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] Language Technologies Institute, Carnegie Mellon University [2] University of Waterloo
[3] University of California, Los Angeles [4] School of Computing and Information, University of Pittsburgh
syfeng@cs.cmu.edu, kevin.lu1@uwaterloo.ca, z24tao@g.ucla.edu, malihe@pitt.edu
teruko@cs.cmu.edu, hovy@cs.cmu.edu, vgangal@cs.cmu.edu

Abstract

We investigate the use of multimodal information contained in images as an effective method for enhancing the commonsense of Transformer models for text generation. We perform experiments using BART and T5 on concept-to-text generation, specifically the task of generative commonsense reasoning, or *CommonGen*. We call our approach *VisCTG: Visually Grounded Concept-to-Text Generation*. VisCTG involves captioning images representing appropriate everyday scenarios, and using these captions to enrich and steer the generation process. Comprehensive evaluation and analysis demonstrate that VisCTG noticeably improves model performance while successfully addressing several issues of the baseline generations, including poor commonsense, fluency, and specificity.

1 Introduction

Transformer-based models have seen increasing popularity for NLP tasks and applications. This includes SOTA text generation models such as BART (Lewis et al. 2020) and T5 (Raffel et al. 2020). Larger corpora and better pretraining losses are major reasons driving these gains. However, despite increasing attention on the commonsense of models through works like COMET (Bosselut et al. 2019), studies have shown that even large pretrained models still struggle with commonsense tasks that humans can reason through very easily (Talmor et al. 2020). We believe that there is commonsense information in other modalities like vision, beyond what is reported (Gordon and Van Durme 2013) in text, which can possibly augment commonsense and enhance decision-making processes of text generation models.

In this paper, we show this is true by improving the performance of Transformer-based text generation models on concept-to-text generation using visual grounding, which we call *VisCTG: Visually Grounded Concept-to-Text Generation*. Concept-to-text generation is a high-level formulation of several constrained text generation and data-to-text natural language generation (NLG) tasks. These are challenging tasks that have seen increasing interest, and involve generating natural language outputs given certain pre-conditions, e.g. specific words in the outputs, and structured or semi-structured inputs. They typically involve converting a set of inputs into natural language text. These inputs can normally be thought of as *concepts*, or high-level words or structures, that play an important role in the generated text. Multimodal work has seen increasing popularity, but its exploration for constrained and data-to-text NLG has been limited (Baltrusaitis, Ahuja, and Morency 2019; Gao et al. 2020).[1]

{*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.

Table 1: Examples of retrieved images, captions, baseline and VisCTG (our model's) generations. The images and captions are used as an intermediary to guide the final generation and it need not be faithful to them. E.g. nobody is petting the cat in the image, but since the VisCTG output is conditioned on both the concept set and caption, it includes *being petted*.

We investigate the task of generative commonsense reasoning, or CommonGen (Lin et al. 2020), which involves generating sentences that effectively describe everyday scenarios from concepts sets, which are words that must appear in the output. CommonGen is challenging as effective relational reasoning ability using commonsense knowledge is required. Models must also possess the compositional generalization capabilities to piece together different concepts. CommonGen is an effective benchmark for constrained text generation and commonsense as its task formulation and evaluation methodology are rather broadly applicable.

We experiment on CommonGen using BART and T5. An

[1]Code: https://github.com/styfeng/VisCTG

Dataset Stats	Train$_{CG}$	Dev$_O$	Test$_O$	Dev$_{CG}$	Test$_{CG}$
# concept sets	32,651	993	1,497	240	360
size = 3	25,020	493	-	120	-
size = 4	4,240	250	747	60	180
size = 5	3,391	250	750	60	180

Table 2: Statistics of CommonGen dataset splits.

initial analysis (§3.1) of baseline generations shows several issues related to commonsense, specificity, and fluency. We hypothesize that these can be addressed through image captions (§3.2). Images representing everyday scenarios are commonplace, and typically logical and grounded in commonsense. Captioning models can also normally produce decent captions for everyday images, which can be used to guide and enhance the generation process. See Table 1 for examples.

Expounding on this, we use a pretrained image captioning model on MSCOCO captions (Lin et al. 2014) to caption the top retrieved images for each concept set (§4.1,4.2). We use these captions as additional information to augment inputs to our generation models (§4.3). Extensive evaluation (§6) demonstrates that VisCTG improves model performance and commonsense while addressing the baseline inadequacies.

2 Dataset, Models, and Metrics

2.1 CommonGen Dataset

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. Since the original test set is hidden, we partition the original dev set into new dev and test splits for the majority of our experiments. We do, however, ask the CommonGen authors to evaluate our best VisCTG models on the original test set (more in §6). The training set remains the same. We refer to the original dev and test sets as dev$_O$ and test$_O$, and these new splits as train$_{CG}$, dev$_{CG}$, and test$_{CG}$. Table 2 contains information about these splits. Their relative sizes and distribution of concept set sizes within each are kept similar to the originals.

2.2 Models: T5 and BART

We use pretrained text generation models T5 and BART, both the base and large versions. Both are seq2seq Transformer models. T5 has strong multitask pretraining. BART is pretrained as a denoising autoencoder to reproduce original from noised text. We use their HuggingFace implementations.

We train two seeded versions of each model on train$_{CG}$ and evaluate their performance on dev$_O$. These serve as the baselines for our experiments. Using the numbers in Lin et al. (2020) as comparison, we validate our implementations. We use the hyperparameters from Lin et al. (2020), beam search for decoding, and select the final epoch as the one reaching maximum ROUGE-2 (Lin and Hovy 2003) on the dev split. From Table 3, we observe that our re-implementations reach or exceed reported results in Lin et al. (2020) on most metrics.

2.3 Evaluation Metrics

We use several evaluation metrics, including those in Lin et al. (2020) such as BLEU (Papineni et al. 2002), CIDEr (Vedantam, Lawrence Zitnick, and Parikh 2015), SPICE (Anderson et al. 2016), and coverage (cov). These (other than cov) assess similarity between human references and generations. In particular, CIDEr captures a combination of sentence similarity, grammaticality, saliency, importance, and accuracy. SPICE maps texts to semantic scene graphs and calculates an F-score over these graphs' tuples. Lin et al. (2020) note that SPICE correlates highest with human judgment for CommonGen. Cov measures the average percentage of input concepts covered by the output text in any form.

We also use BERTScore (Zhang et al. 2019) and Perplexity (PPL). BERTScore measures BERT (Devlin et al. 2019) embeddings similarity between individual tokens, serving as a more semantic rather than surface-level similarity measure. We multiply by 100 when reporting BERTScore. PPL serves as a measure of fluency, with lower values representing higher fluency. We use GPT-2 (Radford et al. 2019) for PPL. For all metrics other than PPL, higher means better performance.

Model\Metrics	BLEU-4	CIDEr	SPICE
Reported BART-large	27.50	14.12	30.00
Reported T5-base	18.00	9.73	23.40
Reported T5-Large	30.60	15.84	31.80
Our BART-base	28.30	15.07	30.35
Our BART-large	**30.20**	**15.72**	**31.20**
Our T5-base	**31.00**	**16.37**	**32.05**
Our T5-large	**33.60**	**17.02**	**33.45**

Table 3: Comparing dev$_O$ performance of our re-implemented models to those in Lin et al. (2020). Bold represents where we reach/exceed reported numbers. Results averaged over two seeds for our models. Lin et al. (2020) did not report BART-base. See Appendix A for all metrics.

3 Initial Analysis and Motivation

3.1 Baseline Model Generations

We conduct an initial analysis of the baseline model outputs, and observe that several lack fluency. Some are more like phrases than full coherent sentences, e.g. *"body of water on a raft"*. Others miss important words, e.g. *"A listening music and dancing in a dark room"* misses a noun before *listening*. A large portion of generations are generic and bland, e.g. *"Someone sits and listens to someone talk"*. This may be an instance of the *dull response problem* faced by generation models (Du and Black 2019; Li et al. 2016), where they prefer safe and frequent responses independent of the input.

Many generations also lack commonsense. For example, *"body of water on a raft"* is illogical as the phrases *"body of water"* and *"a raft"* are pieced together incorrectly. A similar issue occurs with the {*horse, carriage, draw*} example in Table 4. At times the models also cannot understand what certain nouns can do, e.g. *"A dog checking his phone on a pier."* Several other examples of this can be found in Table 4.

3.2 Images and Captions

Images that represent everyday scenarios are quite prevalent for almost any reasonable concept set. Further, the images are typically grounded in commonsense. For example, searching

Concept Set	Baseline Generation	Human Reference
{horse, carriage, draw}	horse drawn in a carriage	The carriage is drawn by the horse.
{dog, house, eat}	A dog eats hay in a house	The dog eats food inside the house.
{cow, horse, lasso}	A cow is lassoing a horse.	A group of men riding horses lassoing a cow.

Table 4: Example generations from our baseline models versus human references.

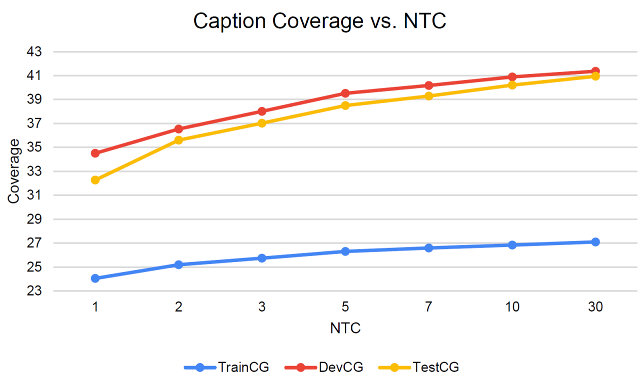

Figure 1: Graph displaying the average coverage (out of 100) by the top NTC captions in aggregate per concept set.

{*cow, horse, lasso*} will result in many images of cowboys riding horses and lassoing cows, rather than the illogical situation of *"A cow is lassoing a horse."* described by the baseline generation in Table 4. Many everyday images are relatively similar to those in image captioning datasets such as MSCOCO, so pretrained captioning models should work quite effectively. We thus hypothesize that using images and their captions to visually ground concept-to-text generation can potentially deal with issues mentioned in 3.1. Retrieved images with corresponding captions generated by a pretrained image captioning model (see §4.2) and final baseline and VisCTG generations for select concept sets are in Table 1.

Textual corpora also suffer from *reporting bias* (Gordon and Van Durme 2013), where everyday, commonsense albeit "uninteresting" actions (walking), objects (bench) and facts (bananas are yellow) are underrepresented compared to real-world frequency, while "newsworthy" actions (murdering), objects (spaceships) and facts (blue GMO bananas) are exaggerated. This seeps into large pretrained text models (Shwartz and Choi 2020). Using visual data and models dampens this bias, likely improving the commonsense of generations.

4 Methodology

4.1 Image Retrieval

We first obtain images for each concept set in our three splits. Image captioning datasets such as MSCOCO and Flickr are typically too small and focused to be effective for our purposes since we must cover numerous different concept sets. Further, a search engine is more generalizable.

We decide to use Google Images. On a sample of concept sets, the retrieved images using other search engines were inappropriate; they did not incorporate most input keywords nor handle homonyms well. For example, *"sports+fan+watch"* yields images of fans watching a sports game on Google images, but images of hand watches on Bing and DuckDuckGo.

We queried input concept sets by concatenating keywords with plus signs (+), and used *simple-image-scraper*[2] to obtain URLs of the top 30 results. The image was scraped only if the URL ended in *.png*, *.jpeg*, *.jpg*, or *.gif*. The received content was verified to be valid images using *pillow*[3], otherwise skipped. Retrieved images were typically of high quality and corresponded well to the concepts. See Table 1 for examples.

4.2 Image Captioning

After retrieving images, we use a PyTorch-based implementation[4] of the FC image captioning model (Luo et al. 2018; Rennie et al. 2017), which generates a caption via an LSTM initialized with a pseudo token obtained by feeding the image into a deep CNN followed by a linear projection. We use a pretrained FC model trained on the MSCOCO dataset with pretrained Resnet-101 image features.[5] As most of our retrieved images represent everyday scenarios and are relatively similar to those in MSCOCO, the pretrained model performs quite well. See example captions in Table 1.

4.3 Caption Selection and Input Augmentation

After we have captions $S_c = \{c_1, c_2, ..., c_n\}$ for each concept set in all three splits, we reorder them by descending coverage to the concept set to obtain $S_{c'} = \{c'_1, c'_2, ..., c'_n\}$. If two captions are tied for coverage, we keep them in their original search result order. This allows us to select the captions that have highest coverage and are most relevant.

Since most retrieved images and corresponding captions cover only a fraction of the entire concept set, and the quality of each varies, we hypothesize that using multiple captions for generation may lead to more robust and higher-quality outputs with more coverage. The models may learn to piece together information from caption(s) while generating final texts. Hence, we try experiments using different numbers of top captions within $S_{c'}$, a parameter we call NTC (Number of Top Captions). We try $NTC = 1, 2, 3, 5, 7, 10$, and do not go above $NTC = 10$ as Figure 1 shows that coverage gains from $10 \rightarrow 30$ are minor. Figure 1 also illustrates that captions have relatively low individual coverage, especially compared with outputs from models trained on CommonGen, which is why we do not use them as a baseline.

The captions are concatenated together and onto the concept set using <s> separator tokens. These serve as augmented inputs to BART and T5. They learn to convert these

[2]https://pypi.org/project/simple-image-download/

[3]https://pypi.org/project/Pillow/

[4]https://github.com/ruotianluo/self-critical.pytorch

[5]See Appendix B for further captioning model details.

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.**

Table 5: Examples of augmented inputs and final generations for varying values of NTC.

augmented inputs to human references during training, and are fed the augmented inputs (corresponding to the value of NTC) during validation and testing. Some examples of augmented inputs and generations can be found in Table 5.

5 Experiments

5.1 Model Training and Selection

For training VisCTG models, we mainly follow baseline hyperparameters, barring learning rate (LR) which is tuned per NTC value, and the maximum encoder length which is chosen depending on the tokenizer and value of NTC to ensure the entire input sequence can fit onto the encoder. We train two seeds per model. See Appendix C for more details.

For each model, we choose the epoch corresponding to highest ROUGE-2 on dev$_{CG}$, and use beam search for decoding. NTC itself is a hyperparameter, so while we train separate versions of each model corresponding to different NTC values, the final chosen models correspond to the NTC values that performed best on dev$_{CG}$ when averaged over both seeds. We then use the final chosen models to generate on both test$_{CG}$ and test$_{O}$, and report the results in §6.

5.2 Human Evaluation

We conduct two human evaluations: one using Amazon Mechanical Turk (AMT), and one using an expert linguist.[6] For the AMT study, we ask annotators to evaluate 86 test$_{CG}$ examples per model. Our evaluation is based on pairwise comparison of VisCTG and baseline model outputs. We ask human annotators to choose which amongst the two outputs (presented in a random order per example) has better *Overall Quality*. There are 3 choices - O1: VisCTG is better, O2: baseline is better, O3: both are indistinguishable. To aggregate multiple annotations per example, we find the fraction of responses towards each outcome value as the per-example distribution. We then find the sample mean of this outcome distribution over all examples. For sample mean and significance testing, we are interested in the values for O1 vs. O2.

For the expert linguist study, our expert is a native English speaker with a graduate degree in linguistics from a North American university. The expert is asked to annotate three aspects for 50 BART-large[7] test$_{CG}$ examples - *Overall Quality (Overall)*, *Commonsense Plausibility (Commonsense)*, and *Fluency (Fluency)*. For all aspects, we have a pairwise-comparison evaluation setup similar to that for AMT.

[6]See Appendix D for further human evaluation details.
[7]Since this is the best performing VisCTG model - see §6.

6 Results and Analysis

Automatic evaluation results on test$_{CG}$ are in Tables 6 and 7, and results on test$_{O}$ in Table 8.[8] Graphs displaying BLEU-4, CIDEr, and SPICE (the metrics on the CommonGen leaderboard[9]) on test$_{CG}$ over different NTC values are in Figure 2. Human evaluation results on test$_{CG}$ are in Tables 9 and 10. Optimal NTC values for BART-base, BART-large, T5-base, and T5-large are 5, 2, 2, and 1, respectively. These are the VisCTG results reported in the aforementioned tables. Table 11 contains qualitative examples, with more in Appendix E.

6.1 Analysis of Automatic Evaluation Results

We see from Tables 6 and 7 that VisCTG outperforms the baselines on all metrics across the models on test$_{CG}$. Performance gains are strong and statistically significant for BART-base, BART-large, and T5-base. VisCTG appears relatively less effective for T5-large which is the strongest baseline, and hence improving its performance may be more difficult.

From Table 8, we see that VisCTG models substantially outperform corresponding baselines reported in Lin et al. (2020) on test$_{O}$. T5-base VisCTG outperforms the reported T5-base and large baselines across metrics, and BART-base VisCTG performs similarly to the reported BART-large baseline. BART-large VisCTG outperforms the reported baseline, EKI-BART (Fan et al. 2020), and KG-BART (Liu et al. 2021). These are SOTA published CommonGen BART models that use external knowledge from corpora and KGs. We show that visual grounding is more effective, and BART-large VisCTG places high on the leaderboard.[9] T5-large VisCTG outperforms the reported baseline, but lags behind SAPPHIRE (Feng et al. 2021b) and RE-T5 (Wang et al. 2021).

Figure 2 shows that as NTC increases, metrics increase to a peak and taper off after. As we saw in Figure 1, the rate of increase of coverage declines with larger NTC. The latter images and captions are thus of diminishing quality, and hence using too many negatively affects model performance.

We also computed ROUGE between captions and outputs over test$_{CG}$. $ROUGE1/2/L = 36.2/12.3/33.5$ are modestly valued. Our models do not simply copy caption content.

6.2 Analysis of Human Evaluation Results

Table 9 shows that VisCTG outperforms the baseline on all four models based on human annotators (with high IAA). Annotators, on average, prefer VisCTG outputs over baseline outputs on overall quality, especially for BART-large. Table

[8]Evaluated by the CommonGen authors on their hidden test set.
[9]https://inklab.usc.edu/CommonGen/leaderboard.html

	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

Table 6: Auto eval results for BART on test$_{CG}$ over two seeds. Bold corresponds to best performance. We include p-values (from Pitman's permutation test (Pitman 1937)) for VisCTG compared to the baseline. Insignificant ones ($\alpha = 0.1$) marked with *.

	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

Table 7: Auto eval results for T5 on test$_{CG}$ over two seeds. Bold corresponds to best performance. We include p-values (from Pitman's permutation test (Pitman 1937)) for VisCTG compared to the baseline. Insignificant ones ($\alpha = 0.1$) marked with *.

10 illustrates that VisCTG outperforms the baseline model for BART-large based on an expert linguist's perspective. VisCTG outputs are highly preferred, on average, over the baseline on all three aspects of overall quality, commonsense, and fluency. This aligns with our automatic results in §6.1.

6.3 Qualitative Analysis

Table 11 shows several baseline outputs that contain issues from §3.1, e.g. incomplete and/or illogical sentences. Human references are all fluent and logical. VisCTG can usually generate much higher-quality text than the baselines.

The baseline outputs for ex. 1-2 are phrases lacking arguments, and all illogical for ex. 1-3. Using captions, VisCTG successfully adjusts semantic roles of entities, replaces incorrect subjects, fixes dependency structure, and grounds generations in commonsense. For ex. 1, captions are of the form "{X} *sitting on a chair with* {Y}", where {X} is a subject and {Y} an object. VisCTG output has similar structure, being fluent and logical with higher coverage. The baseline output also has an incorrect subject of *"hands"*. Our VisCTG output contains an additional entity (not present in the input set) of *"boy"* as subject, likely since it is a subject in the captions. This highlights the usefulness of visual grounding, as the image space can provide additional commonsense information not present in the text (e.g. toys are associated with children/boys). For ex. 2, the baseline output treats *"hand of a bird"* as a single entity, the subject. Captions separate *"bird"* and *"hand"* into two, likely guiding the VisCTG output to do so. For ex. 3, the baseline misplaces *"bus"* as subject. Captions are of form "{X} *sitting on a bench* {Y}", where {X} is a logical subject and {Y} is an expression. The VisCTG output has this structure, with correct subject and commonsense, and higher coverage. Overall, we see that visual grounding guides the model to learn which nouns/subjects can perform which actions (e.g. *"hands"* cannot sit on a chair but a *"boy"* can), which is a major baseline deficiency discussed in §3.1.

For ex. 4, the baseline output lacks a subject that the captions contain, likely guiding the VisCTG output to contain one: *"a man"*. For ex. 5, the baseline output is generic due to uses of *"someone"*. VisCTG's output is more specific and refers to *"man"*, likely because the caption (though not very fitting) includes a *"man"* subject. Even for captions that fit the concepts less, structure and fluency can still be exploited.

Overall, we see that the baselines simply try to piece to-

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	-
SAPPHIRE (T5-large) (Feng et al. 2021b)	-	-	-	37.119	-	16.901	29.751	-
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

Table 8: Auto eval results of VisCTG on test$_O$, evaluated by CommonGen authors. We compare to reported baseline numbers in Lin et al. (2020) (they did not evaluate BART-base), and models on their leaderboard[9] with publications at time of writing. Their leaderboard reports BLEU-4, CIDEr, and SPICE. Bold corresponds to best performance (for those three) per model type+size.

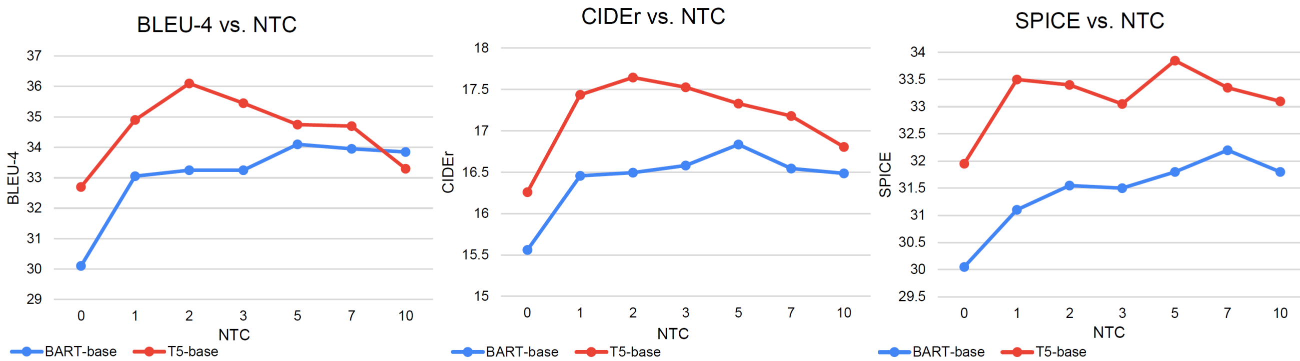

Figure 2: BLEU-4, CIDEr, and SPICE on test$_{CG}$ over different values of NTC for BART-base and T5-base.

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: indistinguishable. All results are stat sig based on paired two-tailed t-tests and $\alpha = 0.1$. Inter-annotator agreement (IAA) is the avg. direct fractional agreement (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: indistinguishable. See §5.2 and Appendix D for further details.

gether the input concepts into a form of English syntax, often failing to do so effectively. VisCTG models can produce more grammatical, fluent, and logical text by exploiting the syntactic and dependency structures of the captions. Further, the visual grounding improves the commonsense of the generations. The images inherently capture commonsense by representing everyday scenarios, and this commonsense info is rarely explicitly included in text. Hence, large text-based models such as our baselines tend to not know this info, whereas VisCTG models learn it through the grounding.

VisCTG is, however, imperfect. For ex. 6, its output is less logical and lower coverage than the baseline's. The captions are all simplistic and low coverage; the first is illogical, and some others are of the form *"a bunch of apples* {...} *on a tree"*, likely negatively impacting the generation. Ex. 4's human reference is creative, which is an area where VisCTG still lacks in comparison. For ex. 5, while VisCTG edits *"someone"* to *"man"*, it is unable to merge the two instances of *"man"* or adjust the sentence to be more coherent. These weaknesses are likely because captions tend to be simplistic (due to the captioning model's training data), limiting VisCTG's ability to make heavier edits. VisCTG, unsurprisingly, appears to depend quite heavily on the captions, and hence the quality of the images and captioning model.

7 Related Work

Constrained Text Generation: There have been several works on constrained text generation. Miao et al. (2019) use Metropolis-Hastings sampling to determine Levenshtein edits per generation step. Feng, Li, and Hoey (2019) pro-

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 <s> a man sitting on a chair with a remote
BART-base-BL	hands sitting on a chair
BART-base-VisCTG	A boy sitting on a chair with a toy in his hand.
Human reference	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
BART-large-BL	hand of a bird eating food
BART-large-VisCTG	A bird eats food from a hand.
Human reference	A small bird eats food from someone's hand.
Concept set	{bench, bus, wait, sit} (example 3)
Captions	a man sitting on a bench with a book <s> a person sitting on a bench with a laptop
T5-base-BL	A bus sits on a bench.
T5-base-VisCTG	A man sits on a bench waiting for a bus.
Human reference	The man sat on the bench waiting for the bus.
Concept set	{jacket, wear, snow, walk} (example 4)
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
BART-large-BL	walking in the snow wearing a furry jacket
BART-large-VisCTG	A man is walking in the snow wearing a jacket.
Human reference	Jamie took a walk out into the snow with only a T shirt on and instantly went back inside to wear his jacket.
Concept set	{hold, hand, stand, front} (example 5)
Captions	a man holding a pair of scissors in front of a wall
T5-large-BL	Someone stands in front of someone holding a hand.
T5-large-VisCTG	A man stands in front of a man holding a hand.
Human reference	A man stands and holds his hands out in front of him.
Concept set	{bag, put, apple, tree, pick} (example 6)
Captions	a person holding a apple in a tree <s> a bunch of apples are growing on a tree <s> a close up of a green apple with a tree <s> a bunch of apples are growing on a tree
BART-base-BL	A man is putting apples in a bag and picking them up from the tree.
BART-base-VisCTG	A man puts a bag of apples on a tree.
Human reference	I picked an apple from the tree and put it in my bag.

Table 11: Qualitative examples for test$_{CG}$. *BL* stands for baseline. *Concept set* refers to the input keywords and *Captions* refers to the captions (separated by <s>) used by the VisCTG model for that particular example to produce its final generation.

pose Semantic Text Exchange to adjust topic-level text semantics. Gangal et al. (2021) introduce narrative reordering (NAREOR) to edit the temporality of narratives.

Data-to-text NLG: E2E-NLG (Dušek, Novikova, and Rieser 2018) and WebNLG (Gardent et al. 2017) are two popular NLG benchmarks with structured inputs - meaning representation (MR) and triple sequences, respectively.

Commonsense Injection and Incorporation: One commonsense knowledge graph (KG) is COMET, trained on KG edges. EKI-BART (Fan et al. 2020) and KG-BART (Liu et al. 2021) use external knowledge to improve CommonGen performance. Distinctly, VisCTG uses visual grounding and shows higher performance (see §6). Visual Commonsense Reasoning (VCR) (Zellers et al. 2019) involves answering commonsense-related MC questions about images. Our work focuses on injecting commonsense into text generators.

Multimodal Machine Learning and NLP: There has been more work on multimodality, in areas like representation and video captioning, but little for constrained and data-to-text NLG (Baltrusaitis, Ahuja, and Morency 2019; Gao et al. 2020). There is work on pretrained multimodal models like ViLBERT (Lu et al. 2019): mainly encoders that jointly represent images and text rather than seq2seq models. Further, unlike these models which are pretrained, VisCTG exploits per-example visual info to fix specific issues per concept set.

8 Conclusion and Future Work

In conclusion, we motivated and explored the use of visual grounding for improving the commonsense of Transformer models for text generation. We investigated this for concept-to-text generation, calling our method VisCTG: Visually Grounded Concept-to-Text Generation. Extensive experiments on BART and T5 showed its efficacy on the CommonGen task. Comprehensive evaluation and analysis showed that VisCTG boosts model performance and commonsense while addressing baseline deficiencies. Potential future work includes using a stronger captioning model, e.g. one based on CLIP (Radford et al. 2021). Video captioning and image generation can also be explored. Further, VisCTG can be investigated for other data-to-text NLG tasks, e.g. WebNLG, and applications like data augmentation for text generation (Feng et al. 2020, 2021a), and enhancing the commonsense reasoning of personalized dialogue agents (Li et al. 2020).

References

Anderson, P.; Fernando, B.; Johnson, M.; and Gould, S. 2016. Spice: Semantic propositional image caption evaluation. In *European conference on computer vision*, 382–398. Springer.

Baltrusaitis, T.; Ahuja, C.; and Morency, L.-P. 2019. Multimodal Machine Learning: A Survey and Taxonomy. *IEEE Trans. Pattern Anal. Mach. Intell.*, 41(2): 423–443.

Bosselut, A.; Rashkin, H.; Sap, M.; Malaviya, C.; Celikyilmaz, A.; and Choi, Y. 2019. COMET: Commonsense Transformers for Automatic Knowledge Graph Construction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 4762–4779. Florence, Italy: Association for Computational Linguistics.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 4171–4186. Minneapolis, Minnesota: Association for Computational Linguistics.

Du, W.; and Black, A. W. 2019. Boosting Dialog Response Generation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 38–43. Florence, Italy: Association for Computational Linguistics.

Dušek, O.; Novikova, J.; and Rieser, V. 2018. Findings of the E2E NLG Challenge. In *Proceedings of the 11th International Conference on Natural Language Generation*, 322–328. Tilburg University, The Netherlands: Association for Computational Linguistics.

Fan, Z.; Gong, Y.; Wei, Z.; Wang, S.; Huang, Y.; Jiao, J.; Huang, X.; Duan, N.; and Zhang, R. 2020. An Enhanced Knowledge Injection Model for Commonsense Generation. In *Proceedings of the 28th International Conference on Computational Linguistics*, 2014–2025. Barcelona, Spain (Online): International Committee on Computational Linguistics.

Feng, S. Y.; Gangal, V.; Kang, D.; Mitamura, T.; and Hovy, E. 2020. GenAug: Data Augmentation for Finetuning Text Generators. In *Proceedings of Deep Learning Inside Out (DeeLIO): The First Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*, 29–42. Online: Association for Computational Linguistics.

Feng, S. Y.; Gangal, V.; Wei, J.; Chandar, S.; Vosoughi, S.; Mitamura, T.; and Hovy, E. 2021a. A Survey of Data Augmentation Approaches for NLP. *ACL 2021 Findings*.

Feng, S. Y.; Huynh, J.; Narisetty, C. P.; Hovy, E.; and Gangal, V. 2021b. SAPPHIRE: Approaches for Enhanced Concept-to-Text Generation. In *Proceedings of the 14th International Conference on Natural Language Generation*, 212–225. Aberdeen, Scotland, UK: Association for Computational Linguistics.

Feng, S. Y.; Li, A. W.; and Hoey, J. 2019. Keep Calm and Switch On! Preserving Sentiment and Fluency in Semantic Text Exchange. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2701–2711.

Gangal, V.; Feng, S. Y.; Hovy, E.; and Mitamura, T. 2021. NAREOR: The Narrative Reordering Problem. *arXiv preprint arXiv:2104.06669*.

Gao, J.; Li, P.; Chen, Z.; and Zhang, J. 2020. A Survey on Deep Learning for Multimodal Data Fusion. *Neural Computation*, 32(5): 829–864.

Gardent, C.; Shimorina, A.; Narayan, S.; and Perez-Beltrachini, L. 2017. The WebNLG challenge: Generating text from RDF data. In *Proceedings of the 10th International Conference on Natural Language Generation*, 124–133.

Gordon, J.; and Van Durme, B. 2013. Reporting bias and knowledge acquisition. In *Proceedings of the 2013 workshop on Automated knowledge base construction*, 25–30.

Lewis, M.; Liu, Y.; Goyal, N.; Ghazvininejad, M.; Mohamed, A.; Levy, O.; Stoyanov, V.; and Zettlemoyer, L. 2020. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 7871–7880. Online: Association for Computational Linguistics.

Li, A. W.; Jiang, V.; Feng, S. Y.; Sprague, J.; Zhou, W.; and Hoey, J. 2020. ALOHA: Artificial Learning of Human Attributes for Dialogue Agents. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05): 8155–8163.

Li, J.; Galley, M.; Brockett, C.; Gao, J.; and Dolan, B. 2016. A Diversity-Promoting Objective Function for Neural Conversation Models. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 110–119. San Diego, California: Association for Computational Linguistics.

Lin, B. Y.; Zhou, W.; Shen, M.; Zhou, P.; Bhagavatula, C.; Choi, Y.; and Ren, X. 2020. CommonGen: A Constrained Text Generation Challenge for Generative Commonsense Reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, 1823–1840. Online: Association for Computational Linguistics.

Lin, C.-Y.; and Hovy, E. 2003. Automatic evaluation of summaries using n-gram co-occurrence statistics. In *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, 150–157.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft COCO: Common Objects in Context. In Fleet, D.; Pajdla, T.; Schiele, B.; and Tuytelaars, T., eds., *Computer Vision – ECCV 2014*, 740–755. Cham: Springer International Publishing. ISBN 978-3-319-10602-1.

Liu, Y.; Wan, Y.; He, L.; Peng, H.; and Yu, P. S. 2021. KG-BART: Knowledge Graph-Augmented BART for Generative Commonsense Reasoning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(7): 6418–6425.

Lu, J.; Batra, D.; Parikh, D.; and Lee, S. 2019. ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks. In Wallach, H.; Larochelle, H.; Beygelzimer, A.; d'Alché-Buc, F.; Fox, E.; and Garnett, R., eds., *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Luo, R.; Price, B.; Cohen, S.; and Shakhnarovich, G. 2018. Discriminability Objective for Training Descriptive Captions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Miao, N.; Zhou, H.; Mou, L.; Yan, R.; and Li, L. 2019. Cgmh: Constrained sentence generation by metropolis-hastings sampling. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, 6834–6842.

Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W.-J. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, 311–318.

Pitman, E. J. 1937. Significance tests which may be applied to samples from any populations. *Supplement to the Journal of the Royal Statistical Society*, 4(1): 119–130.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; Krueger, G.; and Sutskever, I. 2021. Learning Transferable Visual Models From Natural Language Supervision.

Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; and Sutskever, I. 2019. Language models are unsupervised multitask learners. *OpenAI Blog*, 1(8): 9.

Raffel, C.; Shazeer, N.; Roberts, A.; Lee, K.; Narang, S.; Matena, M.; Zhou, Y.; Li, W.; and Liu, P. J. 2020. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *Journal of Machine Learning Research*, 21(140): 1–67.

Rennie, S. J.; Marcheret, E.; Mroueh, Y.; Ross, J.; and Goel, V. 2017. Self-Critical Sequence Training for Image Captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Shwartz, V.; and Choi, Y. 2020. Do Neural Language Models Overcome Reporting Bias? In *Proceedings of the 28th International Conference on Computational Linguistics*, 6863–6870.

Talmor, A.; Elazar, Y.; Goldberg, Y.; and Berant, J. 2020. oLMpics-On What Language Model Pre-training Captures. *Transactions of the Association for Computational Linguistics*, 8: 743–758.

Vedantam, R.; Lawrence Zitnick, C.; and Parikh, D. 2015. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4566–4575.

Wang, H.; Liu, Y.; Zhu, C.; Shou, L.; Gong, M.; Xu, Y.; and Zeng, M. 2021. Retrieval Enhanced Model for Commonsense Generation. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, 3056–3062. Online: Association for Computational Linguistics.

Zellers, R.; Bisk, Y.; Farhadi, A.; and Choi, Y. 2019. From Recognition to Cognition: Visual Commonsense Reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Zhang, T.; Kishore, V.; Wu, F.; Weinberger, K. Q.; and Artzi, Y. 2019. BERTScore: Evaluating Text Generation with BERT. In *International Conference on Learning Representations*.