Describe That GIF

A GIF Description Generator

Presented By:
Chuma Kabaghe
ChumaK@Stanford.edu

MOTIVATION

Finding the right GIF is hard.

Despite their rising popularity, there is a surprising lack of scholarly work on animated GIFs in the computer vision community. We take a step towards improved GIF search and making GIFs more accessible by generating natural language descriptions of GIFs. We trained two models: CNN – LSTM and CNN - LSTM with attention and analyzed their performance on the TGIF dataset.

Results

Model	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE	CIDEr
Baseline (Rand LSTM-CNN)	49.7	27.2	14.5	5.2	13.6	36.6	7.6
CNN-LSTM	40.5	22.8	13.1	**7.4**	**14.8**	33.2	**14.5**
CNN-LSTM with Attention	37.4	18.0	8.7	4.5	13.3	26.6	12.3

The CNN-LSTM model performed the best, **outperforming the baseline** model by **90.8% (6.9) on CIDEr** score, **8.8% (1.2) on METEOR** score and **42% (2.2) on BLEU-4** score.

Data and Features

90,000 GIFs (from Tumblr) with descriptions
82% train, 5.5% dev, and 12.5% test
Training and validation set contain one reference description per GIF
Test set contains three reference descriptions per GIF

an elephant appears to be jumping on a trampoline.

a man in fancy attire is pointing at his bow tie.

a singer with tattoos on his arm is running his fingers through his hair.

a guy wearing a white shirt, brushes his hair back with his hand, as he sings into the microphone.

a guy is moving his hands through his hair while he sings.

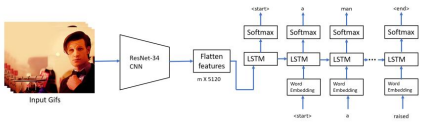
Encoder:

pretrained ResNet-34 Kinetics CNN

- Split GIF into individual frames and encoded every 10th frame
- Capped number of encodings at 10, resulting in a m X 5120 dimension vector which was our input into the LSTM models.

Models

CNN-LSTM Model



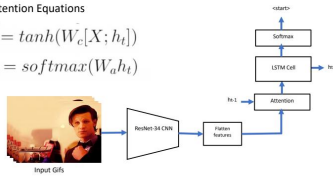
with sampling to minimize the cross entropy loss

CNN-LSTM Model with Attention

Attention Equations

$$\tilde{h}_t = tanh(W_c[X; h_t])$$

$$a_t = softmax(W_a h_t)$$



with sampling and attention to minimize the cross entropy loss

Hyperparameters

Model	batch size	hidden size	learning rate	layers	epochs	dropout	temperature
CNN-LSTM	1000	1024	0.001	1	100	0	0.8
CNN-LSTM with Attention	200	2048	0.001	1	100	0.2	0.8

Future Work

To improve model performance:
Train the CNN-LSTM model on complete GIFs without the 10 frame cap

Encode the GIFs using a deeper CNN model like the ResNext-101 trained on the Kinetics dataset

Experiment with different types of attention and visualize what the model is attending to during decoding.

Analysis

The model produces good, complete, grammatically correct descriptions with few <unk> tokens

GT: three teenagers are dancing on a glass stage.
CL: a group of boys are dancing in rhythm.

...but it is not perfect

GT: a **woman** with **glasses** talked and **moved her fingers**
CL: a **man** is talking to someone else **indoors**.

...maybe far from perfect?

GT: a girl is riding a skateboard down the street.
CL: a woman is holding a glass bong and exhaling smoke.

References

Yuncheng Li et al. "TGIF: A New Dataset and Benchmark on Animated GIF Description".
In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR). June2016

Thang Luong, Hieu Pham, and Christopher D. Manning. "Effective Approaches to Attention-based Neural Machine Translation".
In: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing.
Lisbon, Portugal: Association for
Computational Linguistics, Sept. 2015, pp. 1412–1421.