The winning challenge entry

Ruotian Luo[1], Gilad Vered[2], Lior Bracha[2], Gal Chechik[2], Greg Shakhnarovich[1]

TTI-Chicago[1]

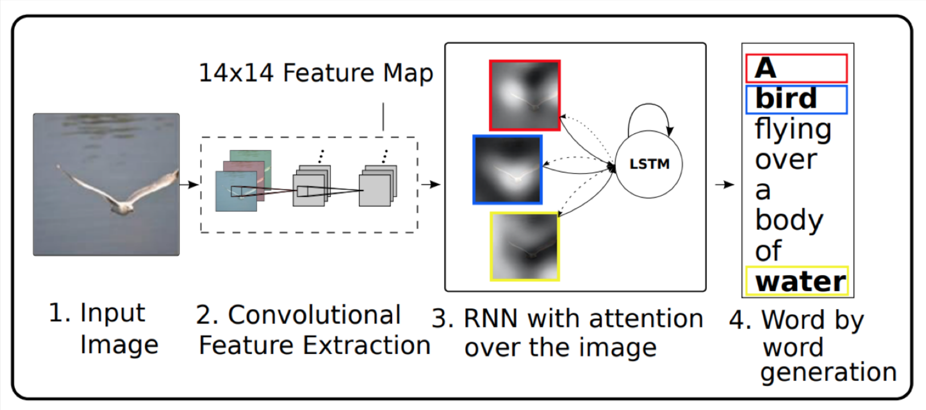
Bar Ilan University [2]

Overview of our submissions

- Two types of captioning models:
  - Attention based **LSTM**
  - **Transformer**
- Novel component: *"drop worst"* mechanism to make learning more robust in presence of many poorly grounded captions
- Two submissions:
  - Ensemble of 3 models: LSTM and two transformer models
    -- trained with CIDEr optimization (reinforcement learning)
    -- **second place on CIDEr (0.99); top human rating**
  - Ensemble of 5 models (including both LSTMs and transformers)
    -- trained with cross-entropy loss +drop worst
    -- **first place on CIDEr (1.04); ranked 4th in human rating**

Model type 1: Attention LSTM

- We use att2in model proposed in Rennie et al.
- A variant of the original attention-LSTM captioner in Xu et al.

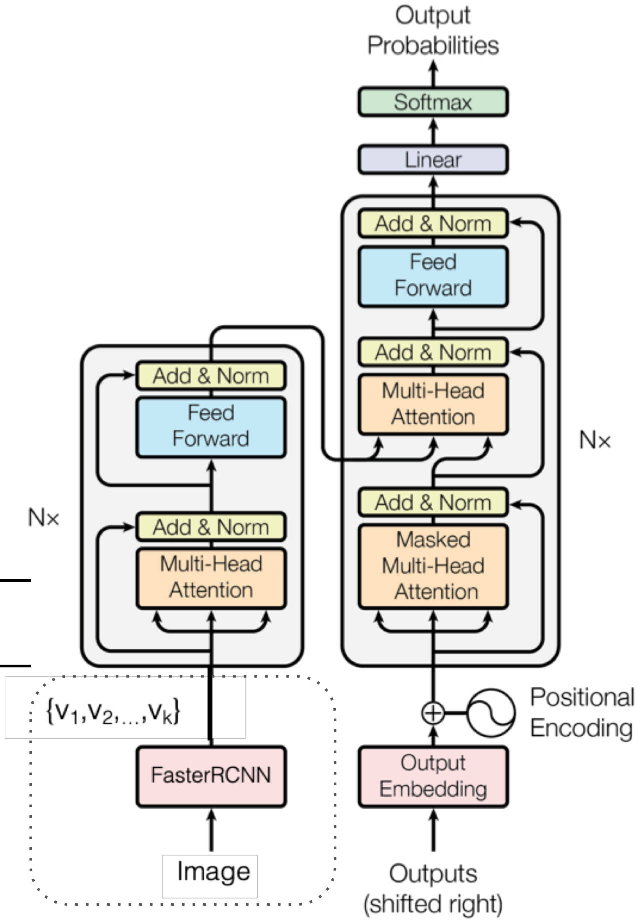


Rennie, Steven J., et al. "Self-critical sequence training for image captioning." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017.
Image credit: Xu, Kelvin, et al. "Show, attend and tell: Neural image caption generation with visual attention." International conference on machine learning. 2015.

Model type 2: Transformer

- State of the art seq2seq model
- Base model is the same as in Vaswani et al.
- Huge model has larger hidden size.

	N	d_{model}	d_{ff}	h	d_k	d_v	P_{drop}
base	6	512	2048	8	64	64	0.1
huge	6	**1024**	**4096**	8	64	64	0.1



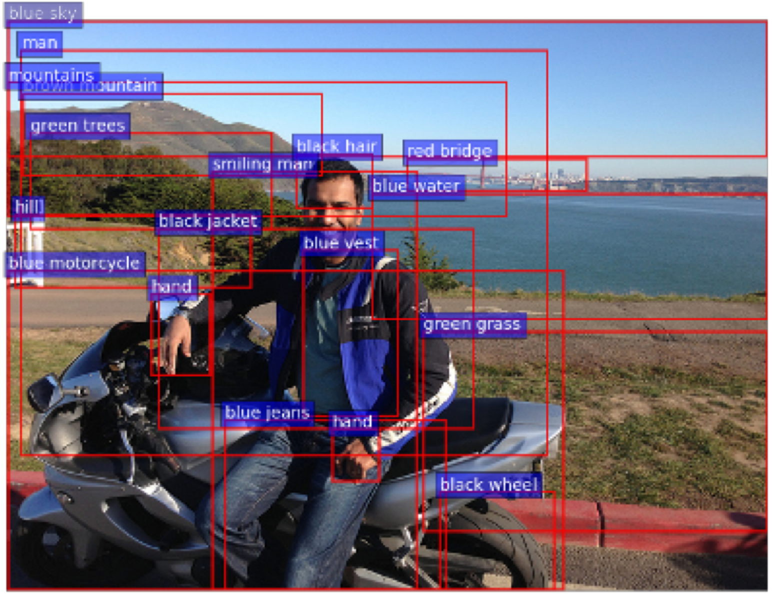
Vaswani, Ashish, et al. "Attention is all you need." Advances in neural information processing systems. 2017.

Image Encoder

Image features in both types of models: following Anderson et al., 2018 (bottom-up attention)

- Image encoding size: K x 2048
- K : number of detection boxes scoring above threshold,

  $10 \leq K \leq 100$



Anderson, Peter, et al. "Bottom-up and top-down attention for image captioning and visual question answering." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018.

Drop worst: motivation

- Many captions in the dataset appear to be poorly grounded

`football team will play for the first time next season`

this princess cross stitch pattern is special because it is modern minimalist suitable for both children and adults

reach new heights on your trip with an adventure

In contrast: grounded/descriptive captions

green basket with yellow flowers of dandelions on the brown wooden background

starfish and seashell with hearts on the sandy beach by the ocean

Drop worst cross entropy

- Normal cross entropy: equal impact of all training samples

$$L = -\frac{1}{N} \sum_{i} \log P(c_i | I_i)$$

c_i: caption of image I_i

- Idea: examples with highest loss (lowest probability) may be *too* hard (not grounded) -- so give up on them for now! ("hard negative *culling*")
- For each batch (after certain epoch), drop (ignore) 20% of the examples with the highest cross entropy loss in that batch

$$L = -\mathrm{Mean}\left[\mathrm{Top}_{80\%}\left\{\log P(c_i | I_i)\right\}\right]$$

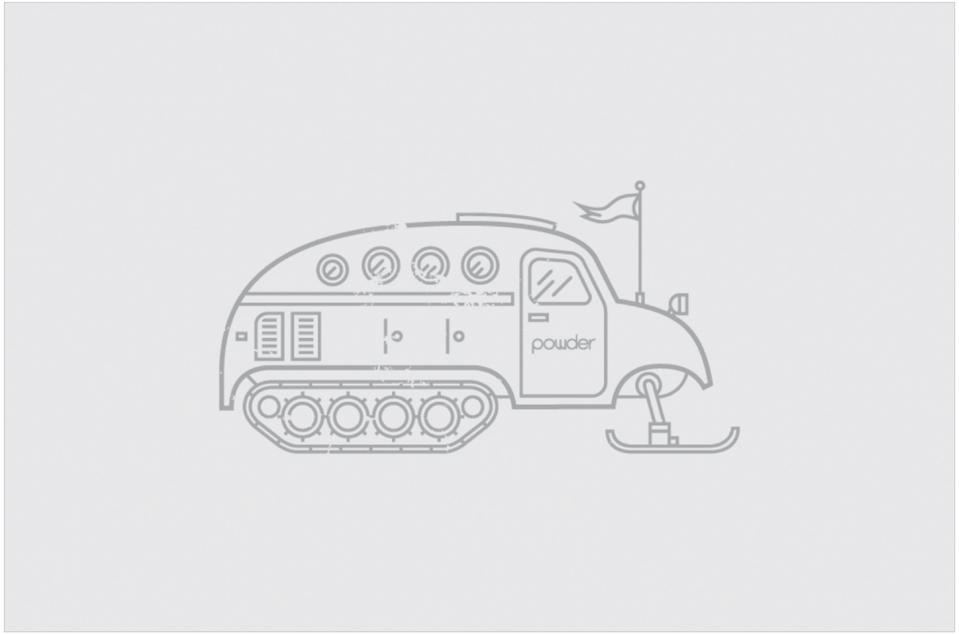
Examples: top probability within a batch

actor during an interview with comedian

football player and battle for the ball

Examples: lowest probability within a batch (dropped)

shirt graphic created for powder



sponsored video this application requires programming language

CIDEr optimization

- We directly optimize CIDEr score of generated captions using Policy Gradient methods.
- This is the loss we are optimizing.
- (R^m is the CIDEr score of sampled caption c^m, b^m is baseline)

$$L = -\frac{1}{M} \sum_{m=1}^{M} (R^m - b^m) \log P(c^m|I)$$

$$b^m = \frac{1}{M-1} \sum R^{\setminus m}$$

Rennie, Steven J., et al. "Self-critical sequence training for image captioning." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017.
Ranzato, Marc'Aurelio, et al. "Sequence level training with recurrent neural networks." arXiv preprint arXiv:1511.06732 (2015).

Other details

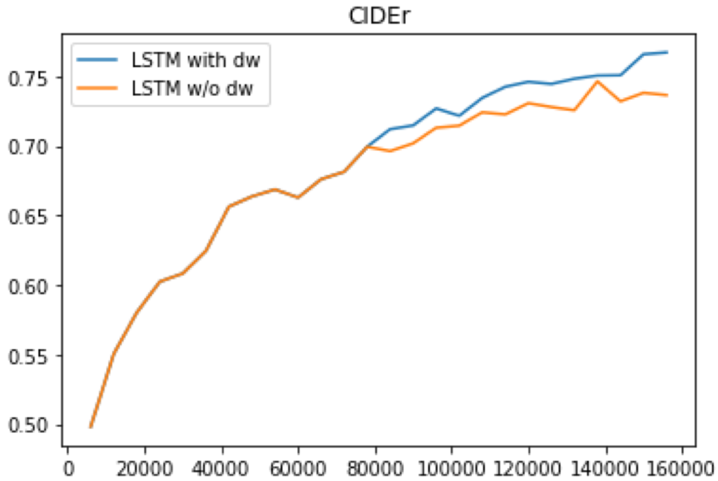
- Training setup:
  - Batch size 250 (drop 20% worst after 6 epochs)
  - Learning rate 5e-4, decay by 0.8 every 3 epochs
  - For transformer, warmup step is 40000 iterations.
  - CIDEr optimization: lr 1e-5; batch size 50.
- At test time (submissions):
  - Beam search with beam size 5
  - Decoding constraints [1]
  - Remove bad endings [2]

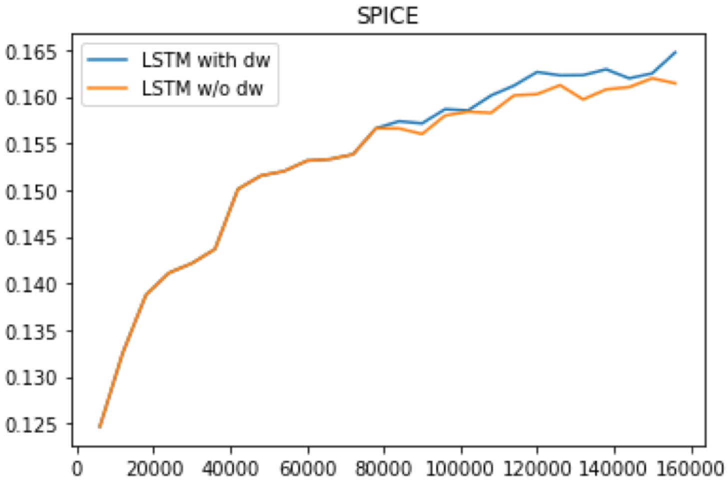
1 Anderson, Peter, et al. "Bottom-up and top-down attention for image captioning and visual question answering." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018.
2 Guo, Tszhang, et al. "Improving Reinforcement Learning Based Image Captioning with Natural Language Prior." arXiv preprint arXiv:1809.06227 (2018).

Analysis

Effect of drop worst

- Results on val set

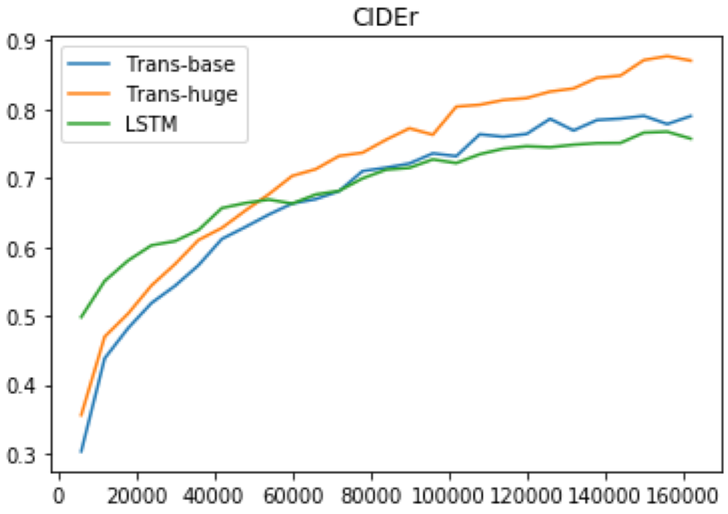


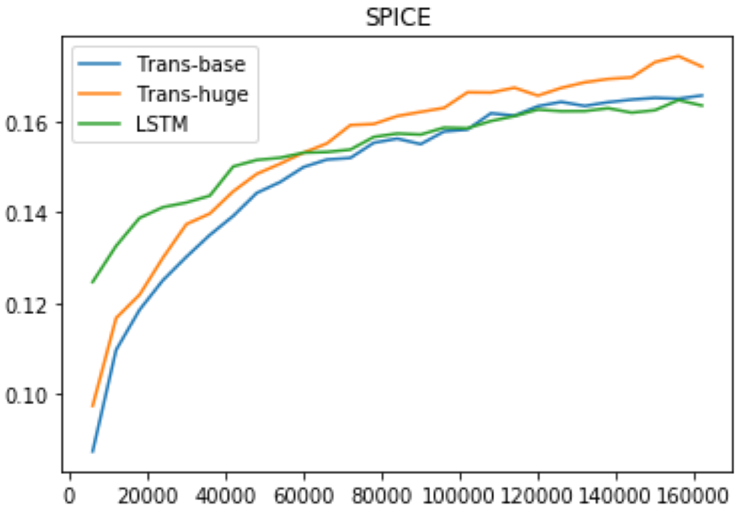


- Baseline model: LSTM
- Consistent improvement with drop worst; use for all models

Results on automatic metrics of different models

- Results on val set with cross-entropy trained models





Combining models (weighted avg. of posteriors)

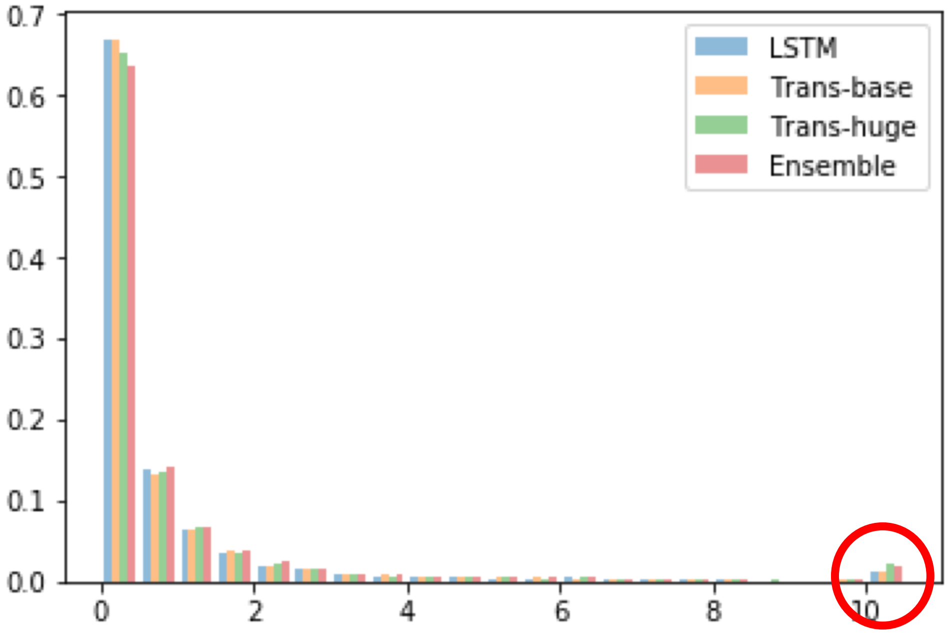
- Results using beam search with beam size 5, combining LSTM and Transformer-huge

LSTM weight	Transformer weight	CIDEr	SPICE	ROUGE_L
1	0	0.7734	0.1586	0.2467
0.8	0.2	0.8501	0.1692	0.2549
0.5	0.5	**0.9235**	**0.1757**	**0.2590**
0.3	0.7	0.9169	0.1750	0.2567
0	1	0.8987	0.1719	0.2520

- Use uniform weights for all ensembles

CIDEr score distribution (val set)

- Frequency of CIDEr scores:
- CIDEr=10 means perfect prediction of GT caption
- Can look in detail at those perfect predictions



The dataset is not balanced

The 10 most frequent captions in training set (counts / frequency)

`actor arrives at the premiere` 7227/0.23

`image may contain person on stage and playing a musical instrument` 4986/0.159

`digital art selected for the #` 4707/0.15

`image may contain person on stage playing a musical instrument and guitar` 2491/0.08

`actor attends the world premiere` 2229/0.07

`image may contain person on stage playing a musical instrument and indoor` 2223/0.07

`a model walks the runway at the fashion show during event` 2037/0.07

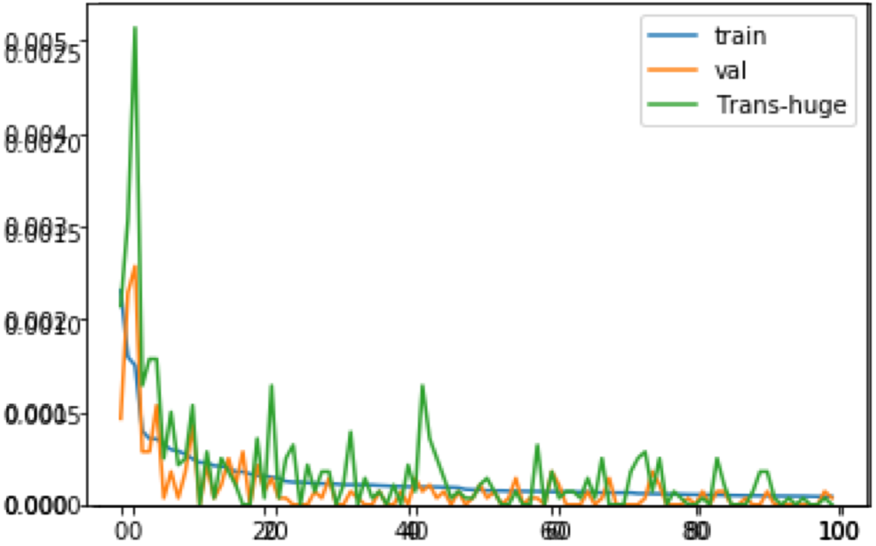
`image may contain person on stage playing a musical instrument and night` 1862/0.06

`football player and battle for the ball` 1811/0.06

`actor attends the premiere during festival` 1701/0.05

Do well on frequent captions

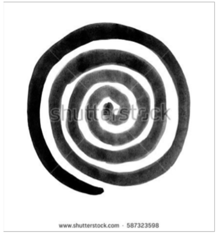
The frequency of top 100 frequent training captions in generated captions.



Other perfectly predicted (unique) captions

- The model is able to generate perfect captions that only appear once or even never in the training set
- Some may be memorization

train



val

`gingerbread little men on the beach`

`statue of builder on the cross`

`spiral in a circle drawn by the brush painted black paint`

Other perfectly predicted unique captions

train

val

`women praying in a mosque`

`builder on the cross stock photo #`

`bicycles parked in the snow`

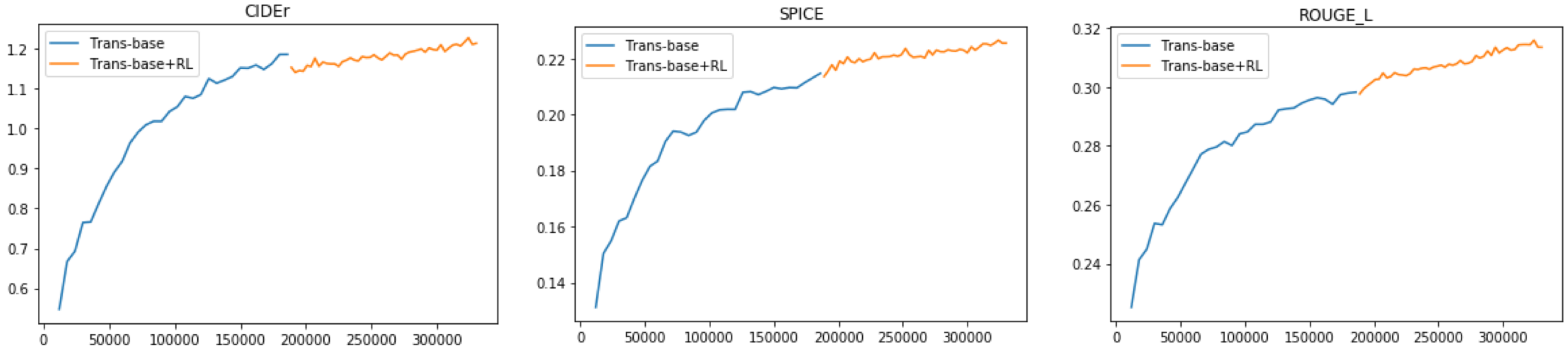
Other perfectly predicted captions

- Can even generate previously unseen GT captions
- Rare: 5 new captions out of total 281 perfectly predicted GT captions in val

black alarm clock on a yellow background royalty free

CIDEr optimization

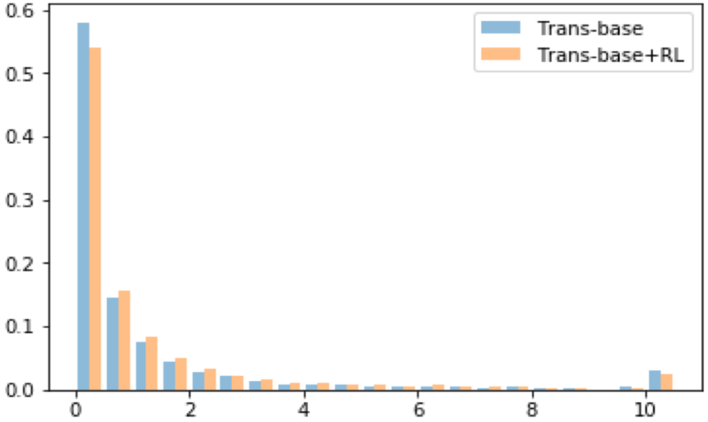
- Direct CIDEr optimization (with RL) did not work as well as on COCO or other datasets. (on COCO, the CIDEr increases drastically once RL kicks in).
- Performance may recover after a while, but that takes much longer. We didn't fully explore this due to time limits

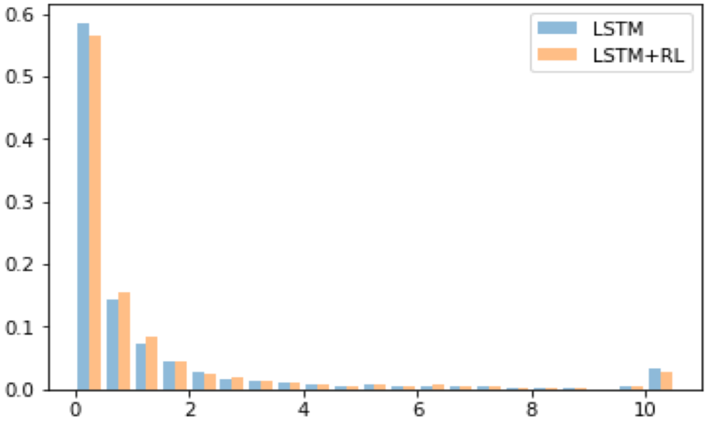


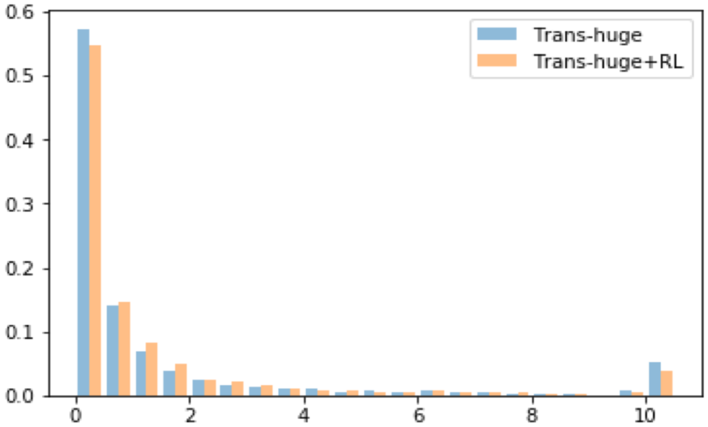
The scores are evaluated on an unofficial train-val split

CIDEr score distribution: a different story?

- Lower proportion on CIDEr 10.
- (For Trans-base+RL, it has higher CIDEr than Trans-base, but it still get lower fraction of CIDEr 10.)







Drop worst for CIDEr optimization?

$$L = -\frac{1}{M}\sum_{m=1}^{M}(R^m - b^m)\log P(c^m|I)$$

$$b^m = \frac{1}{M-1}\sum R^{\setminus m}$$

- Two cases when R^m-b^m will be zero:
  - All the samples are equally bad. (The ground truth is hard.)
  - All the samples are equally good. (The model is confident.)

Qualitative results

Models we will look at

Three individual (single) models trained with cross-entropy+drop-worst:

- LSTM
- Transformer base
- Transformer huge

Ensemble-CE (top CIDEr on test):

- The above three models
- plus another trans-huge and lstm model trained on another train-val split.

Ensemble-RL (top human rating, 2nd CIDEr on test):

- Same three models trained with CIDEr+RL

LSTM: **`a view of the lake`**

Trans-base: **`a city on the water`**

Trans-huge: **`reflections in the early morning`**

Ensemble-CE: **`reflections in the water on a cold winter morning`**

Ensemble-RL: **`a view of the lake in the winter`**

LSTM: **`the road through the forest`**

Trans-base: **`driving through a redwood forest`**

Trans-huge: **`a view of the forest`**

Ensemble-CE: **`a drive through a redwood forest`**

Ensemble-RL: **`a view of the trees in the forest`**

LSTM: **`a helicopter prepares to land`**

Trans-base: **`the amphibious assault ship arrives`**

Trans-huge: **`a helicopter takes off from ship`**

Ensemble-CE: **`a helicopter takes off from the flight deck of the amphibious assault ship`**

Ensemble-RL: **`a helicopter on the flight deck of the ship`**

LSTM: **a table full of food**

Trans-base: **a table full of food**

Trans-huge: **the art of wedding photography**

Ensemble-CE: **breakfast in bed with a dog**

Ensemble-RL: **a woman with her dog at the table**

LSTM: **football player makes a save during the match**

Trans-base: **football player scores the first goal for football team**

Trans-huge: **football player scores the opening goal**

ensemble: **football player scores his team 's first goal during the match**

ensemble: **football player scores his team 's second goal during the match**

LSTM: **`a view of the mountains`**

Trans-base: **`person working in the field`**

Trans-huge: **`the hills are alive with the sound of music`**

Ensemble-CE: **`the hills are alive with the sound of music`**

Ensemble-RL: **`person on the road in the field`**

LSTM: **`a model wears a creation during event`**

Trans-base: **`a model wears a creation as part of fashion collection presented`**

Trans-huge: **`person poses for a photo`**

Ensemble-CE: **`person poses for a photo with a fan before the start of the race`**

Ensemble-RL: **`a model walks the runway at the fashion show during event`**

Resources

- Code available on Github:

  https://github.com/ruotianluo/GoogleConceptualCaptioning

- Docker image: can use to deploy trained models

  Dockerhub: ruotianluo/conceptual_ens3

- Acknowledgement: tools to download the data

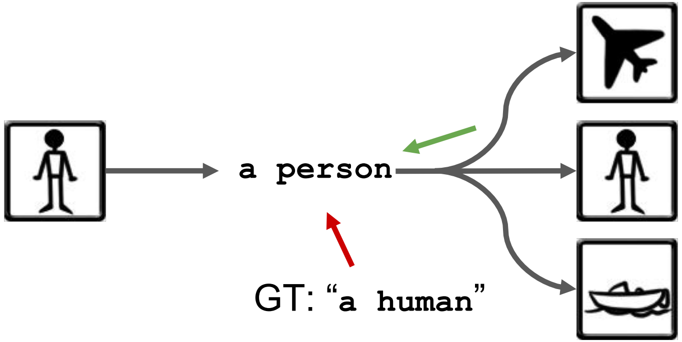
  https://github.com/igorbrigadir/DownloadConceptualCaptions

Additional thoughts

Discriminability objective

- Discriminative captions: allow us to identify the image by its caption

  (see our CVPR'18 paper)



- Did not explore for the challenge; may be useful even for the less grounded captions

Luo, Price, Cohen, Shakhnarovich, "Discriminability Objective for Training Descriptive Captions", CVPR 2018

Mixture models for captioning

- Since there are multiple types of captions in the data set representing *style* of captioning may be helpful

```
person in a gym with towel around neck
the front of the house with the wrap - around deck
mother and child : person was married until last year to ice hockey player
complete your look with a handbag , scarf and belt , and watch heads turn !
this image is described in surrounding text
author usually lets his subjects do the talking
```

- One could apply mixture models to get captions of different styles
- A related issue: diversity of captions
  - Here, can consider diversity of styles

Shen, Tianxiao, et al. "Mixture Models for Diverse Machine Translation: Tricks of the Trade." arXiv preprint arXiv:1902.07816 (2019).

(Yet another) alternative to ImageNet?

- Idea: Use conceptual captions as target to train a backbone CNN model from scratch
- Unlike classification labels, bounding boxes, segmentation masks: a more natural way of providing human supervision?
- Concerns:
    - Noisy ;
    - Arbitrary;
    - Probably expensive?