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

Maha Elbayad[1,2], Laurent Besacier[1],and Jakob Verbeek[2]

[1]LIG , [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.

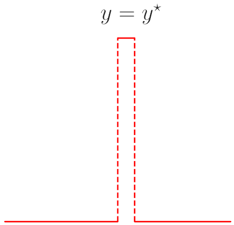
Maximum likelihood estimation (MLE)

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

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

Given the ground truth target sequence $y^\star$:

$$\begin{aligned}
\ell_{\text{MLE}}(y^\star, x) &= -\ln p_\theta(y^\star|x) \\
&= D_{\text{KL}}(\delta(y|y^\star) \| p_\theta(y|x)) && (2) \\
&= \sum_{t=1}^{|y^\star|} D_{\text{KL}}(\delta(y_t|y_t^\star) \| p_\theta(y_t|y_{<t}^\star, x)) && (3)
\end{aligned}$$

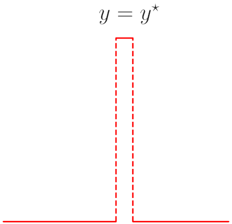


Maximum likelihood estimation (ML)

$$\ell_{\text{MLE}}(y^\star, x) = -\ln p_\theta(y^\star|x)$$

$$= D_{\text{KL}}(\delta(y|y^\star)\|p_\theta(y|x)) \qquad (2)$$

$$= \sum_{t=1}^{T} D_{\text{KL}}(\delta(y_t|y_t^\star)\|p_\theta(y_t|h_t)) \qquad (3)$$



Issues:

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

Loss smoothing

$\delta(y^\star)$ $\quad$ $r_\tau(y|y^\star)$

$$D_{\mathrm{KL}}(\delta(y|y^\star)||p_\theta(y|x))$$

$$\ell^{seq}_{RAML}(y^\star, x) = D_{\mathrm{KL}}(r_\tau(y|y^\star)||p_\theta(y|x))$$ (Norouzi et al, 2016)

Loss smoothing

$\delta(y^\star)$ (*resp.* $\delta(y_t^\star)$) $\qquad r_\tau(y|y^\star)$ (*resp.* $r_\tau(y_t|y_t^\star)$)

$$D_{\mathrm{KL}}(\delta(y|y^\star)||p_\theta(y|x)) \qquad \ell_{RAML}^{seq}(y^\star, x) = D_{\mathrm{KL}}(r_\tau(y|y^\star)||p_\theta(y|x))$$ (Norouzi et al, 2016)

$$\sum_{t=1}^{T} D_{\mathrm{KL}}(\delta(y_t|y_t^\star)||p_\theta(y_t|h_t)) \qquad \ell_{RAML}^{tok}(y^\star, x) = \sum_{t=1}^{T} D_{\mathrm{KL}}(r_\tau(y_t|y_t^\star)||p_\theta(y_t|h_t))$$

Token-level smoothing

Loss smoothing | Token-level

$$\ell^{tok}_{RAML}(y^\star, x) = \sum_{t=1}^{T} D_{\mathrm{KL}}(r_\tau(y_t|y_t^\star) \| p_\theta(y_t|h_t)) \quad (4)$$

- Uniform label smoothing over all words in the vocabulary:

$$r_\tau(y_t|y_t^\star) = \delta(y_t|y_t^\star) + \tau.u(\mathcal{V})$$ (Szegedy et al. 2016)

- We can leverage word co-occurrence statistics to build a non-uniform and "meaningful" distribution.

Loss smoothing

$$\ell^{tok}_{RAML}(y^\star, x) = \sum_{t=1}^{T} D_{\mathrm{KL}}(r_\tau(y_t|y_t^\star)\|p_\theta(y_t|h_t)) \tag{4}$$

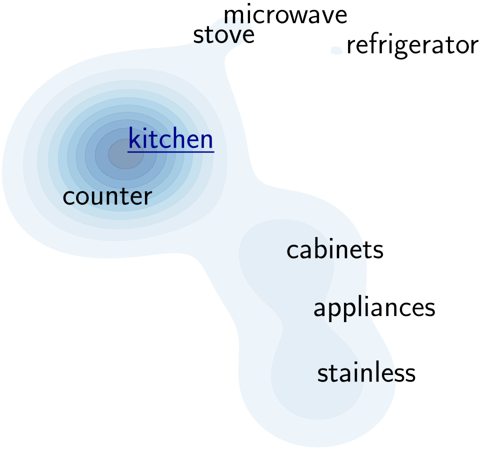
Prerequisite: A word embedding w (e.g. Glove) in the target space and a distance d.

$$r_\tau(y_t|y_t^\star) = \frac{1}{Z}\exp\left(\frac{-\,\mathrm{d}(w(y_t), w(y_t^\star))}{\tau}\right),$$

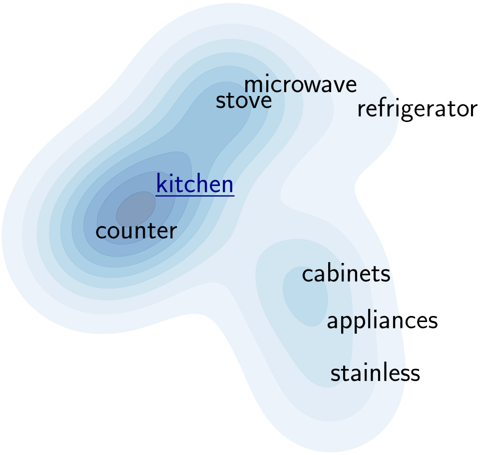
with a temperature τ st. $r_\tau \xrightarrow[\tau\to 0]{} \delta$.

$$Z \text{ st.} \sum_{y_t\in\mathcal{V}} r_\tau(y_t|y_t^\star) = 1$$

Loss smoothing | Token-level



$\tau = 0.12$



$\tau = 0.70$

Loss smoothing | Token-level

$$\ell^{tok}_{RAML}(y^\star, x) = \sum_{t=1}^{T} D_{\mathrm{KL}}\left(r_\tau(y_t|y_t^\star) \| p_\theta(y_t|h_t)\right) \tag{4}$$

$$= \sum_{t=1}^{T} \sum_{y_t \in \mathcal{V}} r_\tau(y_t|y_t^\star) \log\left(\frac{r_\tau(y_t|y_t^\star)}{p_\theta(y_t|h_t)}\right) \tag{5}$$

We can estimate the exact KL divergence for every target token.
No approximation needed.

Sequence-level smoothing

$$\ell^{seq}_{RAML}(y^\star, x) = D_{\mathrm{KL}}(r_\tau(y|y^\star) \| p_\theta(y|x)) \tag{6}$$

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

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

$$Z \text{ st. } \sum_{y \in \mathcal{V}^n, n \in \mathbb{N}} r_\tau(y|y^\star) = 1$$

Possible (pseudo-)distances:

- Hamming
- Edit
- 1−BLEU
- 1−CIDEr

Loss smoothing | Sequence-level

Can we evaluate the partition function Z for a given reward?

$$r_\tau(y_t|y_t^\star) = \frac{1}{\color{red}{Z}} \exp\left(\frac{-\,\mathrm{d}(y, y^\star)}{\tau}\right),$$

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

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 $y^\star$:

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

- 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_τ with Hamming distance:

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

Monte Carlo estimation:

$$\ell^{seq}_{RAML}(y^\star, x) = D_{\mathrm{KL}}(r_\tau(y|y^\star) \| p_\theta(y|x)) \tag{6}$$

$$= -\mathbb{E}_{r_\tau}[\log p_\theta(.|x)] + cst \tag{7}$$

$$(y^l \sim r_\tau) \qquad \approx -\frac{1}{L}\sum_{l=1}^{L} \log p_\theta(y^l|x) \tag{8}$$

We cannot "easily" sample from more complicated rewards such as BLEU or CIDEr.

Importance sampling:

$$\ell^{seq}_{RAML}(y^\star, x) = -\mathbb{E}_{r_\tau}[\log p_\theta(.|x)] \tag{9}$$

$$= -\mathbb{E}_q[\frac{r_\tau}{q}\log p_\theta] \tag{10}$$

$$(y^l \sim q) \approx -\frac{1}{L}\sum_{l=1}^{L}\omega_l \log p_\theta(y^l|x) \tag{11}$$

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

Choose q the reward distribution relative to Hamming distance.

$$\ell^{seq}_{RAML}(y^\star, x) = D_{\mathrm{KL}}(r_\tau(y|y^\star)\|p_\theta(y|x)) \tag{6}$$

Can we reduce the support of r_τ?

$$r_\tau(y|y^\star) = \frac{1}{Z}\exp\left(\frac{-\,\mathrm{d}(y, y^\star)}{\tau}\right), \quad Z = \sum_{y \in \mathcal{V}^T} \exp\left(\frac{-\,\mathrm{d}(y, y^\star)}{\tau}\right)$$

Reduce the support from $\mathcal{V}^{|y^\star|}$ to $\mathcal{V}_{sub}^{|y^\star|}$ where $\mathcal{V}_{sub} \subset \mathcal{V}$.

- $\mathcal{V}_{sub} = \mathcal{V}_{batch}$: tokens occuring in the SGD mini-batch.
- $\mathcal{V}_{sub} = \mathcal{V}_{refs}$: tokens occuring in the available references.

Default training

$$\ell^{seq}_{RAML}(y^\star, x) = -\mathbb{E}_{r_\tau}[\log p_\theta(.|x)]$$
$$\approx -\frac{1}{L}\sum_{l=1}^{L} \log p_\theta(y^l|x)$$

$\forall l$, y^l is:

1. **forwarded** in the RNN.
2. used as target.

$$\log p_\theta(y_l|y_l, x)$$

Lazy training

$$\ell^{seq}_{RAML}(y^\star, x) = -\mathbb{E}_{r_\tau}[\log p_\theta(.|x)]$$
$$\approx -\frac{1}{L}\sum_{l=1}^{L} \log p_\theta(y^l|x)$$

$\forall l$, y^l is:

1. **not forwarded** in the RNN.
2. used as target.

$$\log p_\theta(y_l|y^\star, x)$$

Default training

$$\ell^{seq}_{RAML}(y^\star, x) = -\mathbb{E}_{r_\tau}[\log p_\theta(.|x)]$$
$$\approx -\frac{1}{L}\sum_{l=1}^{L} \log p_\theta(y^l|x)$$

$\forall l,\ y^l$ is:

1. **forwarded** in the RNN.
2. used as target.

$$\log p_\theta(y_l|y_l, x)$$

Complexity : $\mathcal{O}(2L.\lambda)$

Lazy training

$$\ell^{seq}_{RAML}(y^\star, x) = -\mathbb{E}_{r_\tau}[\log p_\theta(.|x)]$$
$$\approx -\frac{1}{L}\sum_{l=1}^{L} \log p_\theta(y^l|x)$$

$\forall l,\ y^l$ is:

1. **not forwarded** in the RNN.
2. used as target.

$$\log p_\theta(y_l|y^\star, x)$$

Complexity: $\mathcal{O}((L+1)\lambda)$

$\lambda = |y||\theta_{cell}|$, where θ_{cell} are the cell parameters.

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 bal
- two soccer players pushing each other for the ball
- two soccer players appear to be pushing each othe

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 ≥ 5)

	|images|
Train	82k
Dev	5k
Test	5k

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

- Architecture:
  Top-down attention
  (Anderson et al. 2017)

Image captioning on MS-COCO | Results

Loss	Reward	$\mathcal{V}_{sub}$	BLEU-1	BLEU-4	CIDEr
MLE			73.40	33.11	101.63
Tok	Glove, cosine		**74.01**	**33.25**	**102.81**

Loss	Reward	$\mathcal{V}_{sub}$	BLEU-1	BLEU-4	CIDEr	
MLE			73.40	33.11	101.63	
Tok	Glove, cosine		**74.01**	**33.25**	**102.81**	
Seq	Hamming	$\mathcal{V}$	73.12	32.71	101.25	(Norouzi et al. 2016)
Seq	Hamming	$\mathcal{V}_{batch}$	73.26	32.73	101.90	
Seq, lazy	Hamming	$\mathcal{V}_{batch}$	73.43	32.95	102.03	

Image captioning on MS-COCO | Results

Loss	Reward	$\mathcal{V}_{sub}$	BLEU-1	BLEU-4	CIDEr
MLE			73.40	33.11	101.63
Tok	Glove, cosine		**74.01**	**33.25**	102.81
Seq	Hamming	$\mathcal{V}$	73.12	32.71	101.25
Seq	Hamming	$\mathcal{V}_{batch}$	73.26	32.73	101.90
Seq, lazy	Hamming	$\mathcal{V}_{batch}$	73.43	32.95	102.03
Seq	CIDEr	$\mathcal{V}_{batch}$	73.50	33.04	**102.98**
Seq	CIDEr	$\mathcal{V}_{refs}$	73.42	32.91	102.23
Seq, lazy	CIDEr	$\mathcal{V}_{refs}$	73.92	33.10	102.64

Image captioning on MS-COCO | Results

Loss	Reward	$\mathcal{V}_{sub}$	BLEU-1	BLEU-4	CIDEr
MLE			73.40	33.11	101.63
Tok	Glove, cosine		74.01	33.25	102.81
Seq	Hamming	$\mathcal{V}$	73.12	32.71	101.25
Seq	Hamming	$\mathcal{V}_{batch}$	73.26	32.73	101.90
Seq, lazy	Hamming	$\mathcal{V}_{batch}$	73.43	32.95	102.03
Seq	CIDEr	$\mathcal{V}_{batch}$	73.50	33.04	102.98
Seq	CIDEr	$\mathcal{V}_{refs}$	73.42	32.91	102.23
Seq, lazy	CIDEr	$\mathcal{V}_{refs}$	73.92	33.10	102.64
Tok-Seq	CIDEr	$\mathcal{V}_{refs}$	**74.28**	**33.34**	**103.81**

Machine translation | Setup

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

IWSLT'14 DE→EN

	$\|$Pairs$\|$
Train	153k
Dev	7k
Test	7k

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

WMT'14 EN→FR

	$\|$Pairs$\|$
Train	12M
Dev	6k
Test	3k

- $|\mathcal{V}| = 30k$ words.

Machine translation | Results

Loss	Reward	$\mathcal{V}_{sub}$	**WMT'14 En→Fr**	**IWSLT'14 De→En**
MLE			30.03	27.55
tok	Glove, cosine		**30.19**	**27.83**

Loss	Reward	$\mathcal{V}_{sub}$	**WMT'14 En→Fr**	**IWSLT'14 De→En**	
MLE			30.03	27.55	
tok	Glove, cosine		30.19	27.83	
Seq	Hamming	$\mathcal{V}$	30.85	27.98	(Norouzi et al. 2016)
Seq	Hamming	$\mathcal{V}_{batch}$	31.18	**28.54**	
Seq	BLEU-4	$\mathcal{V}_{batch}$	**31.29**	28.53	

Loss	Reward	$\mathcal{V}_{sub}$	**WMT'14 En→Fr**	**IWSLT'14 De→En**
MLE			30.03	27.55
tok	Glove, cosine		30.19	27.83
Seq	Hamming	$\mathcal{V}$	30.85	27.98
Seq	Hamming	$\mathcal{V}_{batch}$	31.18	28.54
Seq	BLEU-4	$\mathcal{V}_{batch}$	31.29	28.53
Tok-Seq	Hamming	$\mathcal{V}_{batch}$	31.36	28.70
Tok-Seq	BLEU-4	$\mathcal{V}_{batch}$	**31.39**	**28.74**

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.

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: $\alpha, \alpha_1, \alpha_2 \in (0,1)$ $(\forall \alpha,\ \bar{\alpha} = 1-\alpha)$.

Combininig ML and RAML:

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

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

Combininig the smoothing schemes:

$$\begin{aligned} \ell^{\text{seq, tok}}_{\text{RAML},\alpha_1,\alpha_2}(y^\star, x) &= \alpha_1 \mathbb{E}_{r_\tau}[\ell^{tok}_{\text{RAML}}(y, x)] + \bar{\alpha}_1 \ell^{tok}_{\text{RAML}}(y^\star, x) \\ &= \alpha_1 \mathbb{E}_{r_\tau}[\alpha_2 \ell^{tok}_{\text{RAML}}(y, x) + \bar{\alpha}_2 \ell_{\text{MLE}}(y, x)] \\ &+ \bar{\alpha}_1(\alpha_2 \ell^{tok}_{\text{RAML}}(y^\star, x) + \bar{\alpha}_2 \ell_{\text{MLE}}(y^\star, x)). \end{aligned} \tag{14}$$

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.

Loss	MLE	Tok	Seq	Seq lazy	Seq	Seq lazy	Seq	Seq lazy	Tok-Seq	Tok-Seq	Tok-Seq
Reward		Glove sim	Hamming								
$\mathcal{V}_{sub}$			$\mathcal{V}$	$\mathcal{V}$	$\mathcal{V}_{batch}$	$\mathcal{V}_{batch}$	$\mathcal{V}_{refs}$	$\mathcal{V}_{refs}$	$\mathcal{V}$	$\mathcal{V}_{batch}$	$\mathcal{V}_{refs}$
ms/batch	347	359	390	349	395	337	401	336	445	446	453

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 par

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 siting in a pile
bananas and an orange are sitting together on the clot
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

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

MS-COCO server results

	BLEU-1		BLEU-2		BLEU-3		BLEU-4		METEOR		ROUGE-L		CIDEr		SPICE	
	c5	c40	c5	c40	c5	c40	c5	c40	c5	c40	c5	c40	c5	c40	c5	c40
Google NIC^{+} (Vinyals et al., 2015)	71.3	89.5	54.2	80.2	40.7	69.4	30.9	58.7	25.4	34.6	53.0	68.2	94.3	94.6	18.2	63.6
Hard-Attention (Xu et al., 2015)	70.5	88.1	52.8	77.9	38.3	65.8	27.7	53.7	24.1	32.2	51.6	65.4	86.5	89.3	17.2	59.8
ATT-FCN^{+} (You et al., 2016)	73.1	90.0	56.5	81.5	42.4	70.9	31.6	59.9	25.0	33.5	53.5	68.2	94.3	95.8	18.2	63.1
Review Net^{+} (Yang et al., 2016)	72.0	90.0	55.0	81.2	41.4	70.5	31.3	59.7	25.6	34.7	53.3	68.6	96.5	96.9	18.5	64.9
Adaptive^{+} (Lu et al., 2017)	74.8	92.0	58.4	84.5	44.4	74.4	33.6	63.7	26.4	35.9	55.0	70.5	104.2	105.9	19.7	67.3
SCST:Att2all$^{+\dagger}$ (Rennie et al., 2017)	78.1	93.7	61.9	86.0	47.0	75.9	35.2	64.5	27.0	35.5	56.3	70.7	114.7	116.7	-	-
LSTM-A3$^{+\dagger\circ}$ (Yao et al., 2017)	78.7	93.7	62.7	86.7	47.6	76.5	35.6	65.2	27.0	35.4	56.4	70.5	116	118	-	-
Up-Down$^{+\dagger\circ}$ (Anderson et al., 2017)	80.2	95.2	64.1	88.8	49.1	79.4	36.9	68.5	27.6	36.7	57.1	72.4	117.9	120.5	-	-
Ours: Tok-Seq CIDEr	72.6	89.7	55.7	80.9	41.2	69.8	30.2	58.3	25.5	34.0	53.5	68.0	96.4	99.4	-	-
Ours: Tok-Seq CIDEr $^{+}$	74.9	92.4	58.5	84.9	44.8	75.1	34.3	64.7	26.5	36.1	55.2	71.1	103.9	104.2	-	-

Table: MS-COCO 's server evaluation . ($^{+}$) for ensemble submissions, (†) for submissions with CIDEr optimization and ($^{\circ}$) for models using additional data.

References I

P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, and L. Zhang. 2017. Bottom-up and top-down attention for image captioning and visual question answering. *arXiv preprint arXiv:1707.07998*.

J. Lu, C. Xiong, D. Parikh, and R. Socher. 2017. Knowing when to look: Adaptive attention via a visual sentinel for image captioning. In *CVPR*.

S. Rennie, E. Marcheret, Y. Mroueh, J. Ross, and V. Goel. 2017. Self-critical sequence training for image captioning. In *CVPR*.

O. Vinyals, A. Toshev, S. Bengio, and D. Erhan. 2015. Show and tell: A neural image caption generator. In *CVPR*.

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*.

References II

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

T. Yao, Y. Pan, Y. Li, Z. Qiu, and T. Mei. 2017. Boosting image captioning with attributes. In *ICLR*.

Q. You, H. Jin, Z. Wang, C. Fang, and J. Luo. 2016. Image captioning with semantic attention. In *CVPR*.