Reinforced Video Captioning with Entailment Rewards

Ramakanth Pasunuru and Mohit Bansal
UNC Chapel Hill
{ram, mbansal}@cs.unc.edu

Abstract

Sequence-to-sequence models have shown promising improvements on the temporal task of video captioning, but they optimize word-level cross-entropy loss during training. First, using policy gradient and mixed-loss methods for reinforcement learning, we directly optimize sentence-level task-based metrics (as rewards), achieving significant improvements over the baseline, based on both automatic metrics and human evaluation on multiple datasets. Next, we propose a novel entailment-enhanced reward (CIDEnt) that corrects phrase-matching based metrics (such as CIDEr) to only allow for logically-implied partial matches and avoid contradictions, achieving further significant improvements over the CIDEr-reward model. Overall, our CIDEnt-reward model achieves the new state-of-the-art on the MSR-VTT dataset.

1 Introduction

The task of video captioning (Fig. 1) is an important next step to image captioning, with additional modeling of temporal knowledge and action sequences, and has several applications in online content search, assisting the visually-impaired, etc. Advancements in neural sequence-to-sequence learning has shown promising improvements on this task, based on encoder-decoder, attention, and hierarchical models (Venugopalan et al., 2015a; Pan et al., 2016a). However, these models are still trained using a word-level cross-entropy loss, which does not correlate well with the sentence-level metrics that the task is finally evaluated on (e.g., CIDEr, BLEU). Moreover, these models suffer from exposure bias (Ranzato et al., 2016), which occurs when a model is only exposed to the training data distribution, instead of its own predictions. First, using a sequence-level training, policy gradient approach (Ranzato et al., 2016), we allow video captioning models to directly optimize these non-differentiable metrics, as rewards in a reinforcement learning paradigm. We also address the exposure bias issue by using a mixed-loss (Paulus et al., 2017; Wu et al., 2016), i.e., combining the cross-entropy and reward-based losses, which also helps maintain output fluency.

Next, we introduce a novel entailment-corrected reward that checks for logically-directed partial matches. Current reinforcement-based text generation works use traditional phrase-matching metrics (e.g., CIDEr, BLEU) as their reward function. However, these metrics use undirected n-gram matching of the machine-generated caption with the ground-truth caption, and hence fail to capture its directed logical correctness. Therefore, they still give high scores to even those generated captions that contain a single but critical wrong word (e.g., negation, unrelated action or object), because all the other words still match with the ground truth. We introduce CIDEnt, which penalizes the phrase-matching metric (CIDEr) based reward, when the entailment score is low. This ensures that a generated caption gets a high re-
ward only when it is a directed match with (i.e., it is logically implied by) the ground truth caption, hence avoiding contradictory or unrelated information (e.g., see Fig. 1). Empirically, we show that first the CIDEr-reward model achieves significant improvements over the cross-entropy baseline (on multiple datasets, and automatic and human evaluation); next, the CIDEnt-reward model further achieves significant improvements over the CIDEr-based reward. Overall, we achieve the new state-of-the-art on the MSR-VTT dataset.

2 Related Work

Past work has presented several sequence-to-sequence models for video captioning, using attention, hierarchical RNNs, 3D-CNN video features, joint embedding spaces, language fusion, etc., but using word-level cross entropy loss training (Venugopalan et al., 2015a; Yao et al., 2015; Pan et al., 2016a,b; Venugopalan et al., 2016).

Policy gradient for image captioning was recently presented by Ranzato et al. (2016), using a mixed sequence level training paradigm to use non-differentiable evaluation metrics as rewards. Liu et al. (2016b) and Rennie et al. (2016) improve upon this using Monte Carlo roll-outs and a test inference baseline, respectively. Paulus et al. (2017) presented summarization results with ROUGE rewards, in a mixed-loss setup.

Recognizing Textual Entailment (RTE) is a traditional NLP task (Dagan et al., 2006; Lai and Hockenmaier, 2014; Jimenez et al., 2014), boosted by a large dataset (SNLI) recently introduced by Bowman et al. (2015). There have been several leaderboard models on SNLI (Cheng et al., 2016; Rocktäschel et al., 2016); we focus on the decomposable, intra-sentence attention model of Parikh et al. (2016). Recently, Pasunuru and Bansal (2017) used multi-task learning to combine video captioning with entailment and video generation.

1 Several papers have presented the relative comparison of image captioning metrics, and their pros and cons (Vedantam et al., 2015; Anderson et al., 2016; Liu et al., 2016b; Hodosh et al., 2013; Elliott and Keller, 2014).

3 Models

Attention Baseline (Cross-Entropy) Our attention-based seq-to-seq baseline model is similar to the Bahdanau et al. (2015) architecture, where we encode input frame level video features \{f_{1:n}\} via a bi-directional LSTM-RNN and then generate the caption \(w_{1:m}\) using an LSTM-RNN with an attention mechanism. Let \(\theta\) be the model parameters and \(w_{1:m}\) be the ground-truth caption, then the cross entropy loss function is:

\[
L(\theta) = -\sum_{t=1}^{m} \log p(w_t^*|w_{1:t-1}, f_{1:n})
\]

where \(p(w_{1:t-1}, f_{1:n}) = \text{softmax}(W^T h_t^d)\), \(W^T\) is the projection matrix, and \(w_t^*\) and \(h_t^d\) are the generated word and the RNN decoder hidden state at time step \(t\), computed using the standard RNN recursion and attention-based context vector \(c_t\). Details of the attention model are in the supplementary (due to space constraints).

Reinforcement Learning (Policy Gradient) In order to directly optimize the sentence-level test metrics (as opposed to the cross-entropy loss above), we use a policy gradient \(p_\theta\), where \(\theta\) represent the model parameters. Here, our baseline model acts as an agent and interacts with its environment (video and caption). At each time step, the agent generates a word (action), and the generation of the end-of-sequence token results in a reward \(r\) to the agent. Our training objective is to minimize the negative expected reward function:

\[
L(\theta) = -\mathbb{E}_{w^* \sim p_\theta}[r(w^*)]
\]

where \(w^*\) is the word sequence sampled from the model. Based on the REINFORCE algorithm (Williams, 1992), the gradients of this non-differentiable, reward-based loss function are:

\[
\nabla_\theta L(\theta) = -\mathbb{E}_{w^* \sim p_\theta}[r(w^*) \cdot \nabla_\theta \log p_\theta(w^*)]
\]

We follow Ranzato et al. (2016) approximating the above gradients via a single sampled word.
Mixed Loss During reinforcement learning, optimizing for only the reinforcement loss (with automatic metrics as rewards) doesn’t ensure the readability and fluency of the generated caption, and there is also a chance of gaming the metrics without actually improving the quality of the output (Liu et al., 2016a). Hence, for training our reinforcement based policy gradients, we use a mixed loss function, which is a weighted combination of the cross-entropy loss (XE) and the reinforcement learning loss (RL), similar to the previous work (Paulus et al., 2017; Wu et al., 2016). This mixed loss improves results on the metric used as reward through the reinforcement loss (and improves relevance based on our entailment-enhanced rewards) but also ensures better readability and fluency due to the cross-entropy loss (in which the training objective is a conditioned language model, learning to produce fluent captions). Our mixed loss is defined as:

\[ L_{\text{MIXED}} = (1 - \gamma)L_{\text{XE}} + \gamma L_{\text{RL}} \]  

where \( \gamma \) is a tuning parameter used to balance the two losses. For annealing and faster convergence, we start with the optimized cross-entropy loss baseline model, and then move to optimizing the above mixed loss function.\(^2\)

4 Reward Functions

Caption Metric Reward Previous image captioning papers have used traditional captioning metrics such as CIDEr, BLEU, or METEOR as reward functions, based on the match between the generated caption sample and the ground-truth reference(s). First, it has been shown by Vedantam et al. (2015) that CIDEr, based on a consensus measure across several human reference captions, has a higher correlation with human evaluation than other metrics such as METEOR, ROUGE, and BLEU. They further showed that CIDEr gets better with more number of human references (and this is a good fit for our video captioning datasets, which have 20-40 human references per video).

More recently, Rennie et al. (2016) further showed that CIDEr as a reward in image captioning outperforms all other metrics as a reward, not just in terms of improvements on CIDEr metric, but also on all other metrics. In line with these above previous works, we also found that CIDEr as a reward (‘CIDEr-RL’ model) achieves the best metric improvements in our video captioning task, and also has the best human evaluation improvements (see Sec. 6.3 for result details, incl. those about other rewards based on BLEU, SPICE).

Entailment Corrected Reward Although CIDEr performs better than other metrics as a reward, all these metrics (including CIDEr) are still based on an undirected n-gram matching score between the generated and ground truth captions. For example, the wrong caption “a man is playing football” w.r.t. the correct caption “a man is playing basketball” still gets a high score, even though these two captions belong to two completely different events. Similar issues hold in case of a negation or a wrong action/object in the generated caption (see examples in Table 1).

We address the above issue by using an entailment score to correct the phrase-matching metric (CIDEr or others) when used as a reward, ensuring that the generated caption is logically implied by (i.e., is a paraphrase or directed partial match with) the ground-truth caption. To achieve an accurate entailment score, we adapt the state-of-the-art decomposable-attention model of Parikh et al. (2016) trained on the SNLI corpus (image caption domain). This model gives us a probability for whether the sampled video caption (generated by our model) is entailed by the ground truth caption as premise (as opposed to a contradiction or neu-

<table>
<thead>
<tr>
<th>Ground-truth caption</th>
<th>Generated (sampled) caption</th>
<th>CIDEr</th>
<th>Ent</th>
</tr>
</thead>
<tbody>
<tr>
<td>a man is spreading some butter in a pan</td>
<td>puppies is melting butter on the pan</td>
<td>140.5</td>
<td>0.07</td>
</tr>
<tr>
<td>a panda is eating some bamboo</td>
<td>a panda is eating some fried</td>
<td>256.8</td>
<td>0.14</td>
</tr>
<tr>
<td>a monkey pulls a dogs tail</td>
<td>a monkey pulls a woman</td>
<td>116.4</td>
<td>0.04</td>
</tr>
<tr>
<td>a man is cutting the meat</td>
<td>a man is cutting meat into potato</td>
<td>114.3</td>
<td>0.08</td>
</tr>
<tr>
<td>the dog is jumping in the snow</td>
<td>a dog is jumping in cucumbers</td>
<td>126.2</td>
<td>0.03</td>
</tr>
<tr>
<td>a man and a woman is swimming in the pool</td>
<td>a man and a whale are swimming in a pool</td>
<td>192.5</td>
<td>0.02</td>
</tr>
</tbody>
</table>

Table 1: Examples of captions sampled during policy gradient and their CIDEr vs Entailment scores.

\( ^2 \)We also experimented with the curriculum learning ‘MIXER’ strategy of Ranzato et al. (2016), where the XE+RL annealing is based on the decoder time-steps; however, the mixed loss function strategy (described above) performed better in terms of maintaining output caption fluency.
Similar to the traditional metrics, the overall ‘Ent’ score is the maximum over the entailment scores for a generated caption w.r.t. each reference human caption (around 20/40 per MSR-VTT/YouTube2Text video). CIDEr is defined as:

\[ \text{CIDEr} = \begin{cases} \text{CIDEr} - \lambda, & \text{if } \text{Ent} < \beta \\ \text{CIDEr}, & \text{otherwise} \end{cases} \]

which means that if the entailment score is very low, we penalize the metric reward score by decreasing it by a penalty \( \lambda \). This agreement-based formulation ensures that we only trust the CIDEr-based reward in cases when the entailment score is also high. Using CIDEr – \( \lambda \) also ensures the smoothness of the reward w.r.t. the original CIDEr function (as opposed to clipping the reward to a constant). Here, \( \lambda \) and \( \beta \) are hyperparameters that can be tuned on the dev-set; on light tuning, we found the best values to be intuitive: \( \lambda = 0.45 \) for CIDEr on MSR-VTT dataset; and \( \beta = 0.33 \) (i.e., the 3-class entailment classifier chose contradiction or neutral label for this pair). Table 1 shows some examples of sampled generated captions during our model training, where CIDEr was misleadingly high for incorrect captions, but the low entailment score (probability) helps us successfully identify these cases and penalize the reward.

5 Experimental Setup

Datasets We use 2 datasets: MSR-VTT (Xu et al., 2016) has 10,000 videos, 20 references/video; and YouTube2Text/MSVD (Chen and Dolan, 2011) has 1970 videos, 40 references/video. Standard splits and other details in supp.

Automatic Evaluation We use several standard automated evaluation metrics: METEOR, BLEU-4, CIDEr-D, and ROUGE-L (from MS-COCO evaluation server (Chen et al., 2015)).

Human Evaluation We also present human evaluation for comparison of baseline-XE, CIDEr-RL, and CIDEnt-RL models, esp. because the automatic metrics cannot be trusted solely. Relevance measures how related is the generated caption w.r.t. to the video content, whereas coherence measures readability of the generated caption.

6 Results

6.1 Primary Results

Table 2 shows our primary results on the popular MSR-VTT dataset. First, our baseline attention model trained on cross entropy loss (‘Baseline-XE’) achieves strong results w.r.t. the previous state-of-the-art methods.\(^4\) Next, our policy gradient based mixed-loss RL model with reward as CIDEr (‘CIDEr-RL’) improves significantly\(^3\) over the baseline on all metrics, and not just the CIDEr metric. It also achieves statistically significant improvements in terms of human relevance evaluation (see below). Finally, the last row in Table 2 shows results for our novel CIDEnt-reward RL model (‘CIDEnt-RL’). This model achieves statistically significant\(^6\) improvements on top of the strong CIDEr-RL model, on all automatic metrics (as well as human evaluation). Note that in Table 2, we also report the CIDEr reward scores, and the CIDEnt-RL model strongly outperforms CIDEr and baseline models on this entailment-corrected measure. Overall, we are also the new Rank1 on the MSR-VTT leaderboard, based on their ranking criteria.

Human Evaluation We also perform small human evaluation studies (250 samples from the MSR-VTT test set output) to compare our 3 models pairwise.\(^7\) As shown in Table 3 and Table 4, in terms of relevance, first our CIDEr-RL model statistically significantly outperforms the baseline model (\( p < 0.02 \)); next, our CIDEnt-RL model statistically outperforms the CIDEr-RL model (\( p < \))

---

4 We list previous works’ results as reported by the MSR-VTT dataset paper itself, as well as their 3 leaderboard winners (http://ms-multimedia-challenge.com/leaderboard), plus the 10-ensemble video-entailment generation multi-task model of Pasunuru and Bansal (2017).

3 Statistical significance of \( p < 0.01 \) for CIDEr, METEOR, and ROUGE, and \( p < 0.05 \) for BLEU, based on the bootstrap test (Noreen, 1989; Efron and Tibshirani, 1994).

6 Statistical significance of \( p < 0.01 \) for CIDEr, BLEU, ROUGE, and CIDEr, and \( p < 0.05 \) for METEOR.

7 We randomly shuffle pairs to anonymize model identity and the human rater then chooses the better caption based on relevance and coherence (see Sec. 5). ‘Not Distiguishable’ are cases where the annotator found both captions to be equally good or equally bad.

---

\( \text{CIDEr} \) based on Parikh et al. (2016) is 92% accurate on entailment in the caption domain, hence serving as a highly accurate reward score. For other domains in future tasks such as new summarization, we plan to use the new multi-domain dataset by Williams et al. (2017).
Table 2: Our primary video captioning results on MSR-VTT. All CIDEr-RL results are statistically significant over the baseline XE results, and all CIDEnt-RL results are stat. signif. over the CIDEr-RL results. Human* refers to the ‘pairwise’ comparison of human relevance evaluation between CIDEr-RL and CIDEnt-RL models (see full human evaluations of the 3 models in Table 3 and Table 4).

<table>
<thead>
<tr>
<th>Models</th>
<th>BLEU-4</th>
<th>METEOR</th>
<th>ROUGE-L</th>
<th>CIDEr-D</th>
<th>CIDEnt</th>
<th>Human*</th>
</tr>
</thead>
<tbody>
<tr>
<td>PREVIOUS WORK</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Venugopalan (2015b)*</td>
<td>32.3</td>
<td>23.4</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Yao et al. (2015)*</td>
<td>35.2</td>
<td>25.2</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Xu et al. (2016)</td>
<td>36.6</td>
<td>25.9</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Pasunuru and Bansal (2017)</td>
<td><strong>40.8</strong></td>
<td><strong>28.8</strong></td>
<td>60.2</td>
<td>47.1</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Rank1: v2navigator</td>
<td><strong>40.8</strong></td>
<td>28.2</td>
<td>60.9</td>
<td>44.8</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Rank2: Aalto</td>
<td>39.8</td>
<td>26.9</td>
<td>59.8</td>
<td>45.7</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Rank3: VideoLAB</td>
<td>39.1</td>
<td>27.7</td>
<td>60.6</td>
<td>44.1</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>OUR MODELS</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cross-Entropy (Baseline-XE)</td>
<td>38.6</td>
<td>27.7</td>
<td>39.5</td>
<td>44.6</td>
<td>34.4</td>
<td>-</td>
</tr>
<tr>
<td>CIDEr-RL</td>
<td>39.1</td>
<td>28.2</td>
<td>60.9</td>
<td>51.0</td>
<td>37.4</td>
<td>11.6</td>
</tr>
<tr>
<td>CIDEnt-RL (New Rank1)</td>
<td><strong>40.5</strong></td>
<td><strong>28.4</strong></td>
<td><strong>61.4</strong></td>
<td><strong>51.7</strong></td>
<td><strong>44.0</strong></td>
<td><strong>18.4</strong></td>
</tr>
</tbody>
</table>

Table 5: Results on YouTube2Text (MSVD) dataset. CE = CIDEnt score. H* refer to the pairwise human comparison of relevance.

We also experimented with the new SPICE metric (Anderson et al., 2016) as a reward, but this produced long repetitive phrases (as also discussed in Liu et al. (2016b)).

6.4 Analysis

Fig. 1 shows an example where our CIDEnt-reward model correctly generates a ground-truth style caption, whereas the CIDEr-reward model produces a non-entailed caption because this caption will still get a high phrase-matching score. Several more such examples are in the supp.

7 Conclusion

We first presented a mixed-loss policy gradient approach for video captioning, allowing for metric-based optimization. We next presented an entailment-corrected CIDEnt reward that further improves results, achieving the new state-of-the-art on MSR-VTT. In future work, we are applying our entailment-corrected rewards to other directed generation tasks such as image captioning and document summarization (using the new multi-domain NLI corpus (Williams et al., 2017)).

Acknowledgments

We thank the anonymous reviewers for their helpful comments. This work was supported by a Google Faculty Research Award, an IBM Faculty Award, a Bloomberg Data Science Research Grant, and NVidia GPU awards.
References


Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In ICLR.

Samuel R Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. 2015. A large annotated corpus for learning natural language inference. In EMNLP.


Tim Rocktäschel, Edward Grefenstette, Karl Moritz Hermann, Tomáš Kočíský, and Phil Blunsom. 2016. Reasoning about entailment with neural attention. In ICLR.


Subhashini Venugopalan, Lisa Anne Hendricks, Raymond Mooney, and Kate Saenko. 2016. Improving lstm-based video description with linguistic knowledge mined from text. In EMNLP.


