss functions are then defined as

$$
\ell_{\text{Seq}}^\alpha(y^*, x) = \alpha \ell_{\text{Seq}}(y^*, x) + \bar{\alpha} \ell_{\text{MLE}}(y^*, x) \quad (15)
$$

$$
\ell_{\text{Tok}}^\alpha(y^*, x) = \alpha \ell_{\text{Tok}}(y^*, x) + \bar{\alpha} \ell_{\text{MLE}}(y^*, x) \quad (16)
$$

To benefit from both sequence-level and token-level loss smoothing, we also combine them by applying token-level smoothing to the different sequences sampled for the sequence-level smoothing. We introduce two mixing parameters $\alpha_1$ and $\alpha_2$. The first controls to what extent sequence-level smoothing is used, while the second controls to what extent token-level smoothing is used. The combined loss is defined as

$$
\ell_{\text{Seq Tok}}^{\alpha_1, \alpha_2}(y^*, x, r) = \alpha_1 \ell_{\text{Seq}}(y^*, x) + \bar{\alpha}_1 \ell_{\text{Tok}}(y^*, x) = \\
\alpha_1 \mathbb{E}_r[\alpha_2 \ell_{\text{Tok}}(y^*, x) + \bar{\alpha}_2 \ell_{\text{MLE}}(y^*, x)] + \\
\bar{\alpha}_1 (\alpha_2 \ell_{\text{Tok}}(y^*, x) + \bar{\alpha}_2 \ell_{\text{MLE}}(y^*, x)). \quad (17)
$$

In our experiments, we use held out validation data to set mixing and temperature parameters.

#### Algorithm 1 Sequence-level smoothing algorithm

**Input:** $x, y^*$

**Output:** $\ell_{\text{Seq}}^\alpha(x, y^*)$

Encode $x$ to initialize the RNN

Forward $y^*$ in the RNN to compute the hidden states $h_t^*$

Compute the MLE loss $\ell_{\text{MLE}}(y^*, x)$

for $l \in \{1, \ldots, L\}$ do

Sample $y^*_l \sim r(\cdot)$

if Lazy then

Compute $\ell(y^*_l, x) = -\sum \log p_\theta(y^*_l|h_t^*)$

else

Forward $y^*_l$ in the RNN to get its hidden states $h_t^*$

Compute $\ell(y^*_l, x) = \ell_{\text{MLE}}(y^*_l, x)$

end if

end for

$$
\ell_{\text{Seq}}^\alpha(x, y^*) = \bar{\alpha} \ell_{\text{MLE}}(y^*, x) + \frac{\alpha}{T} \sum \ell(y^*_l, x) \quad (19)
$$

#### Lazy sequence smoothing

Although sequence-level smoothing is computationally efficient compared to reinforcement learning approaches (Ranzato et al., 2016; Rennie et al., 2017), it is slower compared to MLE. In particular, we need to forward each of the samples $y^*$ through the RNN in teacher-forcing mode so as to compute its hidden states $h_t^*$, which are used to compute the sequence MLE loss as

$$
\ell_{\text{MLE}}(y^*, x) = -\sum_{t=1}^T \ln p_\theta(y_t^*|h_t^*). \quad (18)
$$

To speed up training, and since we already forward the ground truth sequence in the RNN to evaluate the MLE part of $\ell_{\text{Seq}}^\alpha(y^*, x)$, we propose to use the same hidden states $h_t^*$ to compute both the MLE and the sequence-level smoothed loss. In this case:

$$
\ell_{\text{lazy}}(y^*, x) = -\sum_{t=1}^T \ln p_\theta(y_t^*|h_t^*) \quad (19)
$$

In this manner, we only have a single instead of $L + 1$ forwards-passes in the RNN. We provide the pseudo-code for training in Algorithm 1.
4 Experimental evaluation

In this section, we compare sequence prediction models trained with maximum likelihood (MLE) with our token and sequence-level loss smoothing on two different tasks: image captioning and machine translation.

4.1 Image captioning

4.1.1 Experimental setup.

We use the MS-COCO dataset (Lin et al., 2014), which consists of 82k training images each annotated with five captions. We use the standard splits of Karpathy and Li (2015), with 5k images for validation, and 5k for test. The test set results are generated via beam search (beam size 3) and are evaluated with the MS-COCO captioning evaluation tool. We report CIDE\(R\) and BLEU scores on this internal test set. We also report results obtained on the official MS-COCO server that additionally measures METEOR (Denkowski and Lavie, 2014) and ROUGE-L (Lin, 2004).

We experiment with both non-attentive LSTMs (Vinyals et al., 2015) and the ResNet baseline of the state-of-the-art top-down attention (Anderson et al., 2017).

The MS-COCO vocabulary consists of 9,800 words that occur at least 5 times in the training set. Additional details and hyperparameters can be found in Appendix B.1.

4.1.2 Results and discussion

Restricted vocabulary sampling In this section, we evaluate the impact of the vocabulary subset from which we sample the modified sentences for sequence-level smoothing. We experiment with two rewards: CIDE\(R\), which scores w.r.t. all five available reference sentences, and Hamming distance reward taking only a single reference into account. For each reward we train our (Seq) models with each of the three subsets detailed previously in Section 3.2, Restricted vocabulary sampling.

From the results in Table 1 we note that for the inattentive models, sampling from \(V_{\text{refs}}\) or \(V_{\text{batch}}\) has a better performance than sampling from the full vocabulary on all metrics. In fact, using these subsets introduces a useful bias to the model and improves performance. This improvement is most notable using the CIDE\(R\) reward that scores candidate sequences w.r.t. to multiple references, which stabilizes the scoring of the candidates.

With an attentive decoder, no matter the reward, re-sampling sentences with words from \(V_{\text{ref}}\) rather than the full vocabulary \(V\) is better for both reward functions, and all metrics. Additional experimental results, presented in Appendix B.2, obtained with a BLEU-4 reward, in its single and
Table 2: MS-COCO’s server evaluation. (+) for ensemble submissions, (†) for submissions with CIDEr optimization and (◦) for models using additional data.

<table>
<thead>
<tr>
<th>Model</th>
<th>BLEU-1 c5</th>
<th>BLEU-2 c5</th>
<th>BLEU-3 c5</th>
<th>BLEU-4 c5</th>
<th>METEOR c5</th>
<th>ROUGE-L c5</th>
<th>CIDEr c5</th>
<th>SPICE c5</th>
</tr>
</thead>
<tbody>
<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>58.2</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>33.2</td>
<td>56.4</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>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>

*Lazy training. From the results of Table 1, we see that lazy sequence-level smoothing is competitive with exact non-lazy sequence-level smoothing, while requiring roughly equivalent training time as MLE. We provide detailed timing results in Appendix B.3.

For reference, we include in Table 1 baseline results obtained using MLE, and our implementation of MLE with entropy regularization (MLE+γH) (Pereyra et al., 2017), as well as the RAML approach of Norouzi et al. (2016) which corresponds to sequence-level smoothing based on the Hamming reward and sampling replacements from the full vocabulary (Seq, Hamming, V)

We observe that entropy smoothing is not able to improve performance much over MLE for the model without attention, and even deteriorates for the attention model. We improve upon RAML by choosing an adequate subset of vocabulary for substitutions.

We also report the performances of token-level smoothing, where the promotion of rare tokens boosted the scores in both attentive and non-attentive models.

For sequence-level smoothing, choosing a task-relevant reward with importance sampling yielded better results than plain Hamming distance.

Moreover, we used the two smoothing schemes (Tok-Seq) and achieved the best results with CIDEr as a reward for sequence-level smoothing combined with a token-level smoothing that promotes rare tokens improving CIDEr from 93.59 (MLE) to 99.92 for the model without attention, and improving from 101.63 to 103.81 with attention.

Qualitative results. In Figure 1 we showcase captions obtained with MLE and our three variants of smoothing i.e token-level (Tok), sequence-level (Seq) and the combination (Tok-Seq). We note that the sequence-level smoothing tends to generate lengthy captions overall, which is maintained in the combination. On the other hand, the token-level smoothing allows for a better recognition of objects in the image that stems from the robust training of the classifier e.g. the ’cement block’ in the top right image or the carrots in the bottom right. More examples are available in Appendix B.4

Comparison to the state of the art. We compare our model to state-of-the-art systems on the MS-COCO evaluation server in Table 2. We submitted a single model (Tok-Seq, CIDEr, Vrefs, Vrefs) as well as an ensemble of five models with different initializations trained on the training set plus 35k images from the dev set (a total of 117k images) to the MS-COCO server. The three best results on the server (Rennie et al., 2017; Yao et al., 2017; Anderson et al., 2017) are trained in two stages where they first train using MLE, before switching to policy gradient methods based on CIDEr. Anderson et al. (2017) reported an increase of 5.8% of CIDEr on the test split after the CIDEr optimization. Moreover, Yao et al. (2017) uses additional information about image regions to train the attributes classifiers, while Anderson et al. (2017) pre-trains its bottom-up attention model on the Visual Genome dataset (Krishna et al., 2017). Lu et al. (2017); Yao et al. (2017) use the same CNN encoder as ours (ResNet-152), (Vinyals et al., 2015; Yang et al., 2016) use Inception-v3 (Szegedy et al., 2016) for image encoding and Rennie et al. (2017); Anderson et al.
We present our results in Table 3. On both benchmarks, we improve on both MLE and RAML approach of Norouzi et al. (2016) (Seq, Hamming, $\mathcal{V}$): using the smaller batch-vocabulary for replacement improves results, and using importance sampling based on BLEU-4 further boosts results. In this case, unlike in the captioning experiment, token-level smoothing brings smaller improvements. The combination of both smoothing approaches gives best results, similar to what was observed for image captioning, improving the MLE BLEU-4 from 30.03 to 31.39 on WMT’14 and from 27.55 to 28.74 on IWSLT’14. The outputs of our best model are compared to the MLE in some examples showcased in Appendix C.

5 Conclusion

We investigated the use of loss smoothing approaches to improve over maximum likelihood estimation of RNN language models. We generalized the sequence-level smoothing RAML approach of Norouzi et al. (2016) to the token-level by smoothing the ground-truth target across semantically similar tokens. For the sequence-level, which is computationally expensive, we introduced an efficient “lazy” evaluation scheme, and introduced an improved re-sampling strategy. Experimental evaluation on image captioning and machine translation demonstrates the complementarity of sequence-level and token-level loss smoothing, improving over both the maximum likelihood and RAML.

Acknowledgment. This work has been partially supported by the grant ANR-16-CE23-0006 “Deep in France” and LabEx PERSYVAL-Lab (ANR-11-LABX-0025-01).
References


A. Karpathy and Fei-Fei Li. 2015. Deep visual-semantic alignments for generating image descriptions. In CVPR.


R. Kiros, R. Salakhutdinov, and R. Zemel. 2014. Multimodal neural language models. In ICML.


T. Mikolov, K. Chen, G. Corrado, and J. Dean. 2013. Efficient estimation of word representations in vector space. In ICLR.


Z. Xie, S. Wang, J. Li, D. Lévy, A. Nie, D. Jurafsky, and A. Ng. 2017. Data noising as smoothing in neural network language models. In ICLR.

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.

Z. Yang, Y. Yuan, Y. Wu, R. Salakhutdinov, and W. Cohen. 2016. Encode, review, and decode: Reviewer module for caption generation. In NIPS.


A Stratified sampling of the Hamming distance

Norouzi et al. (2016) detail the steps to drawing samples from the reward distribution based on the edit distance with stratified sampling. In this section we show how we sample from the Hamming distance (a special case of the edit distance) reward and how we handle large vocabularies to generate reasonable candidates. To draw from the Hamming distance reward, we proceed as follows:

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 a subset $V_{sub}$ of the vocabulary ($|V_{sub}| = V_{sub}$).

To sample a distance, we partition the set of sequences in $V_{sub}$ terms $V_{sub}^T$ with respect to their distance to the ground truth $y^*$:

$$S_d = \{y \in V_{sub}^T | d(y, y^*) = d\},$$
$$V_{sub}^T = \bigcup_d S_d,$$
$$\forall d, d': S_d \cap S_{d'} = \emptyset.$$

To each distance $d$ in $\{0, \ldots, T\}$ we assign the portion of rewards covered by $S_d$ in $V_{sub}^T$. Since all elements of $S_d$ are assigned the same reward $e^{-\frac{d}{\tau}}$; we need only to multiply it by the set’s size $|S_d|$. Counting the elements of $|S_d|$ is straight-forward: we choose $d$ elements to alter in $y^*$ ($\binom{T}{d}$ combinations) and at each position we have $(V_{sub} - 1)$ possibilities. The sampling distribution is obtained as:

$$p(d) = \frac{r(S_d)}{r(V_{sub}^T)}$$
$$\quad = \frac{\sum_{y \in S_d} r(y | y^*)}{\sum_{y \in V_{sub}^T} r(y | y^*)}.\quad (20)$$

Given that $\{S_d\}_d$ form a partition of $V_{sub}^T$,

$$\sum_{y \in V_{sub}^T} r(y | y^*) = \sum_d \sum_{y \in S_d} e^{-\frac{d}{\tau}}$$
$$\quad = \sum_d \binom{T}{d} (V_{sub} - 1)^d e^{-\frac{d}{\tau}}$$
$$\quad = (V_{sub} - 1)e^{-\frac{1}{\tau} + 1})^T,\quad (23)$$

we find:

$$p(d) = \binom{T}{d} \frac{(V_{sub} - 1)^d e^{-\frac{d}{\tau}}}{((V_{sub} - 1)e^{-\frac{1}{\tau} + 1})^T}.\quad (25)$$

B Captioning

B.1 Experimental setup

Out of vocabulary words are replaced by $<$UNK$>$ token and the captions longer than 16 words are truncated. As image encoding, we average-pool the features in the last convolutional layer of ResNet-152 (He et al., 2016) pre-trained on ImageNet. The 2048-dimensional image signature is further mapped to $\mathbb{R}^{512}$ to fit the word-embedding dimension, so it can be used as the first token fed to the RNN decoder. We use a single-layer RNN with $d = 512$ LSTM units.

For optimization, we use Adam (Kingma and Ba, 2015) with a batch size of 10 images, i.e. 50 sentences. We follow Lu et al. (2017); Pedersoli et al. (2017) and train in two stages: the first, optimizing
the language model alone with an initial learning rate of 5e-4 annealed by a factor of 0.6 every 3 epochs starting from the 5th one. We train for up to 20 epochs with early stopping if CIDE\textsubscript{R} score on the validation set does not improve. In the second stage, we optimize the language model conjointly with \textit{conv}_4 and \textit{conv}_5 (the last 39 building blocks of ResNet-152) of the CNN model. The initial learning rate is of 6e-5 and diminishes by a factor of 0.8 every 4 epochs. The same early-stopping strategy is applied. For the token-level reward, we use GloVe (Pennington et al., 2014) as our word embedding, which we train on the captions in the MS-COCO training set. In preliminary experiments using the publicly available 300-dimensional GloVe vectors trained on Wikipedia 2014 + Gigaword worsens the model’s results.

B.2 Restricted vocabulary sampling - supplementary results

In Table 4 we provide additional results for sequence-level smoothing, when using the BLEU-4 as reward function. We include results when computing BLEU-4 w.r.t. all reference sequences (like also done for CIDE\textsubscript{R}), and when computing w.r.t. a single randomly selected reference sentence (as is the case for Hamming). With the BLEU-4 reward, using \textit{\nu}_{refs} yields best results in all but a single case. This underlines the effectiveness of sampling replacement words that are relevant to the task in sequence-level smoothing.

<table>
<thead>
<tr>
<th>Reward</th>
<th>\textit{\nu}_{sub}</th>
<th>BLEU-1</th>
<th>BLEU-4</th>
<th>CIDE\textsubscript{R}</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dirac</td>
<td>70.63</td>
<td>30.14</td>
<td>93.59</td>
<td></td>
</tr>
<tr>
<td>Hamming</td>
<td>\nu</td>
<td>71.76</td>
<td>31.16</td>
<td>96.37</td>
</tr>
<tr>
<td>Hamming</td>
<td>\nu_{batch}</td>
<td>71.46</td>
<td>31.15</td>
<td>96.53</td>
</tr>
<tr>
<td>Hamming</td>
<td>\nu_{refs}</td>
<td>71.80</td>
<td>31.63</td>
<td>96.22</td>
</tr>
<tr>
<td>BLEU-4 single</td>
<td>\nu</td>
<td>71.30</td>
<td>31.11</td>
<td>96.11</td>
</tr>
<tr>
<td>BLEU-4 single</td>
<td>\nu_{batch}</td>
<td>71.13</td>
<td>30.84</td>
<td>94.74</td>
</tr>
<tr>
<td>BLEU-4 single</td>
<td>\nu_{refs}</td>
<td>71.78</td>
<td>31.63</td>
<td>97.29</td>
</tr>
<tr>
<td>BLEU-4</td>
<td>\nu</td>
<td>71.56</td>
<td>31.47</td>
<td>96.56</td>
</tr>
<tr>
<td>BLEU-4</td>
<td>\nu_{batch}</td>
<td>71.41</td>
<td>30.87</td>
<td>95.69</td>
</tr>
<tr>
<td>BLEU-4</td>
<td>\nu_{refs}</td>
<td>72.08</td>
<td>31.41</td>
<td>97.21</td>
</tr>
<tr>
<td>CIDE\textsubscript{R}</td>
<td>\nu</td>
<td>71.05</td>
<td>30.46</td>
<td>94.40</td>
</tr>
<tr>
<td>CIDE\textsubscript{R}</td>
<td>\nu_{batch}</td>
<td>71.51</td>
<td>31.17</td>
<td>95.78</td>
</tr>
<tr>
<td>CIDE\textsubscript{R}</td>
<td>\nu_{refs}</td>
<td>71.93</td>
<td>31.41</td>
<td>96.81</td>
</tr>
</tbody>
</table>

Table 4: Captioning performance on MSCOCO when training with sequence-level loss smoothing.

B.3 Training time

We report below (Table 5) the average wall time to process a single batch (10 images i.e. 50 captions) when training the RNN language model with fixed CNN (without attention) on a Titan X GPU. We can clearly see that the lazy training is faster compared to the standard sequence smoothing and that the token-level smoothing does not hinder the training speed.

<table>
<thead>
<tr>
<th>Loss</th>
<th>MLE</th>
<th>Tok</th>
<th>Seq</th>
<th>Seq lazy</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>
</tr>
<tr>
<td>\textit{\nu}_{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>
</tr>
</tbody>
</table>

Table 5: Average training time per batch for different losses
B.4 Additional examples

Ground truth: a man holding his dog and sitting on a wooden table
- a flower vase with a glass to it sitting on a table
- a small, gray dog sits holding a single tail in a cat bowl
- a single trip is seen in a small vase

Ground truth: a bunch of green bananas hanging from a tree
- a bunch of green bananas hanging from a tree
- a bunch of green bananas sitting on a table

Ground truth: a person standing in front of a street vendor, a woman behind the counter
- a man preparing food on a street cart stands by a building
- a man running a street stand in front of a yellow van
- a man and a woman stand at a pedestrian crossing
two women eat a burger next to a street vendor

Ground truth: a man is standing next to a bicycle
- a man on a bike with a man on the book
- a couple of men standing next to each other
- a woman standing next to a woman in front of a store

Ground truth: a colorful walk sign in the city on a post
- a sidewalk with signs and illuminated cross signs
- a pedestrian walking under a street sign
- traffic signal for walking on green street

Ground truth: there are two men standing on the beach with surfboards
- two surfers are standing on a beach carrying their surfboards
- a pair of surfers goes on a pedestrian walkway
- two surfers with surfboards on a bench

Ground truth: a display case filled with blueberries goods in front of a store
- a display case filled with different types of cloudberry
- a display case filled with lots of fruit and vegetables
- a display case filled with lots of flowers
- a display case filled with lots of green bananas

Ground truth: a low room with a sectional couch, a sofa, and a glass desk covered in paper
- a low room with a yellow chair and a black footrest
- a small dog is sitting on a floor
- a book with a cover on top of a table
- a coffee table on a sofa

Ground truth: a cat fixing a couch next to a computer wire while holding a red stuffed toy
- a black and white cat jumping in a chair holding a stuffed animal
- a cat sitting on a back pack with a stuffed animal
- a black and white cat sitting on a bed
- a black and white cat with a red stuffed toy

Ground truth: a man riding a motorcycle down a street
- a group of people riding motorcycles down a street
- a group of people riding motorcycles down a street

Ground truth: a man riding a bicycle down a street
- a group of people riding bicycles down a street
- a group of people riding bicycles down a street

Ground truth: a man riding a motorcycle down a street
- a group of people riding motorcycles down a street
- a group of people riding motorcycles down a street

Ground truth: a man standing in front of a street vendor, a woman behind the counter
- a man preparing food on a street cart stands by a building
- a man running a street stand in front of a yellow van
- a man and a woman stand at a pedestrian crossing
two women eat a burger next to a street vendor

Ground truth: a man is holding a bottle in a field
- a man in a hat holding a tennis racket
- a man in a hat holding a frisbee
- a man holding a frisbee in his hand

Ground truth: a woman looks apathetically as she prepares to cut her own hair with scissors.
- a woman that is holding her hair and a pair of scissors
- a lady cutting her own hair with a pair of huge scissors
- a woman in a salon with a pair of scissors
- a young woman is cutting her long hair

Ground truth: a black and white image of white umbrellas hanging below light
- the umbrellas are hanging from a lamppost
- a bunch of umbrellas are hanging from the ceiling
- there are many lights hanging from the ceiling

Ground truth: the man wearing a red tie is standing near an animal's plastic house
- a young girl wearing a red tie is adjusting her glasses
- the man is in the red tie is putting on his glasses
- a person with glasses and a tie in a room
- a woman wearing glasses and a tie

Ground truth: a bathroom with a red wall and toilet in front of a toilet
- this bathroom has red walls and a toilet
- a small toilet with a toilet paper over it
- a toilet paper over a toilet
- a bathroom with a toilet and a sign

Ground truth: a man holding a bicycle next to a car
- a man in a shirt holding a bicycle
- a man in a shirt holding a bicycle
- a man in a shirt holding a bicycle
- a man in a shirt holding a bicycle

Ground truth: a man riding a bicycle down a street
- a group of people riding motorcycles down a street
- a group of people riding motorcycles down a street
- a group of people riding motorcycles down a street

Ground truth: a man riding a motorcycle down a street
- a group of people riding motorcycles down a street
- a group of people riding motorcycles down a street
- a group of people riding motorcycles down a street

Ground truth: a man sitting next to a car
- a man sitting next to a car
- a man sitting next to a car
- a man sitting next to a car
- a man sitting next to a car

Ground truth: a man sitting next to a car
- a man sitting next to a car
- a man sitting next to a car
- a man sitting next to a car
- a man sitting next to a car

Ground truth: a man holding his dog and sitting on a wooden table
- a flower vase with a glass to it sitting on a table
- a small, gray dog sits holding a single tail in a cat bowl
- a single trip is seen in a small vase

Ground truth: a bunch of green bananas hanging from a tree
- a bunch of green bananas hanging from a tree
- a bunch of green bananas sitting on a table

Ground truth: a man standing in front of a street vendor, a woman behind the counter
- a man preparing food on a street cart stands by a building
- a man running a street stand in front of a yellow van
- a man and a woman stand at a pedestrian crossing
two women eat a burger next to a street vendor

Ground truth: a man is standing next to a bicycle
- a man on a bike with a man on the book
- a couple of men standing next to each other
- a woman standing next to a woman in front of a store

Ground truth: a colorful walk sign in the city on a post
- a sidewalk with signs and illuminated cross signs
- a pedestrian walking under a street sign
- traffic signal for walking on green street

Ground truth: there are two men standing on the beach with surfboards
- two surfers are standing on a beach carrying their surfboards
- a pair of surfers goes on a pedestrian walkway
- two surfers with surfboards on a bench

Ground truth: a display case filled with blueberries goods in front of a store
- a display case filled with different types of cloudberry
- a display case filled with lots of fruit and vegetables
- a display case filled with lots of flowers
- a display case filled with lots of green bananas

Ground truth: the blossoms of a plant hanging near large green leaves
- small bananas growing on top of the banana tree
- a small bouquet tree filled with little miniature bananas
- a large bunch of bananas growing on it
- a tree with tiny bananas growing on it

Ground truth: a display case filled with different types of cloudberry
- a display case filled with lots of fruit and vegetables
- a display case filled with lots of flowers
- a display case filled with lots of flowers

Ground truth: a living room with a sectional couch, a sofa, and a glass desk covered in paper
- a low room with a yellow chair and a black footrest
- a small dog is sitting on a floor
- a book with a cover on top of a table
- a coffee table on a sofa

Ground truth: a cat fixing a couch next to a computer wire while holding a red stuffed toy
- a black and white cat jumping in a chair holding a stuffed animal
- a cat sitting on a back pack with a stuffed animal
- a black and white cat sitting on a bed
- a black and white cat with a red stuffed toy

Ground truth: a man sitting next to a car
- a man sitting next to a car
- a man sitting next to a car
- a man sitting next to a car
- a man sitting next to a car

Ground truth: a man holding a bicycle next to a car
- a man in a shirt holding a bicycle
- a man in a shirt holding a bicycle
- a man in a shirt holding a bicycle
- a man in a shirt holding a bicycle

Ground truth: a man riding a motorcycle down a street
- a group of people riding motorcycles down a street
- a group of people riding motorcycles down a street
- a group of people riding motorcycles down a street

Ground truth: a man sitting in front of a street vendor, a woman behind the counter
- a man preparing food on a street cart stands by a building
- a man running a street stand in front of a yellow van
- a man and a woman stand at a pedestrian crossing
two women eat a burger next to a street vendor

Ground truth: a man is standing next to a bicycle
- a man on a bike with a man on the book
- a couple of men standing next to each other
- a woman standing next to a woman in front of a store

Ground truth: a colorful walk sign in the city on a post
- a sidewalk with signs and illuminated cross signs
- a pedestrian walking under a street sign
- traffic signal for walking on green street

Ground truth: there are two men standing on the beach with surfboards
- two surfers are standing on a beach carrying their surfboards
- a pair of surfers goes on a pedestrian walkway
- two surfers with surfboards on a bench

Ground truth: a display case filled with blueberries goods in front of a store
- a display case filled with different types of cloudberry
- a display case filled with lots of fruit and vegetables
- a display case filled with lots of flowers
- a display case filled with lots of green bananas

Ground truth: the blossoms of a plant hanging near large green leaves
- small bananas growing on top of the banana tree
- a small bouquet tree filled with little miniature bananas
- a large bunch of bananas growing on it
- a tree with tiny bananas growing on it

Ground truth: a close-up of a bunch of green bananas
- a close-up of a bunch of green bananas
- a close-up of a bunch of green bananas
- a close-up of a bunch of green bananas
- a close-up of a bunch of green bananas

Ground truth: a living room with a sectional couch, a sofa, and a glass desk covered in paper
- a low room with a yellow chair and a black footrest
- a small dog is sitting on a floor
- a book with a cover on top of a table
- a coffee table on a sofa

Ground truth: a cat fixing a couch next to a computer wire while holding a red stuffed toy
- a black and white cat jumping in a chair holding a stuffed animal
- a cat sitting on a back pack with a stuffed animal
- a black and white cat sitting on a bed
- a black and white cat with a red stuffed toy
C Neural machine translation

C.1 Experimental setups

**WMT14 English-to-French**  We use the same experimental setting as Bahdanau et al. (2015): 12M paired sentences are used for training, 6,003 pairs for validation (news-test-2012 and news-test-2013) and 3,003 test pairs (news-test-2014). After tokenization, 30k most frequent tokens are selected for the model’s vocabulary. We use an attentive encoder-decoder with a 2-layers bi-directional encoder of dimension $d = 2000$ and a single-layer decoder of dimension $d = 2000$ as well. We use batches of size 80 and train for 3 epochs with Adam (Kingma and Ba, 2015) starting with a learning rate of 2e-4. To generate translations we use beam search of size five.

**IWSLT14 German-to-English**  We use the same settings as (Ranzato et al., 2016); the training set consists of 153k sentence pairs and 7k pairs are assigned to the validation and test sets. After tokenization and lower-casing, we remove sentences longer than 50 tokens. The English vocabulary has 22,822 words while the German has 32,009 words. We use an attentive encoder-decoder with a single bi-directional encoder and decoder of dimension $d = 128$. To generate translations we use beam search of size five. We use batches of size 32 and train for 40 epochs with Adam (Kingma and Ba, 2015) starting with a learning rate of 1e-3.

C.2 Examples

<table>
<thead>
<tr>
<th>Source (en)</th>
<th>Target (fr)</th>
<th>MLE</th>
<th>Tok-Seq</th>
</tr>
</thead>
<tbody>
<tr>
<td>I think it’s conceivable that these data are used for mutual benefit.</td>
<td>J'estime qu'il est concevable que ces données soient utilisées dans leur intérêt mutuel.</td>
<td>Je pense qu'il est possible que ces données soient utilisées à des fins réciproques.</td>
<td>Je pense qu'il est possible que ces données soient utilisées pour le bénéfice mutuel.</td>
</tr>
<tr>
<td>However, given the ease with which their behaviour can be recorded, it will probably not be long before we understand why their tails sometimes go one way, sometimes the other.</td>
<td>Toutefois, étant donné la facilité avec laquelle leurs comportements peuvent être enregistrés, il ne faudra sûrement pas longtemps avant que nous comprenions pourquoi leur queue bouge parfois d'un côté et parfois de l'autre.</td>
<td>Cependant, compte tenu de la facilité avec laquelle on peut enregistrer leur comportement, il ne sera probablement pas temps de comprendre pourquoi leurs contemporains vont parfois une façon, parfois l’autre.</td>
<td>Cependant, compte tenu de la facilité avec laquelle leur comportement peut être enregistré, il ne sera probablement pas long avant que nous ne comprenions la raison pour laquelle il arrive parfois que leurs agresseurs suivent un chemin, parfois l’autre.</td>
</tr>
<tr>
<td>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.</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>
<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>
<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>

Table 7: WMT’14 English-to-French examples
<table>
<thead>
<tr>
<th>Source (de)</th>
<th>und ja, vieles von dem, was heute gesagt wurde, berührt mich sehr, weil viele, viele schöne äußerungen dabei waren, die ich auch durchlebt habe.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Target (en)</td>
<td>and yes, a lot of what is said today really moves me, because many, many nice statements were made, which i also was part of.</td>
</tr>
<tr>
<td>MLE</td>
<td>and yes, a lot of what 's been said to me today, i got very, very, very, very beautiful expressions that i used to live through.</td>
</tr>
<tr>
<td>Tok-Seq</td>
<td>and yes, a lot of what 's been told today is very, very much, because many, lots of beautiful statements that i 've been through.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Source (de)</th>
<th>sie repräsentieren teile der menschlichen vorstellungskraft, die in vergangene zeiten &lt;UNK&gt; . und für alle von uns, werden die träume dieser kinder, wie die träume unserer eigenen kinder teil der geographie der hoffnung .</th>
</tr>
</thead>
<tbody>
<tr>
<td>Target (en)</td>
<td>they represent branches of the human imagination that go back to the dawn of time . and for all of us , the dreams of these children , like the dreams of our own children , become part of the naked geography of hope .</td>
</tr>
<tr>
<td>MLE</td>
<td>they represent parts of the human imagination that were blogging in the past time , and for all of us , the dreams of these children will be like the dreams of our own children part of hope .</td>
</tr>
<tr>
<td>Tok-Seq</td>
<td>and they represent parts of the human imagination that live in past times , and for all of us , the dreams of these children , like the dreams of our own children are part of hope .</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Source (de)</th>
<th>noch besser, er wurde in &lt;UNK&gt; nach den angeblich höchsten standards der nachhaltigkeit gezüchtet.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Target (en)</td>
<td>even better, it was &lt;UNK&gt; to the supposed highest standards of sustainability .</td>
</tr>
<tr>
<td>MLE</td>
<td>even better , he has been bred in love with the highest highest standards of sustainability .</td>
</tr>
<tr>
<td>Tok-Seq</td>
<td>even better , he was raised in terms of dignity , the highest standards of sustainability .</td>
</tr>
</tbody>
</table>

Table 8: IWSLT’14 German-to-English examples