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

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

AAAI 2022

What is Concept-to-Text Generation?

- Constrained text generation: produce natural language outputs under certain pre-conditions (e.g. particular words must appear in the outputs)
- Data-to-text NLG: produce natural language descriptions of structured or semi-structured data
- Common task formulation: set of inputs → natural language
  - Inputs can be thought of as concepts, e.g. higher-level words or structures that play an important role in the generated text
  - Think of these tasks as "concept-to-text generation"

Motivation for Visual Grounding

- Are there simple and effective approaches to improving performance on concept-to-text generation that comes from:

  Visual grounding or multimodal information in images?

- Large pretrained NLP models still struggle with commonsense tasks that humans can reason through easily[1]
- Hypothesis: commonsense information contained in modalities like vision beyond text that can be exploited
- VisCTG: Visually Grounded Concept-to-Text Generation

Generative Commonsense Reasoning

- AKA ***CommonGen*** task
- Generate logical sentences from given sets of input concepts
- Examples:
  - {horse, carriage, draw} → The carriage is drawn by the horse.
  - {listen, talk, sit} → The man told the boy to sit down and listen to him talk.

Figure 1: **An example of the dataset of CommonGen.** GPT-2, UniLM, BART and T5 are large pre-trained text generation models, *fine-tuned* on the proposed task.

Lin et al., 2020. CommonGen: A Constrained Text Generation Challenge for Generative Commonsense Reasoning. EMNLP 2020 Findings.

Why CommonGen?

- Difficult instance of concept-to-text generation that assesses:
    1. Relational reasoning abilities using commonsense knowledge
    2. Compositional generalization capabilities to piece together different/unseen concept combos
- Broadly applicable and encompassing task formulation and evaluation methodology
- Growing interest in the commonsense capabilities of NLP models

Dataset Splits and Baseline Models

- Created new dev, test splits (dev_{CG}, $test_{CG}$) from original dev set (dev_O) since original test set ($test_O$) is hidden. Training set ($train_{CG}$) was unaltered

Stats	$Train_{CG}$	Dev_O	$Test_O$	Dev_{CG}	$Test_{CG}$
# concept sets	32,651	993	1,497	240	360
# sentences	67,389	4,018	7,644	984	1583

- Baselines: trained 4 seq2seq Transformer models: BART-base, BART-large, T5-base, T5-large. Performance exceeded original reported scores

Thorough Baseline Analysis – Qualitative Study (1)

- Many baseline generations contain following issues:
  1. Lack commonsense and logic
     1. Improper ordering/piecing of sentence segments
        - “**body of water** <u>on a</u> **raft**”
     2. Does not understand what certain nouns can/cannot do
        - “A **dog checking his phone** on a pier”

Thorough Baseline Analysis – Qualitative Study (2)

- Many baseline generations contain following issues:

2. Not fluent or coherent, e.g. phrases and not full sentences
3. Missing important words such as nouns
    - "A **[?]** listening music and dancing in a dark room"
4. Generally generic and bland (dull response problem[2])
    - "**Someone** sits and listens to **someone** talk"

Motivation for Images and Captions (1)

- Images representing everyday scenarios prevalent for diff. concept sets
- E.g. searching "{cow, horse, lasso} → images of cowboys riding horses and lassoing cows, unlike baseline generation of "A cow is lassoing a horse."
- Everyday images similar to those in captioning datasets like MSCOCO, so pretrained captioning models should work well
- Textual corpora suffer from "reporting bias"[3]
  - Everyday things underrepresented compared to "newsworthy" things
  - Bias can be possibly dampened using visual data and models

Motivation for Images and Captions (2)

{stand, hold, umbrella, street}

baseline: A holds an umbrella while standing on the street
capt: a woman walking down a street holding an umbrella
VisCTG: A woman stands on a street holding an umbrella.

{food, eat, hand, bird}

baseline: hand of a bird eating food
capt: a person holding a small bird in their hand
VisCTG: A bird eats food from a hand.

{cat, bed, pet, lay}

baseline: A cat is laying on a bed and petting it.
capt: a cat laying on a bed with a stuffed animal
VisCTG: A cat laying on a bed being petted.

{fence, jump, horse, rider}

baseline: A rider jumps over a fence.
capt: a horse is jumping over a wooden fence
VisCTG: A rider jumps a fence on a horse.

Image Retrieval and Captioning

- Retrieve images for the concept sets in our three dataset splits
- Search engine is more generalizable and can cover more concept sets
- Google Images performs better compared to Bing and DuckDuckGo
  - Many input keywords not included and homonyms not handled well
- PyTorch-based implementation[4] of the FC image captioning model[5]
  - Image into deep CNN → caption generation via LSTM
  - Pretrained on the MSCOCO dataset with Resnet-101 image features

Caption Selection and Input Augmentation

- Captions $S_c = \{c_1, c_2, \ldots, c_n\}$ for each concept set are sorted by descending coverage to the concept set to obtain $S_{c'} = \{c_1', c_2', \ldots, c_n'\}$
- If two captions tied for coverage, kept in original order (by relevance)
- Retrieved images and captions cover fraction of concept set and quality varies → using multiple captions for generation may be better
- Try using different numbers of top captions within $S_{c'}$ – a parameter called Number of Top Captions (NTC); we try NTC = 1, 2, 3, 5, 7, 10
- Captions are used to augment the inputs to the models: {concept_set} <s> {caption_1} <s> {caption_2} ….

Augmented Input → Final Generation
wave fall board surfer <s> a surfer riding a wave on a surfboard → **A surfer is falling off his board into the waves.**
dance stage front crowd <s> a crowd of people watching a man on a stage <s> a man is holding a microphone in front of a crowd → **A man dances in front of a crowd on stage.**
stand hold umbrella street <s> a woman walking down a street holding an umbrella <s> a woman walking down a street holding an umbrella <s> a girl holding a pink umbrella in a city <s> a man holding an umbrella in a city <s> a group of people standing under a umbrella → **A group of people standing on a street holding umbrellas.**

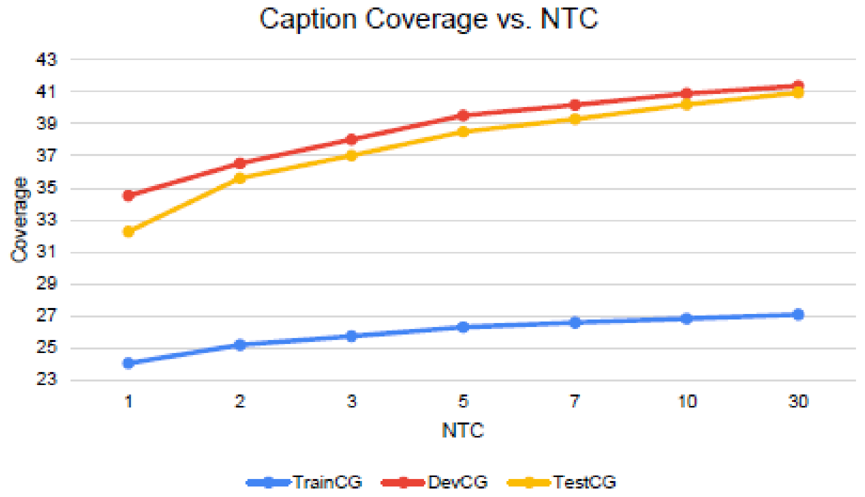

Figure 1: Graph displaying the average coverage (out of 100) by the top NTC captions in aggregate per concept set.

Experimental Setup

- Epochs with best ROUGE-2 score on the dev split are chosen for beam-search decoding on the test splits ($test_{CG}$ and $test_{O}$)
- NTC is a hyperparam; only best value per model is selected and reported
- Conduct two human evaluations: AMT and expert linguist
  - Pairwise comparison of VisCTG and baseline model outputs
  - AMT: choose which of the two has better "Overall Quality"
  - Expert linguist: "Overall Quality", "Commonsense Plausibility", and "Fluency"
  - Three options: O1 – VisCTG better, O2 – baseline better, O3 – both indistinguishable

Automatic Evaluation Results on test$_{CG}$

	BART-base ($NTC = 5$)			**BART-large** ($NTC = 2$)		
Metrics	Baseline	VisCTG	p-value	Baseline	VisCTG	p-value
ROUGE-1	43.96±0.03	**45.44**±0.08	1.58E-05	45.67±0.25	**46.91**±0.31	1.58E-05
ROUGE-2	17.31±0.02	**19.15**±0.21	1.58E-05	18.77±0.04	**20.36**±0.05	1.58E-05
ROUGE-L	36.65±0.00	**38.43**±0.07	1.58E-05	37.83±0.29	**39.23**±0.01	1.58E-05
BLEU-1	73.20±0.28	**75.65**±0.78	6.94E-05	74.45±0.21	**78.80**±0.28	6.94E-05
BLEU-2	54.50±0.14	**59.05**±0.07	6.94E-05	56.25±0.78	**61.60**±0.85	6.94E-05
BLEU-3	40.40±0.14	**44.90**±0.42	6.94E-05	42.15±0.49	**47.00**±0.71	6.94E-05
BLEU-4	30.10±0.14	**34.10**±0.57	3.82E-03	32.10±0.42	**36.25**±0.78	2.08E-04
METEOR	30.35±0.35	**31.95**±0.07	6.94E-05	31.70±0.14	**34.00**±0.14	6.94E-05
CIDEr	15.56±0.10	**16.84**±0.05	6.94E-05	16.42±0.09	**18.35**±0.13	6.94E-05
SPICE	30.05±0.07	**31.80**±0.28	6.94E-05	31.85±0.21	**34.60**±0.28	6.94E-05
BERTScore	59.19±0.32	**61.44**±0.02	1.58E-05	59.95±0.29	**62.85**±0.30	1.58E-05
Coverage	90.43±0.17	**90.66**±1.39	0.33*	94.49±0.53	**96.49**±0.24	1.58E-05
PPL	80.39±3.65	**72.45**±0.79	1.58E-05	80.37±4.51	**68.46**±5.90	1.58E-05

	T5-base ($NTC = 2$)			**T5-large** ($NTC = 1$)		
Metrics	Baseline	VisCTG	p-values	Baseline	VisCTG	p-values
ROUGE-1	44.63±0.13	**46.26**±0.07	1.58E-05	46.32±0.26	**46.93**±0.22	7.26E-04
ROUGE-2	18.40±0.14	**19.78**±0.30	1.58E-05	19.59±0.12	**20.01**±0.23	0.02
ROUGE-L	37.60±0.16	**38.91**±0.27	1.58E-05	39.20±0.21	**39.52**±0.43	0.06
BLEU-1	73.60±0.85	**76.80**±0.28	6.94E-05	77.55±0.35	**78.65**±0.21	4.65E-03
BLEU-2	57.00±0.71	**60.30**±0.28	6.94E-05	60.80±0.28	**61.55**±0.35	0.07
BLEU-3	42.75±0.49	**46.25**±0.64	6.94E-05	46.50±0.00	**47.10**±0.57	0.11*
BLEU-4	32.70±0.42	**36.10**±0.85	6.94E-05	36.20±0.14	**36.40**±0.28	0.21*
METEOR	31.05±0.49	**32.70**±0.00	6.94E-05	33.20±0.00	**33.65**±0.49	0.49*
CIDEr	16.26±0.25	**17.65**±0.02	6.94E-05	17.79±0.01	**17.94**±0.25	0.23*
SPICE	31.95±0.07	**33.40**±0.28	6.94E-05	33.90±0.42	**34.55**±0.21	0.03
BERTScore	61.40±0.34	**62.42**±0.17	1.58E-05	62.67±0.09	**62.72**±0.03	0.34*
Coverage	90.96±1.77	**94.48**±1.39	1.58E-05	94.40±0.02	**95.95**±0.45	1.58E-05
PPL	83.04±1.62	**77.50**±3.86	3.16E-05	81.78±4.63	**73.41**±4.32	1.58E-05

Trends of Automatic Metrics vs. NTC

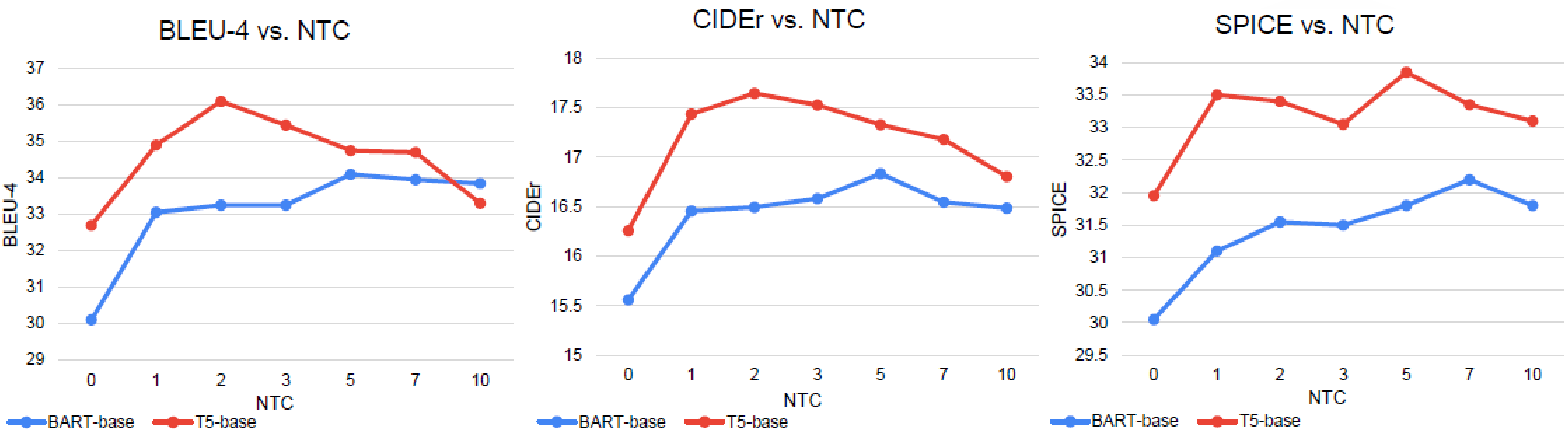

Figure 2: BLEU-4, CIDEr, and SPICE on test$_{CG}$ over different values of NTC for BART-base and T5-base.

Human Evaluation Results on test_{CG}

Model	O1	O2	O3	IAA
BART-base	**0.45**	0.33	0.22	0.72
BART-large	**0.62**	0.18	0.20	0.55
T5-base	**0.46**	0.33	0.21	0.72
T5-large	**0.46**	0.34	0.20	0.74

Table 9: Avg. AMT eval results on test_{CG} for *overall quality*. O1: VisCTG wins, O2: baseline wins, O3: both indistinguishable. Bold corresponds to higher fractional outcome between O1 and O2. All results are statistically significant based on paired two-tailed t-tests and $\alpha = 0.1$. The inter-annotator agreement (IAA) is the average direct fractional agreement (where both annotators choose O1 or O2) over all examples. See §5.2 and Appendix D for further details.

Model	Aspect	O1	O2	O3
BART-large	Overall	**0.44**	0.24	0.32
	Commonsense	**0.32**	0	0.68
	Fluency	**0.56**	0.12	0.32

Table 10: Avg. expert linguist eval results on test_{CG} for BART-large. O1: VisCTG wins, O2: baseline wins, O3: both indistinguishable. Bold corresponds to higher fractional outcome between O1 and O2 per aspect. See §5.2 and Appendix D for further details.

Automatic Evaluation Results on test$_O$

Models\Metrics	ROUGE-2/L		**BLEU-3/4**		METEOR	**CIDEr**	**SPICE**	Coverage
T5-base (reported baseline)	14.63	34.56	28.76	18.54	23.94	9.40	19.87	76.67
T5-large (reported baseline)	21.74	42.75	43.01	31.96	31.12	15.13	28.86	95.29
BART-large (reported baseline)	22.02	41.78	39.52	29.01	31.83	13.98	28.00	97.35
EKI-BART (Fan et al. 2020)	-	-	-	35.945	-	16.999	29.583	-
KG-BART (Liu et al. 2021)	-	-	-	33.867	-	16.927	29.634	-
RE-T5 (Wang et al. 2021)	-	-	-	**40.863**	-	**17.663**	**31.079**	-
T5-base VisCTG	22.83	44.98	45.749	**34.722**	31.809	**16.173**	**28.808**	92.92
T5-large VisCTG	23.83	45.76	47.376	36.409	33.012	16.815	29.629	95.54
BART-base VisCTG	21.73	43.43	43.235	**32.291**	30.86	**15.187**	**27.403**	88.98
BART-large VisCTG	23.68	45.07	48.031	**36.939**	33.215	**17.199**	**29.973**	94.86

Analysis of Results

- Both automatic and human evaluation results show that VisCTG greatly outperforms the baselines across all metrics and models
- Most outperforming model is BART-large, which is why we ask the expert linguist to evaluate BART-large on three aspects
- BART-large VisCTG outperforms EKI-BART[6] and KG-BART[7], two SOTA published CommonGen models that use external knowledge
- BLEU-4, CIDEr, and SPICE increase to a peak NTC value and taper off

Qualitative Analysis

Concept Set	{sit, chair, toy, hand}
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 <s> a man sitting on a chair with a remote
Baseline	hands sitting on a chair
VisCTG	A boy sitting on a chair with a toy in his hand.

Concept Set	{jacket, wear, snow, walk}
Captions	a young boy in a red jacket is standing in the snow <s> a man in a red jacket is standing in the snow
Baseline	walking in the snow wearing a furry jacket
VisCTG	A man is walking in the snow wearing a jacket.

Concept Set	{rock, water, stand, body}
Captions	a bird sitting on a rock in a body of water
Baseline	a body of water standing on rocks
VisCTG	A man standing on a rock near a body of water.

Concept Set	{bench, bus, wait, sit}
Captions	a man sitting on a bench with a book <s> a person sitting on a bench with a laptop
Baseline	A bus sits on a bench.
VisCTG	A man sits on a bench waiting for a bus.

Concept Set	{hold, hand, stand, front}
Captions	a man holding a pair of scissors in front of a wall
Baseline	Someone stands in front of someone holding a hand.
VisCTG	A man stands in front of a man holding a hand.

Concept Set	{bag, put, apple, tree, pick}
Captions	a person holding a apple in a tree <s> a bunch of apples are growing on a tree
Baseline	A man is putting apples in a bag and picking them up from the tree.
VisCTG	A man puts a bag of apples on a tree.

Conclusion and Future Work

- Explored the use of visual grounding for improving the commonsense of Transformer models for concept-to-text generation, calling our method VisCTG: Visually Grounded Concept-to-Text Generation
- Showed its effectiveness on the CommonGen task using BART and T5
- Can improve image search and captioning, e.g. stronger captioning model or better selection of images during retrieval
- Can explore video captioning and image generation rather than retrieval
- Can investigate VisCTG for other NLG tasks such as WebNLG

References

1. Talmor et al., 2020. oLMpics -- On what Language Model Pre-training Captures. TACL 2020.
2. Du and Black. 2019. Boosting Dialog Response Generation. ACL 2019.
3. Gordon and Van Durme. 2013. Reporting bias and knowledge acquisition. 2013 Workshop on Automated Knowledge Base Construction.
4. https://github.com/ruotianluo/self-critical.pytorch
5. Rennie et al., 2017. Self-Critical Sequence Training for Image Captioning. CVPR 2017.
6. Fan et al., 2020. An Enhanced Knowledge Injection Model for Commonsense Generation. COLING 2020.
7. Liu et al., 2021. KG-BART: Knowledge Graph-Augmented BART for Generative Commonsense Reasoning. AAAI 2021.

Thanks for Listening!

https://github.com/styfeng/VisCTG

https://arxiv.org/abs/2109.03892

- Steven Y. Feng: syfeng@cs.cmu.edu
  Website: https://styfeng.github.io/
  Twitter: @stevenyfeng
- Kevin Lu: kevin.lu1@uwaterloo.ca
  Website: https://kevin-lu.tech/
  Twitter: @KevinLu45010771
- Zhuofu Tao: z24tao@g.ucla.edu
  Website: https://www.linkedin.com/in/z24tao/
- Malihe Alikhani: malihe@pitt.edu
  Website: https://www.malihealikhani.com/
  Twitter: @malihealikhani
- Teruko Mitamura: teruko@cs.cmu.edu
- Eduard Hovy: hovy@cs.cmu.edu
- Varun Gangal: vgangal@cs.cmu.edu
  Website: https://vgtomahawk.github.io/
  Twitter: @VarunGangal