Retrieve, Caption, Generate: Visual Grounding for Enhancing Commonsense in Text Generation Models

Carnegie Mellon University Language Technologies Institute

UNIVERSITY OF WATERLOO

Steven Y. Feng[1], Kevin Lu[2], Zhuofu Tao[3], Malihe Alikhani[4], Teruko Mitamura[1], Eduard Hovy[1], Varun Gangal[1]

[1]Carnegie Mellon University, [2]University of Waterloo

[3]University of California Los Angeles, [4]University of Pittsburgh

Overview

Generative Commonsense Reasoning (CommonGen): produce logical sentences from keywords.

Challenges: Learn commonsense reasoning from images and transform into coherent sentences.

Baseline: BART-base, T5-base, Bart-large, T5-large.

Improvement: Obtain images from keywords and use image captions to guide sentence generation.

Results: Improve model performance and commonsense of generated sentences.

Dataset: CommonGen

The original **CommonGen** dataset is made up of 35,141 concept sets (consisting of 3 to 5 keywords each) and 79,051 sentences, split into train, dev, and test splits.

Concept set: a collection of objects / actions, for example: {*dog, frisbee, catch, throw*}

Human sentences (follow common sense):

- A **dog** leaps to **catch** a **thrown frisbee**.
- The **dog catches** the **frisbee** when the boy **throws** it.

Machine sentences (do not follow common sense):

- A **dog throws** a **frisbee** at a football player.
- Two **dogs** are **throwing frisbees** at each other.

Baseline Models

BART is a denoising autoencoder for pretraining sequence-to-sequence models. It is trained by (1) corrupting text with an arbitrary noising function, and (2) learning a model to reconstruct the original text.

T5 is a unified framework that converts all text-based language problems into a text-to-text format. It first trains a large text-to-text model, then transfers the learned model to other tasks.

Email: syfeng@andrew.cmu.edu

Proposed Improvements

We improve baseline models by **(1) retrieving images** of given concept sets from Google, **(2) captioning retrieved images** using a pretrained captioning model (MSCOCO), and **(3) generating sentences** given the concept set plus image captions (concatenated to the concept set) as input to the model. Since each image and caption differs in its quality and coverage of the input keywords, we try different numbers of captions for each example, a parameter called **Number of Top Captions (NTC)**. We try NTC = 1, 2, 3, 5, 7, 10.

Figure 1: Sample retrieved images

Fig. 1 (a)	concept set	{stand, hold, umbrella, street}
	baseline	A holds an umbrella while standing on the street
	caption	a woman walking down a street holding an umbrella
	VisCTG	a woman stands on a street holding an umbrella
Fig. 1 (b)	concept set	{food, eat, hand, bird}
	baseline	a hand of a bird eating food
	caption	a person holding a small bird in their hand
	VisCTG	a bird eats food from a hand

Qualitative Examples

Method	**Text**
Concept set	sit, chair, toy, hand (example 1)
Captions	a little girl sitting on a chair with a teddy bear <s> a small child sitting on a chair with a teddy bear <s> a young boy sitting on a chair with a skateboard
Baseline	hands sitting on a chair
VisCTG	A boy sitting on a chair with a toy in his hand.
Human	A baby sits on a chair with a toy in one of its hands.
Concept set	food, eat, hand, bird (example 2)
Captions	a bird is perched on a branch with a hand <s> a person holding a small bird in their hand
Baseline	hand of a bird eating food
VisCTG	A bird eats food from a hand.
Human	A small bird eats food from someone's hand.

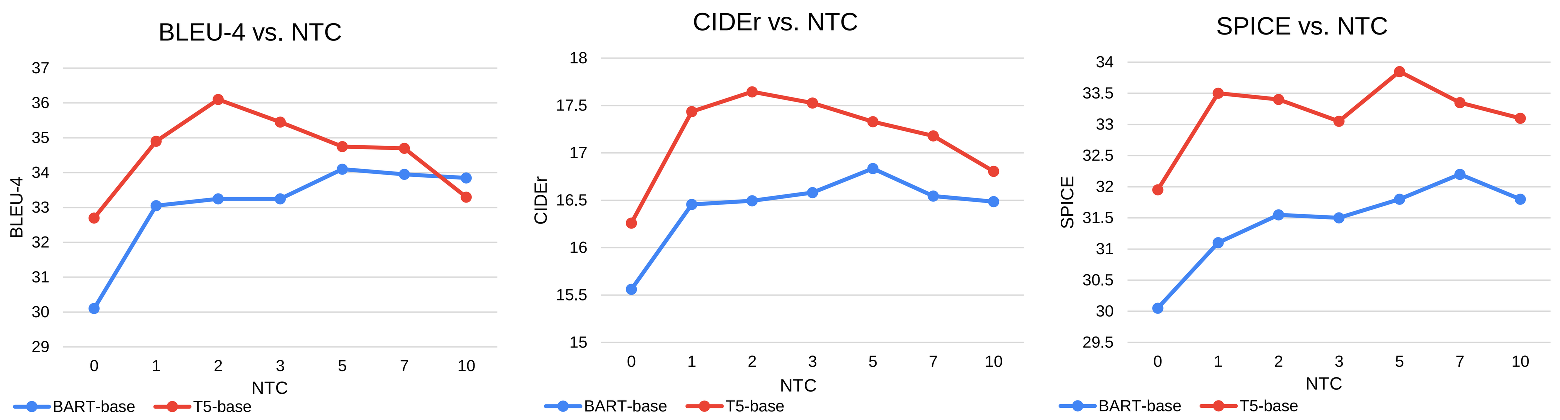
Evaluation Metrics

- BLEU, ROUGE, and METEOR measure similarity between the generated and human reference sentences, at a more token-level
- CIDEr captures a combo of sentence similarity, grammaticality, etc.
- SPICE maps text to semantic scene graphs and calculates an F-score over the graphs' tuples
- Coverage measures the average percentage of input concepts covered by the generated text

Results

Below are the metrics in comparison with other models. T5-base, T5-large, and BART-large refer to baselines reported in the original paper.

Models	ROUGE-2	**BLEU-4**	**CIDEr**	**SPICE**	Coverage
T5-base	14.63	18.54	9.40	19.87	76.67
T5-large	21.74	31.96	15.13	28.86	95.29
BART-large	22.02	29.01	13.98	28.00	97.35
EKI-BART	-	35.945	16.999	29.583	-
KG-BART	-	33.867	16.927	29.634	-
RE-T5	-	**40.863**	**17.663**	**31.079**	-
T5-base VisCTG	22.83	**34.722**	**16.173**	**28.808**	92.92
T5-large VisCTG	23.83	36.409	16.815	29.629	95.54
BART-base VisCTG	21.73	**32.291**	**15.187**	**27.403**	88.98
BART-large VisCTG	23.68	**36.939**	**17.199**	**29.973**	94.86



BLEU-4, CIDEr, and SPICE on test_{CG} over different values of NTC for BART-base and T5-base.

Future Work

- Improve image search and captioning, e.g. better selection of images or using a stronger captioning model.
- Video captioning and image generation can be explored.
- Extend VisCTG to other data-to-text NLG tasks, e.g. WebNLG.

GitHub: `https://github.com/styfeng/VisCTG`