Discriminability objective for training descriptive image captions

Ruotian Luo
TTI-Chicago

Joint work with
Brian Price (Adobe)
Scott Cohen (Adobe)
Greg Shakhnarovich (TTIC)
Image captioning

The man at bat readies to swing at the pitch while the umpire looks on.

A large bus sitting next to a very tall building.
Example

**ATTN+CIDER**: a large airplane is flying in the sky

**Human**: a large jetliner taking off from an airport runway

**Ours**: a large airplane taking off from runway

**ATTN+CIDER**: a large airplane is flying in the sky

**Human**: a jet airplane flying above the clouds in the distance

**Ours**: a plane flying in the sky with a cloudy sky

**Discriminability objective for training descriptive image captions**
Example

**ATTN+CIDER**: a large airplane is flying in the sky

**Human**: a large jetliner taking off from an airport runway

**ATTN+CIDER**: a large airplane is flying in the sky

**Human**: a jet airplane flying above the clouds in the distance
Example

**ATTN+CIDER**: a large airplane is flying in the sky

**Human**: a large jetliner taking off from an airport runway

**Ours**: a large airplane taking off from runway

**ATTN+CIDER**: a large airplane is flying in the sky

**Human**: a jet airplane flying above the clouds in the distance

**Ours**: a plane flying in the sky with a cloudy sky
Our approach

- We propose a discriminability objective which encourages the output captions to be more discriminative.
- The discriminability objective is derived from how well a (pre-trained) image caption retrieval model can match the images and the generated captions.
Captioning models\textsuperscript{1}

- Basic form: an RNN decoder, producing generative model

\[ c = (w_0 = \langle \text{BOS} \rangle, w_1, \ldots, w_T = \langle \text{EOS} \rangle) \]

\[ p(c | I; \theta) = \prod_t p(w_t | w_{t-1}, I; \theta) \]

- FC model initialized with visual features (CNN, mapped to word representation)

- ATTN model Image is encoded into a set of spatially anchored features; attention, modeled as weights on visual features, evolves with the sequence.

Captioning models training

- **Maximum Likelihood Estimation:**
  \[
  \max_{\theta} \log p(c|I; \theta) = \sum_t \log p(w_t|w_{t-1}, I; \theta)
  \]

- **CIDEr optimization:**
  \[
  \max_{\theta, \hat{c} \sim p(c|I; \theta)} \text{CIDEr}(\hat{c})
  \]

- Use self-critical sequence training method to optimize. (REINFORCE with baseline)

---

Retrieval model²

- Image encoder: image \( I \rightarrow f(I) \)
- Text encoder: caption \( c \rightarrow g(c) \)
- Similarity between image and caption: cosine

\[
s(I, c) = \frac{f(I) \cdot g(c)}{\|f(I)\| \|g(c)\|}
\]

Retrieval model

- Image encoder: image $I \rightarrow f(I)$
- Text encoder: caption $c \rightarrow g(c)$
- Similarity between image and caption: cosine
  \[ s(I, c) = \frac{f(I) \cdot g(c)}{\|f(I)\| \|g(c)\|} \]

- Contrastive loss:
  \[ L_{CON}(c, I) = \max_{c'}[\alpha + s(I, c') - s(I, c)]_+ + \max_{I'}[\alpha + s(I', c) - s(I, c)]_+ \]

- Intuition: score of correct match $(I, c)$ should be higher than score of any mismatched pair $(I, c')$ or $(I', c)$

---

Retrieval model

- Image encoder: image \( I \rightarrow f(I) \)
- Text encoder: caption \( c \rightarrow g(c) \)
- Similarity between image and caption: cosine

\[
s(I, c) = \frac{f(I) \cdot g(c)}{\|f(I)\| \|g(c)\|}
\]

- Contrastive loss:

\[
L_{\text{CON}}(c, I) = \max_{c'}[\alpha + s(I, c') - s(I, c)]_+ + \max_{I'}[\alpha + s(I', c) - s(I, c)]_+
\]

- Intuition: score of correct match \((I, c)\) should be higher than score of any mismatched pair \((I, c')\) or \((I', c)\)

\(\max\) is taken over some set of captions/images

---

Discriminability objective

\[ L_{\text{CON}}(c, I) = \max_{c'} [\alpha + s(I, c') - s(I, c)]_+ + \max_{I'} [\alpha + s(I', c) - s(I, c)]_+ \]

- Can be seen as a discriminability measurement, the lower the more discriminative.
Discriminability objective

\[ L_{\text{CON}}(c, I) = \max_{c'} [\alpha + s(I, c') - s(I, c)]_+ + \max_{I'} [\alpha + s(I', c) - s(I, c)]_+ \]

- Can be seen as a discriminability measurement, the lower the more discriminative.
- Increase discriminability of the generated caption:
  \[ \min L_{\text{CON}}(\hat{c}, I), \hat{c}(I) \sim p_\theta(c|I) \]
Discriminability objective

\[ L_{\text{CON}}(c, I) = \max_{c'}[\alpha + s(I, c') - s(I, c)]_+ + \max_{I'}[\alpha + s(I', c) - s(I, c)]_+ \]

- Can be seen as a discriminability measurement, the lower the more discriminative.
- Increase discriminability of the generated caption:
  \[ \min L_{\text{CON}}(\hat{c}, I), \; \hat{c}(I) \sim p_\theta(c|I) \]
- Optimizing \( L_{\text{CON}} \) leads to unfluent captions:
  \[ \max \text{CIDEr}(\hat{c}) - \lambda L_{\text{CON}}(\hat{c}, I) \]
Discriminability objective

\[ L_{\text{CON}}(c, I) = \max_{c'} [\alpha + s(I, c') - s(I, c)]_+ + \max_{I'} [\alpha + s(I', c) - s(I, c)]_+ \]

- Can be seen as a discriminability measurement, the lower the more discriminative.
- Increase discriminability of the generated caption:
  \[ \min L_{\text{CON}}(\hat{c}, I), \hat{c}(I) \sim p_\theta(c|I) \]
- Optimizing \( L_{\text{CON}} \) leads to unfluent captions:
  \[ \max \text{CIDEr}(\hat{c}) - \lambda L_{\text{CON}}(\hat{c}, I) \]
  \[ \lambda = 0, \text{ identical to SCST.} \]
Experiments

- **Dataset:** MSCOCO
  - training set of 113,287 images,
  - validation set of 5,000 images,
  - 5,000 images used to set up test set: 1000 challenging image pairs

- **Metrics:**
  - Automatic measurements: BLEU, METEOR, ROUGE, CIDEr, SPICE
  - Human experiment on discriminability:

---

Validation experiments

- Discriminability objective works agnostically to the choice of captioning model.
- ATTN models better than FC models.

<table>
<thead>
<tr>
<th></th>
<th>CIDEr</th>
<th>SPICE</th>
<th>Acc</th>
<th>3 in 5</th>
<th>4 in 5</th>
<th>5 in 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>FC+CIDER</td>
<td>1.0154</td>
<td>0.1899</td>
<td>74.00%</td>
<td>73.04%</td>
<td>50.58%</td>
<td>24.83%</td>
</tr>
<tr>
<td>FC+CIDER+DISC (1)</td>
<td>1.0231</td>
<td>0.1939</td>
<td>79.26%</td>
<td>74.26%</td>
<td>55.53%</td>
<td>24.13%</td>
</tr>
<tr>
<td>ATTN+CIDER</td>
<td>1.1332</td>
<td>0.2083</td>
<td>71.05%</td>
<td>69.97%</td>
<td>51.34%</td>
<td>27.34%</td>
</tr>
<tr>
<td>ATTN+CIDER+DISC (1)</td>
<td>1.1406</td>
<td>0.2113</td>
<td>75.74%</td>
<td>72.70%</td>
<td>53.23%</td>
<td>34.33%</td>
</tr>
</tbody>
</table>
Validation experiments

- Moderate $\lambda = 1$ produces improves both non-discriminative and discriminative score.
- Higher $\lambda$ make captions more discriminative to machine and to humans, but at the cost of fluency.
- Especially surprising result: CIDEr (ostensibly we focus less on maximizing it during training.)

![Graph showing BLEU4, CIDEr, and SPICE scores](image)

**Discriminability objective for training descriptive image captions**
Scores on test set

- Test on ATTN models.
- CACA$^3$: requires seeing the “distractor images” before producing caption for the target

<table>
<thead>
<tr>
<th></th>
<th>BLEU4</th>
<th>CIDEr</th>
<th>SPICE</th>
<th>Acc</th>
<th>3 in 5</th>
<th>4 in 5</th>
<th>5 in 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>74.30%</td>
<td>91.14%</td>
<td>82.38%</td>
<td>57.08%</td>
</tr>
<tr>
<td>MLE</td>
<td>0.5956</td>
<td>1.2198</td>
<td>0.2132</td>
<td>68.60%</td>
<td>72.06%</td>
<td>59.06%</td>
<td>44.25%</td>
</tr>
<tr>
<td>CIDER</td>
<td>0.5971</td>
<td>1.2604</td>
<td>0.2260</td>
<td>68.19%</td>
<td>70.07%</td>
<td>55.95%</td>
<td>35.95%</td>
</tr>
<tr>
<td>CACA</td>
<td>0.4719</td>
<td>0.7656</td>
<td>0.1526</td>
<td>75.80%</td>
<td>74.1%</td>
<td>56.88%</td>
<td>35.19%</td>
</tr>
<tr>
<td>CIDER+D(1)</td>
<td>0.3971</td>
<td>1.2770</td>
<td>0.2302</td>
<td>72.63%</td>
<td>76.91%</td>
<td>61.67%</td>
<td>40.09%</td>
</tr>
<tr>
<td>CIDER+D(10)</td>
<td>0.3538</td>
<td>1.1429</td>
<td>0.2204</td>
<td>79.75%</td>
<td>77.70%</td>
<td>64.63%</td>
<td>44.63%</td>
</tr>
</tbody>
</table>

SPICE

- Semantic propositional image caption evaluation score
- Inference over objects, attributes, relations

F-score is calculated for semantic tuples, e.g.,

\{ (girl), (court), (girl, young), (girl, standing) (court, tennis), (girl, on-top-of, court) \}

Can calculate separately for different aspects: color, counting, size, etc.
Results: SPICE

- Results on 5k validation (all scores scaled ×100)

![Graph showing the relationship between lambda and color, attribute, and counting attributes.](image-url)
Effect on caption diversity

- For the validation set, 5000 images
Examples: disambiguation

- Images with identical baseline captions selected for display purposes
- “Ours” = ATTN+CIDER+D(1)

**Human**: a man riding skis next to a blue sign near a forest

**ATTN+CIDER**: a man standing on skis in the snow

**Ours**: a man standing in the snow with a sign

**Human**: the man is skiing down the hill with his goggles up

**ATTN+CIDER**: a man standing on skis in the snow

**Ours**: a man riding skis on a snow covered slope
Examples: disambiguation

Human: a hot dog served with fries and dip on the side
ATTN+CIDER: a plate of food with meat and vegetables on a table
Ours: a hot dog and french fries on a plate

Human: a plate topped with meat and vegetables and sauce
ATTN+CIDER: a plate of food with meat and vegetables on a table
Ours: a plate of food with carrots and vegetables on a plate
Examples: disambiguation

Human: a train on an overpass with people under it
ATTN+CIDER: a train is on the tracks at a train station
Ours: a red train parked on the side of a building

Human: a train coming into the train station
ATTN+CIDER: a train is on the tracks at a train station
Ours: a green train traveling down a train station
Examples: disambiguation

**Human:** people skiing in the snow on the mountainside

**ATTN+CIDER:** a group of people standing on skis in the snow

**Ours:** a group of people skiing down a snow covered slope

**Human:** two skiers travel along a snowy path towards trees

**ATTN+CIDER:** a group of people standing on skis in the snow

**Ours:** two people standing on skis in the snow

Discriminability objective for training descriptive image captions
Examples: model comparison

- Left: the target image; right (for illustration): a distractor (shown to human subjects in Mechanical Turk evaluation)

**ATTN+MLE**: a large clock tower with a clock on it

**ATTN+CIDER**: a clock tower with a clock on the side of it

**ATTN+CIDER+DISC(1)**: a clock tower with bikes on the side of a river

**ATTN+CIDER+DISC(10)**: a clock tower with bicycles on the boardwalk near a harbor

Discriminability objective for training descriptive image captions
Examples: model comparison

- High $\lambda$ often causes “insistence” on distinctive words, at the expense of fluency

**ATTN+MLE**: a view of an airplane flying through the sky
**ATTN+CIDER**: a plane is flying in the sky
**ATTN+CIDER+DISC(1)**: a plane flying in the sky with a sunset
**ATTN+CIDER+DISC(10)**: a sunset of a sunset with a sunset in the sunset
Examples: model comparison

ATTN+MLE: a couple of people standing next to a stop sign
ATTN+CIDER: a stop sign on the side of a street
ATTN+CIDER+DISC(1): a stop sign in front of a store with umbrellas
ATTN+CIDER+DISC(10): a stop sign sitting in front of a store with shops
Examples: model comparison

ATTN+MLE: a large jetliner sitting on top of an airport tarmac
ATTN+CIDER: a large airplane sitting on the runway at an airport
ATTN+CIDER+DISC(1): a blue airplane sitting on the tarmac at an airport
ATTN+CIDER+DISC(10): a blue and blue airplane sitting on a tarmac with a plane
Conclusions

- Propose a discriminability objective, derived from the loss of a trained image/caption retrieval model
- Improves the quality of resulting captions on both discriminability and standard metrics.
Comparison

**CL:** \((I, c) \text{ vs } (I, c')\),

**GAN:** \((I, c) \text{ vs } (I, \hat{c}), \ (I, c) \text{ vs } (I, c')\)

**Ours:** \((I, \hat{c}) \text{ vs } (I, \hat{c}'), \ (I, \hat{c}) \text{ vs } (I', \hat{c})\)