

# Grounded Video Description

MMML Reading Group

20 March 2019

volkan cirik

# Grounding Phrases is Essential in Captioning



A **man** is seen standing in a **room** speaking to the camera while holding a **bike**.

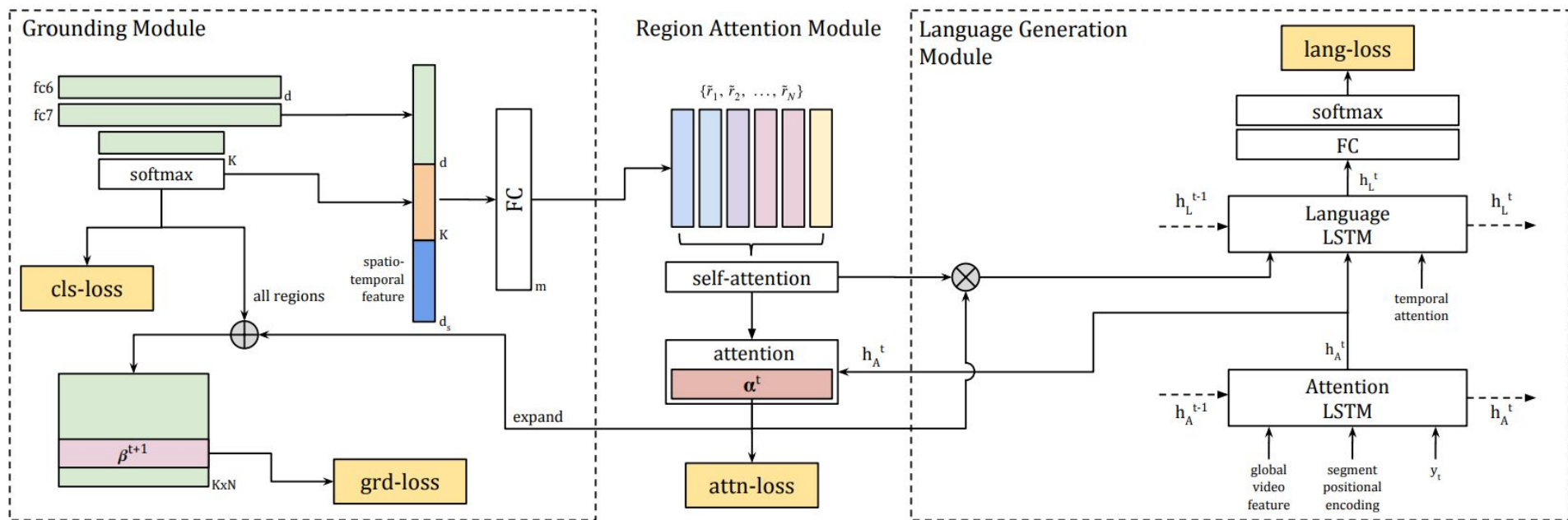


A group of **people** are in a **raft** down a **river**.

# Contribution I: New Set of Annotations

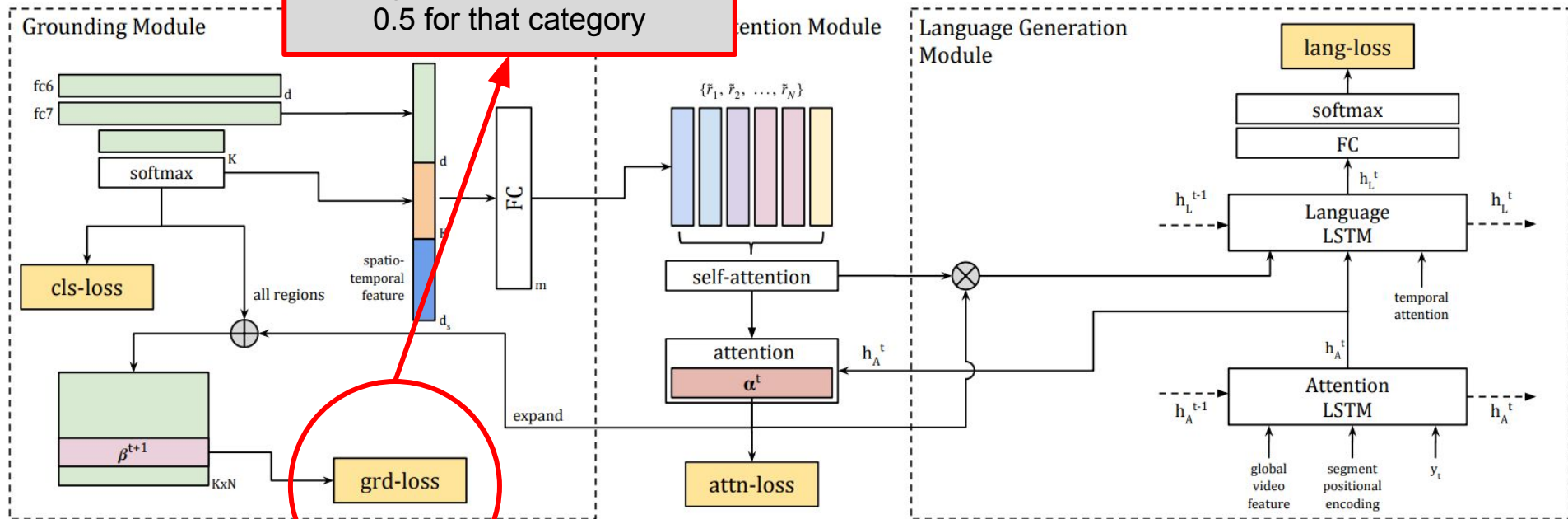
| Dataset                            | Domain       | # Vid/Img  | # Sent     | # Obj      | # BBoxes    |
|------------------------------------|--------------|------------|------------|------------|-------------|
| Flickr30k Entities [26]            | Images       | 32k        | 160k       | 480        | 276k        |
| MPII-MD [30]                       | Video        | ≪1k        | ≪1k        | 4          | 2.6k        |
| YouCook2 [45]                      | Video        | 2k         | 15k        | 67         | 135k        |
| ActivityNet Humans [38]            | Video        | 5.3k       | 5.3k       | 1          | 63k         |
| <b>ActivityNet-Entities (ours)</b> | <b>Video</b> | <b>15k</b> | <b>52k</b> | <b>432</b> | <b>158k</b> |
| –train                             |              | 10k        | 35k        | 432        | 105k        |
| –val                               |              | 2.5k       | 8.6k       | 427        | 26.5k       |
| –test                              |              | 2.5k       | 8.5k       | 421        | 26.1k       |

# Contribution II: Multi-task Objective



# Contribution to Objective

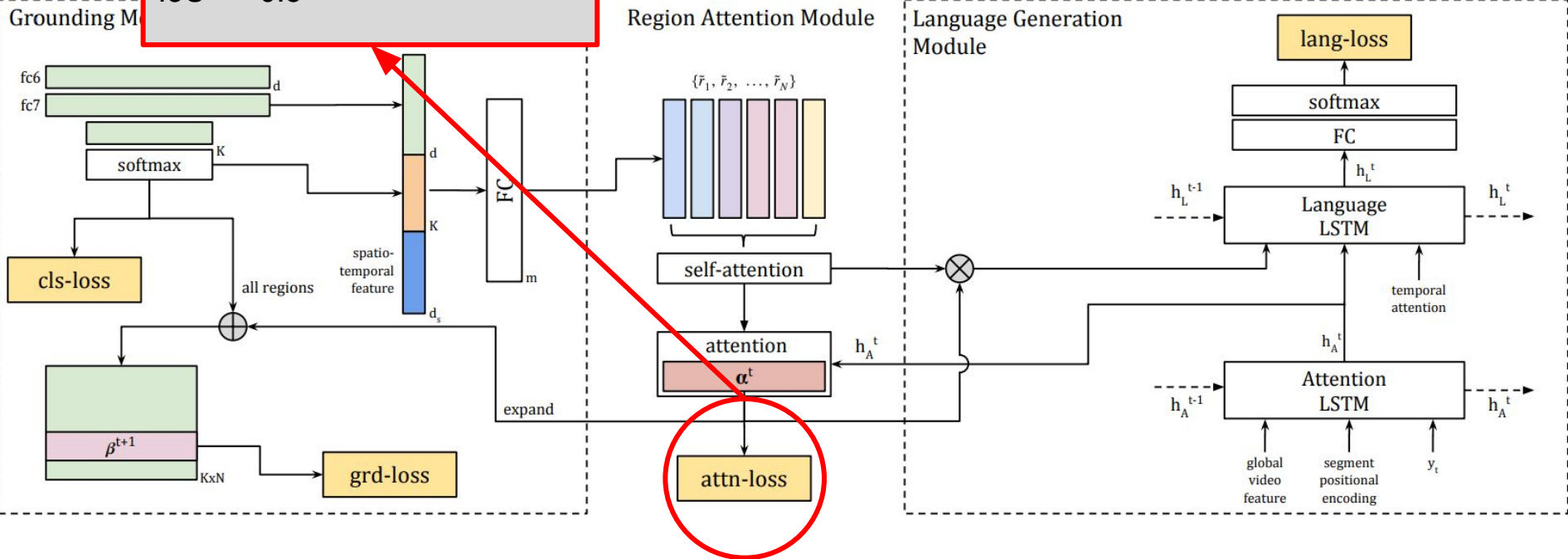
For the object category (e.g. car) of the token generated at  $t+1$ , model should predict 1 for all regions that have IoU  $\geq 0.5$  for that category



# Coarse-grained Region Attention Module

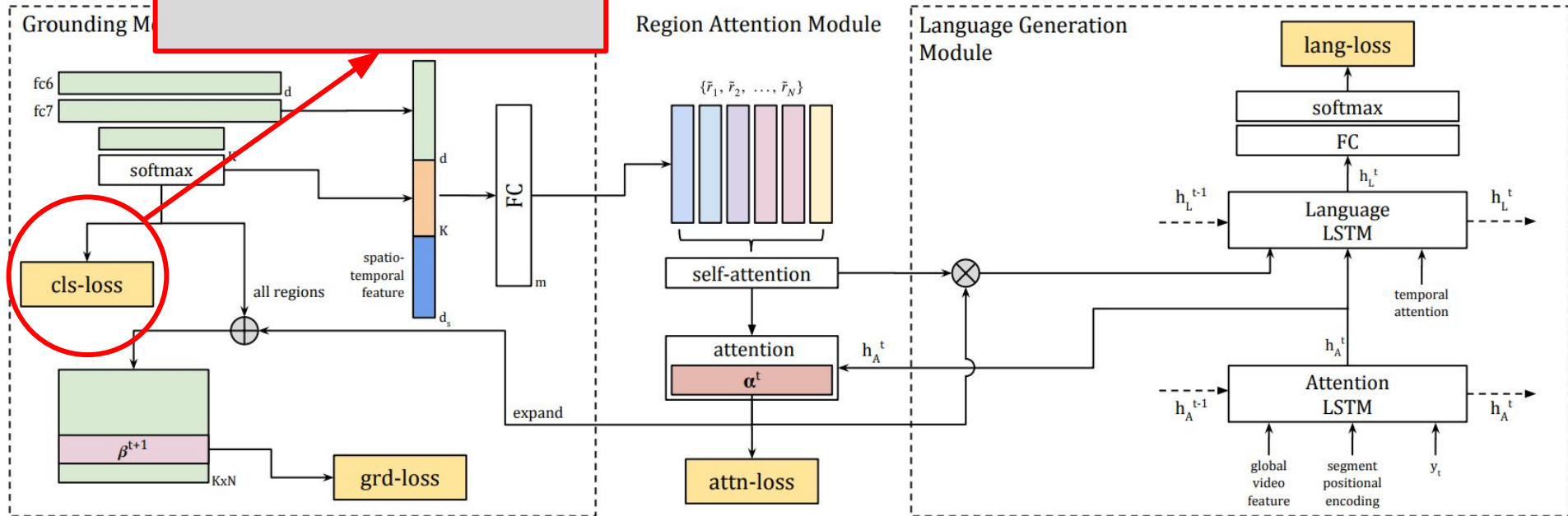
## -task Objective

At time  $t$ , before generating a grounded word, model should attend to all regions that have  $\text{IoU} \geq 0.5$



# Co-task Objective

Model also needs to predict the object categories of the regions.



# Competitive/SOTA Results on Video Captioning

How accurately predicted regions match ground-truth region mentions.

| Method                    | B@1         | B@4         | M           | C           | S           | Attn.       | Grd.        | Prec.       | Recall      | Cls.        |
|---------------------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|
| Masked Transformer [47]   | 22.9        | <b>2.41</b> | 10.6        | <b>46.1</b> | 13.7        | -           | -           | -           | -           | -           |
| Bi-LSTM+TempoAttn [47]    | 22.8        | 2.17        | 10.2        | 42.2        | 11.8        | -           | -           | -           | -           | -           |
| Our Unsup. (w/o SelfAttn) | 23.1        | 2.16        | 10.8        | 44.9        | <b>14.9</b> | 15.9        | 22.2        | 7.43        | 7.00        | 6.50        |
| Our Sup. Attn.+Cls.       | <b>23.6</b> | 2.35        | <b>11.0</b> | 45.5        | 14.7        | <b>34.6</b> | <b>43.2</b> | <b>11.3</b> | <b>15.9</b> | <b>14.6</b> |



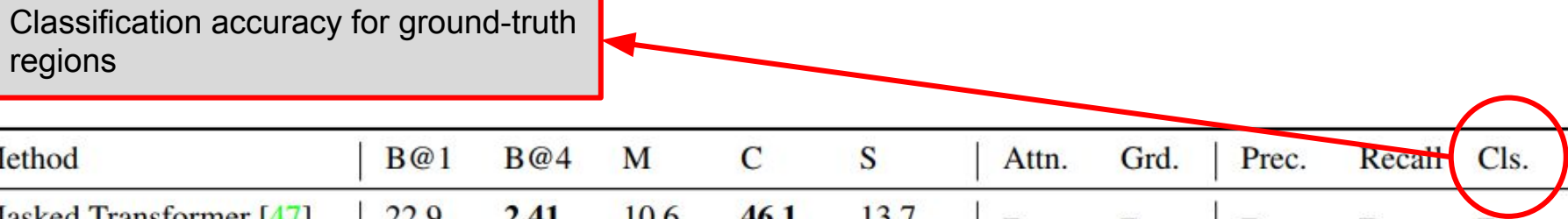
# Competitive/SOTA Results on Video Captioning

Given ground-truth sentence for generation, how accurate in grounding the phrases to ground-truth regions?

| Method                    | B@1         | B@4         | M           | C           | S           | Attn.       | Grd.        | Prec.       | Recall      | Cls.        |
|---------------------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|
| Masked Transformer [47]   | 22.9        | <b>2.41</b> | 10.6        | <b>46.1</b> | 13.7        | –           | –           | –           | –           | –           |
| Bi-LSTM+TempoAttn [47]    | 22.8        | 2.17        | 10.2        | 42.2        | 11.8        | –           | –           | –           | –           | –           |
| Our Unsup. (w/o SelfAttn) | 23.1        | 2.16        | 10.8        | 44.9        | <b>14.9</b> | 15.9        | 22.2        | 7.43        | 7.00        | 6.50        |
| Our Sup. Attn.+Cls.       | <b>23.6</b> | 2.35        | <b>11.0</b> | 45.5        | 14.7        | <b>34.6</b> | <b>43.2</b> | <b>11.3</b> | <b>15.9</b> | <b>14.6</b> |

# Competitive/SOTA Results on Video Captioning

Classification accuracy for ground-truth regions



| Method                    | B@1         | B@4         | M           | C           | S           | Attn.       | Grd.        | Prec.       | Recall      | Cls.        |
|---------------------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|
| Masked Transformer [47]   | 22.9        | <b>2.41</b> | 10.6        | <b>46.1</b> | 13.7        | –           | –           | –           | –           | –           |
| Bi-LSTM+TempoAttn [47]    | 22.8        | 2.17        | 10.2        | 42.2        | 11.8        | –           | –           | –           | –           | –           |
| Our Unsup. (w/o SelfAttn) | 23.1        | 2.16        | 10.8        | 44.9        | <b>14.9</b> | 15.9        | 22.2        | 7.43        | 7.00        | 6.50        |
| Our Sup. Attn.+Cls.       | <b>23.6</b> | 2.35        | <b>11.0</b> | 45.5        | 14.7        | <b>34.6</b> | <b>43.2</b> | <b>11.3</b> | <b>15.9</b> | <b>14.6</b> |

# Competitive/SOTA Results on Image Captioning

| Method                    | VG | Box | B@1         | B@4         | M           | C           | S           | Attn.       | Grd.        | Prec.       | Recall      | Cls.        |
|---------------------------|----|-----|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|
| ATT-FCN* [41]             |    |     | 64.7        | 19.9        | 18.5        | –           | –           | –           | –           | –           | –           | –           |
| NBT* [18]                 |    | ✓   | 69.0        | 27.1        | 21.7        | 57.5        | 15.6        | –           | –           | –           | –           | –           |
| BUTD [2]                  | ✓  |     | 69.4        | <b>27.3</b> | 21.7        | 56.6        | 16.0        | 24.5        | 32.3        | 13.5        | 13.7        | 1.89        |
| Our Unsup. (w/o SelfAttn) | ✓  |     | 69.5        | 27.0        | 22.1        | 60.1        | 16.1        | 21.7        | 25.6        | 11.9        | 11.8        | 18.4        |
| Our Sup. Attn.+Grd.+Cls.  | ✓  | ✓   | <b>69.9</b> | <b>27.3</b> | <b>22.5</b> | <b>62.3</b> | <b>16.5</b> | <b>41.8</b> | <b>51.2</b> | <b>25.0</b> | <b>26.6</b> | <b>19.9</b> |