A Overview

In the supplementary materials of this work, we first discuss the main limitation of CoNT. After that, we describe the detailed experimental setup for the 10 different benchmarks. Finally, we randomly present some generation examples of CoNT on machine translation and summarization.

B Limitation

CoNT practically does not have a negative impact on decoding speed. Compared with the decoding algorithm used in the inference stage of conventional generation models, the additional operations brought by CoNT are reflected in line 4 and line 5, Algorithm 1. The two operations can be efficiently calculated on GPUs. However, an obvious disadvantage of CoNT is the sacrifice of training efficiency. In general, the total training time of CoNT is about 2$\sim$4 times more than that of a MLE based model.

We show the pseudo code of our training procedure in Algorithm 2. As can be seen from this algorithm, there are three main factors that harm the training speed of CoNT: (i) a pre-train stage to ensure meaningful contrastive samples (line 3 in Algorithm 2), (ii) the involvement of token-by-token decoding in training (line 7 in Algorithm 2) which can hardly benefit from the parallel computing power of GPU, and (iii) the calculation of oracle measurement (the nested loop in line 9$\sim$11 Algorithm 2). Regrading the last issue, if we use some lexical matching metrics such as BLEU or ROUGE that always need calculated on CPU, the training speed will be obviously slowed down. During waiting for the results from CPU, the GPU utilization will decrease to 0. We solve the problem with by placing the nested loop on GPU where candidate samples and the ground truth are represented by a long type tensor and the similarity between two sequences are calculated by matrix multiplication which significantly improve the efficiency. The GPU version of getting oracle measurement will also be released along with our source code. As for the first issue and the second issue, we haven't found good alternatives yet so that we advise two ways to make a speed-accuracy trade-off at the following parts.

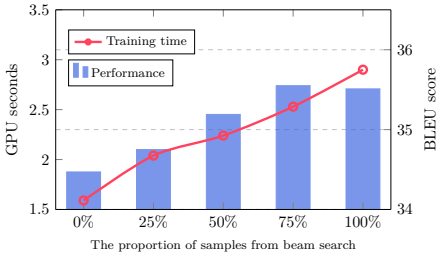
Algorithm 2 Contrastive Text Generation: Given a generation dataset $<\mathcal{X}, \mathcal{Y}>$, a randomly initialized encoder-decoder model $\mathcal{M} = (f, g)$; return a contrastive generation model.

1: **procedure** WARMUP($\mathcal{M}$)
2: Update the parameters of randomly initialized $\mathcal{M}$ with $\nabla_\theta \mathcal{L}_{nll}$ until convergence

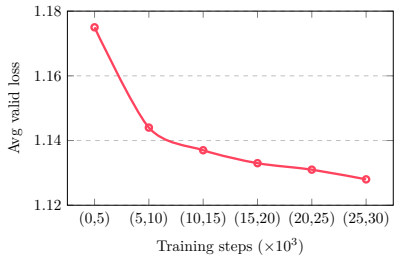
1: **procedure** BEAMSEARCH(g, $\mathbf{H}_X$, b) ▷ beam search algorithm
2: **return** Text, likelihood, logits of the b hypotheses

1: **procedure** TRAIN($\mathcal{M}$, $<\mathcal{X}, \mathcal{Y}>$)
2: $\theta \leftarrow$ Parameters of $\mathcal{M}$, $b \leftarrow$ beam size
3: WARMUP($\mathcal{M}$)
4: **while** not convergence **do**
5: $X^{1:k}, Y^{1:k} \leftarrow$ A minibatch of k datapoints from $<\mathcal{X}, \mathcal{Y}>$
6: $\mathbf{H}_X^{1:k} \leftarrow f\left(X^{1:k}\right)$, $\mathbf{H}_Y^{1:k} \leftarrow g\left(\mathbf{H}_X^{1:k}, Y^{1:k}\right)$ ▷ outputs from the encoder and decoder
7: ${Y'}^{1:k,1:b}, \mathbf{P}_{Y'}^{1:k,1:b}, \mathbf{H}_{Y'}^{1:k,1:b}$ = BEAMSEARCH$(g, \mathbf{H}_X^{1:k}, b)$
8: ${Y'}^{1:k,1:(b+k)} \leftarrow$ Append b self-generated samples to $Y^{1:k}$
9: **for** $i \in 1, 2, \ldots, k$ **do**
10: **for** $j \in 1, 2, \ldots, , b + k$ **do**
11: Do oracle measurement $o({Y'}^{i,j}, Y^i)$ for each element ${Y'}^{i,j}$ in ${Y'}^{1:k,1:(b+k)}$
$\mathcal{L}_{ctr} \leftarrow$ Get pair-wise contrastive loss
12: update parameters using $\nabla_\theta(\mathcal{L}_{nll} + \mathcal{L}_{ctr})$
13: **return** $\mathcal{M}$

Small Beam Size The first trick is to reduce the proportion of self-generated samples. With the increase of beam size, the time consumed by beam search will increase significantly. Remaining the maximum number of contrastive examples for each input unchanged, we can adjust the ratio of self-generated samples and from-batch samples. For IWSLT'14 En-De benchmark with transformer small, the default maximum number of contrastive samples is set to 32 and the proportion of self-generated samples is 75% (settings of other benchmarks are shown in Table 8). Relationship between the



(a) Relationship of GPU seconds (left axis) and BLEU (right axis) with the proportion of self-generated samples on on the test set of IWSLT'14 De-En. We set the maximum size of candidate samples to 32.



(b) Relationship between the contrastive loss on IWSLT'14 De-En validation set and training steps. We calculate the validation loss every 1k steps and report the average results every 5k steps.

Figure 6: Analysis on using speed-accuracy trade-off tricks

training time of each iteration and the proportion of samples returned by beam search on IWSLT'14 De-En benchmark can be seen in Figure 6a[2]. Totally using self-generated samples will double the training speed of the naive contrastive text generation method and will do not further boost the improvement in performance. Reducing the rate of self-generated samples to 50% still resulting in 1.0 BLEU superior to the baseline while saving about 1.0 GPU seconds per iteration compared with totally using self-generated samples.

Early Stop Another way to save training time is early stop. We can see in Figure 6b, on IWSLT'14 De-En, the declining trend of the contrastive loss on validation set allowing us to perform early stop in training. The contrastive loss on valdation set drop rapidly at the first 10k steps, and this decline will slow down in the following steps. In our experiments we train this model for about 40k steps, and early stop after 10k steps is also enough to improve the MLE baseline by 0.8 BLEU but saving 3/4 training time.

C Experimental Setup

In this section, we will introduce more details about experiments and datasets. Our experiments on IWSLT'14 De-En and WMT'14 En-De use fairseq [29] framework. Experiments on other datasets are run with transformers [49]. Table 7 gives an overview of the number of instances in train/validation/test set and its source. The test set of Totto [31] and CommonGen [22] is invisible. We get these results by submitting our generation results to the leaderboards[3][4].

Table 7: The statistics of datasets we use in experiments.

Datasets	**Train(#)**	**Validation(#)**	**Test(#)**	**Source(#)**
WMT'16 Ro-En	610k	2k	2k	Romanian-to-English translation
IWSLT'14 De-En	160k	7k	7k	Germanish-to-English translation
WMT'14 En-De	**4.5M**	3k	3k	English-to-Germanish translation
XSum [26]	204k	11k	11k	One-sentence summary of BBC news articles
Multi-News [8]	45k	5.6k	5.6k	Long summary of multiple news articles
Java [25]	165k	5k	11k	Code comment for java
Python [8]	252k	14k	15k	Code comment for Python
WikiBio [20]	582k	73k	73k	Description of a table from Wiki
Totto [31]	121k	7.7k	7.7k	Description of a table from Wiki
CommonGen [22]	67k	4k	1.5k	A sentence containing all required concepts

[2]We run the two experiments on single NVIDIA Tesla A100 with maximum number of tokens per batch set to 4000 without gradient accumulation

[3]https://inklab.usc.edu/CommonGen/leaderboard.html

[4]https://github.com/google-research-datasets/ToTTo

Table 8: Hyperparameters brought by contrastive learning at training and inference stage. 'scratch' means the transformer model without pre-training. 'α' is the balance factor between likelihood and sequence similarity and m means the maximum number of contrastive samples during training.

Datasets	model	Inference		Training	
		α	beam size	m	beam size
Small-scale model					
WMT'16 Ro-En	scratch	0.5	12	16	12
WMT'16 Ro-En	T5	0.5	12	16	12
IWSLT'14 De-En	scratch	0.5	12	32	24
XSum [26]	T5	0.5	8	16	12
Multi-News [8]	T5	0.5	8	16	12
WikiBio [20]	T5	0.3	8	16	12
Base-scale model					
WMT'14 En-De	scratch	0.2	8	16	12
Java [25]	CodeT5	0.2	8	16	12
Python [25]	CodeT5	0.2	8	16	12
Totto [31]	T5	0.3	8	16	12
CommonGen [22]	T5	0.2	4	16	12
Large-scale model					
XSum [26]	Pegasus	0.3	12	16	12
Multi-News [8]	Pegasus	0.5	4	16	12

C.1 Hyperparameters Brought by Contrastive Learning

Compared with MLE based models, there are three additional hyperparameters introduced by CoNT. The first one is the maximum number of contrastive samples m for an input sequence, the second one is margin strength γ (defined in Section 3.2) and the third one is the balance factor α. γ is set to 0.01 for all datasets. We tune α on the validation set from $[0.2, 0.3, 0.5, 0.7]$. We set the m to 32 on IWSLT'14 and set m to 16 on other benchmarks considering of efficiency. Actually, increasing the number of contrastive samples will continuously boost the performance. On all benchmarks, the beam size for diverse beam search [46] used in training is $0.75 * m$ and the number of groups of diverse beam search is the same with beam size. Details can be found in Table 8. Before adding contrastive learning, we pretrain the generation model with only negative log likelihood loss until the validation loss no longer decreases. All experiments are done on 4 NVIDIA Tesla A100 GPUs.

C.2 Machine Translation

For WMT'16 Ro-En, We use the Adafactor optimizer following [34] to finetune transformer-small and T5-small with learning rate = 1×10^{-3}. The pre-trained checkpoints of T5 are provided by transformers[5]. We limit the maximum input/output length to 128. Validation step is performed every 2000 training steps. We train our model for 2 epochs and get the best model at step 8000. It takes about 2 hours on 4 NVIDIA Tesla A100 GPUs with a batch size of 32. The dimension of hidden state of small-scale model is 512. So the dimension of representations from the encoder and decoder $\mathbf{z}$ is as the same. At inference stage we set the length penalty to 1.0.

Previous work mainly report their results on IWSLT'14 De-En and WMT'14 En-De based on fairseq [29] library[6]. We also implement CoNT on the two datasets with fairseq. For IWSLT'14 De-En, we use the small setup of the Transformer model. The model has 6 layers where model dimension for each layer is 512 and feed-forward dimension is set to 1024. The batch size is up to

[5] https://github.com/huggingface/transformers

[6] https://github.com/facebookresearch/fairseq

4000 tokens and we update our model every 4 backwards. On WMT'14 En-De, we use the base setup of the Transformer model where the dimension of feed-forward layer is set to 2048. The embedding for decoder input and output is shared. The batch size is also set to 4000 tokens but we update every 20 backwards to simulate large batch size which is very crucial to WMT'14 benchmark. In addition, we find that open the dropout module in decoder during inference will help the performance for CoNT on WMT'14 En-De.

We train our model for 20 epochs on IWSLT'14 De-En and 10 epochs on WMT'14 En-De. We use 4 GPUs for the model training. The average running time for IWSLT is about 8 hours and for WMT is around 32 hours. We use FP16 to accelerate our training. Other settings are the same with the default settings recommended by the instruction of fairseq official site to re-produce the neural machine translation results [7]. For both IWSLT'14 De-En and WMT'14 En-De, we use Adam optimizer with learning rate 5×10^{-4} with the inverse sqrt learning rate scheduler to optimize the models.

C.3 Summarization

We use the Adafactor optimizer with learning rate = 1×10^{-3} to finetune T5-small model and Pegasus-large [52] model. For XSum [26], we limit the input length to 512 and output length to 64. The input length is extended to 1024 and the output length is extended to 300 for multi-document summarization benchmark multi-news. The length penalty for XSum is set to 0.8 while for multi-news, which has longer target sequence, the length penalty is set to 2.0 We use a batch size of 32 for small-scale model and 4 for large-scale model. We train our model until the validation loss do not decrease. The total training hours using 4 GPUs is about 6 hours for small model and 12 hours for large model.

C.4 Code Comment Generation

We use the state-of-the-art code comment generation model CodeT5 [48] as our base model. We download their pre-trained checkpoint from transformers. we truncate the input length to 512 and output length to 64. The two benchmark is sensitive to batch size and learning rate. Therefore, we use a smaller learning rate 1×10^{-4} with Adafactor optimizer. The batch size is set to 8 and other settings are the same with the origin paper of CodeT5[8]. We train CoNT on the two benchmark for about 4 hours on 4 GPUs. The length penalty is set to 0.6 during decoding.

C.5 Data-to-text Generation

The input of data-to-text generation tasks is structured data (e.g., table, graph). To input the structured data into a sequence-to-sequence model, we should first the linear the input. We linear the input for WikiBio following Liu et al. [23][9]. For totto, we use preprocess the dataset following the instruction of the official site[10]. We use the T5-small model as base model for WikiBio and T5-base for Totto. We limit the input length to 512 and output length to 128. The length penalty for WikiBio and Totto is set to 2.0. We train our model for about 24 hours on 4 GPUs with a batch size of 32.

C.6 Commonsense Generation

For commonsense generation task, we use the popular benchmark CommonGen Lin et al. [22]. The input of CommonGen is a set of concepts and the output is a fluency sentence mentioning all concepts in the source side. We concatenate these concepts with ',' as separator. We use the base setup of the T5 model for CommonGen [22]. The maximum input length for source and target is limited to 64. Other settings are the same with the settings of Totto. We train our model for 1 epoch upon the checkpoint pre-trained with negative log likelihood loss. Since the scale of the dataset is samll, it only takes about 0.5 hours training to convergence on 4 GPUs.

[7] https://github.com/facebookresearch/fairseq/blob/main/examples/translation/README.md

[8] https://github.com/salesforce/CodeT5

[9] https://github.com/tyliupku/wiki2bio

[10] https://github.com/google-research/language/tree/master/language/totto

D Case Study

We show some randomly selected examples from IWSLT'14 Germanish-to-English translation task and XSum which aims to summarize a news article in table 9, 10, 11.

Table 9: Generation results of IWSLT'14 Germanish-to-English translation task (base model: Transformer-small).

Germanish:	dann kann ich das ganze übertragen.
Ground Truth:	and then i can transfer the whole thing.
CoNT:	and then i can translate the whole thing.
MLE:	then i can transmit this whole thing.
Germanish:	sie machen sich immer sorgen, dass sie regalfläche verlieren.
Ground Truth:	they‘re always worried they’re going to lose shelf space.
CoNT:	they‘re always worried that they’re losing the shelf.
MLE:	they always worry that they lose real galleries.
Germanish:	gewissermassen überflügelt uns die technik also.
Ground Truth:	so in a sense, it’s getting ahead of us.
CoNT:	so, in a sense, technology overwhelms us.
MLE:	so, to some extent, technology overrivers us.
Germanish:	erzählen sie mir über die "warum" phase–was bringt sie uns?
Ground Truth:	tell me about the "why" phase – what does that do for us?
CoNT:	tell me about the "why" phase – what does it bring us?
MLE:	tell me about why – what does it bring us?
Germanish:	und wir können auf diese sehr einfache weise navigieren.
Ground Truth:	and we can just navigate in this very simple way.
CoNT:	and we can navigate in this very simple way.
MLE:	and we can navigate this very simple way.
Germanish:	und wenn ich jemandem sage, gib mir mal salz oder pfeffer, dann wird er beim rechten erstmal überlegen, wo ist was drin.
Ground Truth:	and if i say to someone, pass me the salt or the pepper, they’d have to first think about what’s in what with the right one.
CoNT:	and when i say to someone, i’ll give me salt or pepper, they’ll put it on the right for the first time, where is it in.
MLE:	and when i tell someone, give me salt or pepper, they’ll be thinking about where is there?
Germanish:	und problem , das ist nicht nur ein technisches problem, es kann auch ein gesellschaftliches problem sein, es kann auch einfach ein zugangsproblem sein , was wie dinge vereinfachen, also eine beliebige problemstellung , eine frage aufzuwerfen, und wie kann man das anders oder wie kann man das besser machen.
Ground Truth:	and a problem, so not only a technical problem, it can also be a social problem, it can also just be an access problem that simplifying things, so any way of looking at a problem, of posing a question, asking how you could do something differently or better.
CoNT:	and the problem is, it‘s not just a technical problem, it can also be a social problem, it can be a problem as well as a problem of accessing how things simplify, which is an arbitrary problem of asking a question, and how do you do it differently, or how do you do it better.
MLE:	and the problem is, it‘s not just a technical problem, it can also be a social problem, it can be a problem as well as a problem of accessing how things simplify, which is an arbitrary problem of asking a question, and how do you do it differently, or better.

Table 10: Generation results of XSum (base model: T5-small).

Document:	The 23-year-old from Guernsey appointed Veronelli in December 2013, but he is no longer able to commit to spending up to 40 weeks a year on the road. Veronelli, 36, moved from Florida back to Buenos Aires earlier this year to be with his young family. Watson, the world number 55, won her second WTA tour title at the Hobart International in January. Veronelli, a former world number 150, had notable success in guiding Watson back inside the world's top 50 for a time, after she had slipped down the rankings following a bout of glandular fever in 2013.
Ground Truth:	British number two Heather Watson has parted company with her Argentine coach Diego Veronelli.
CoNT:	British number two Tom Watson has withdrawn from the WTA Tour due to illness.
MLE:	World number one Laura Watson has been reunited with his wife, Veronelli, after a long illness.
Document:	The 26-year-old was released by York City after failing to score in 14 appearances last season. However, he netted 26 times in 69 appearances in a two-season spell at Barnet between 2012 and 2014. Hyde is the Boro's fourth signing of the summer, following left-back Andrew Fox and forwards Matt Godden and Rowan Liburd. "I know this league inside out now and any team can go on a run, but it's who does it for the longest period that counts. "It's about winning games and fingers crossed I can help Stevenage do that this season," he told the club website. Details of his contract with Stevenage have not been disclosed. Find all the latest football transfers on our dedicated page.
Ground Truth:	League Two side Stevenage have signed their third striker of the summer by bringing in free agent Jake Hyde.
CoNT:	League Two side Stevenage have signed striker Jordan Hyde on a two-year contract.
MLE:	Stevenage have signed forward Ryan Hyde on a two-year deal. Accrington Stanley loan deal.
Document:	Both sides had chances before the Pars' Ryan Wallace drilled a low shot into the bottom corner with 15 minutes left. The lead last just two minutes as Ross Davidson got the last touch on a free-kick into the area. The home side dominated the closing stages but could could not deny Rovers, who remain in fifth place. Rovers remain level with Airdrieonians, who drew at home with Forfar Athletic.
Ground Truth:	Scottish League One leaders Dunfermline Athletic were held at home by Albion Rovers but still moved 11 points clear at the top of the table.
CoNT:	Tranmere Rovers slipped to the bottom of Scottish League One as they were held to a draw by play-off chasing Pars.
MLE:	Dundee Rovers dominated the Scottish Championships with a 2-0 win over Forfar Athletic.
Document:	Reports in France suggest Toulon coach Diego Dominguez's job is under threat, and Lancaster, 46, is viewed as a potential successor. He left his role with England after their exit from the 2015 World Cup. Since his departure, Lancaster has been an adviser to the Football Association, the NFL and British Cycling. He is interested in the Toulon job, but is understood to still be keen on a role in the southern hemisphere, with posts in Australia at the Queensland Reds and the Western Force both available.
Ground Truth:	Former England boss Stuart Lancaster is meeting Toulon president Mourad Boudjellal this week as he seeks a return to full-time coaching.
CoNT:	Former England coach Stuart Lancaster is interested in a role in the southern hemisphere, according to the Football Association.
MLE:	England's Chris Lancaster has been appointed as the new head coach of the Toulon Football Association.

Table 11: Generation results of XSum (base model: T5-small).

Document:	The men were believed to have been working on the northbound carriageway of the A361 at Gornhay Cross, Tiverton, when they were hit. One was flown to Plymouth's Derriford Hospital, while a second Devon Air Ambulance flew the other man to the Royal Devon and Exeter Hospital. The van driver has been arrested on suspicion of dangerous driving. The condition of the two injured men, who are both in their 40s, is not known. Devon and Cornwall Police said a third man was hurt in the incident, but he is described as the "walking wounded". The Barnstaple-bound carriageway is expected to remain closed for several hours and diversions have been put in place.
Ground Truth:	ref: Two road workers have been seriously hurt in an accident involving a van in mid Devon.
CoNT:	Two men have been seriously injured in a collision involving a van in Devon.
MLE:	A man has been arrested on suspicion of dangerous driving after two men were hit by a van on the M4 in Devon.
Document:	Officers were called at about 02:30 BST following reports that a man in his 30s had been attacked in Greenbrow Road, Wythenshawe. He was taken to hospital but he died from his injuries. Supt Steve Howard, from Greater Manchester Police, said: "We are working hard to piece together what happened to the man."
Ground Truth:	A murder investigation has been launched after a man was found stabbed in Manchester, police have said.
CoNT:	A murder inquiry has been launched after a man was stabbed to death in Manchester.
MLE:	An 81-year-old man has died after being attacked in a street in Greater Manchester.
Document:	Damage to to overhead wires meant the line is blocked north of Morpeth. Virgin East Coast, Northern Rail, and Cross Country services were affected, with reports of large queues at Newcastle Central Station. Buses were organised to take passengers between Newcastle and Edinburgh, with people advised to avoid travelling if possible. Services resumed late on Friday.
Ground Truth:	Rail passengers travelling between Newcastle and Scotland faced severe disruption on Friday.
CoNT:	Rail services between Newcastle and Edinburgh have been disrupted after a power cut led to delays.
MLE:	Trains across the UK have been cancelled due to a disruption to the main line in the Highlands.
Document:	Rashan Charles, 20, was wrestled to the ground in Dalston, east London, on 22 July, and died about an hour later. On Friday, clashes broke out in Hackney as protesters blocked part of Kingsland Road and set mattresses alight. A spokesman for Mr Charles's family said they understood the anger but called for "dignified" protest. "Burning down homes will not give justice," he said. Mr Charles was pursued by officers and became ill after trying to swallow an object, the Met has said. He died soon after in hospital. The Independent Police Complaints Commission is investigating. Police warned that anyone using Mr Charles's death "as an excuse to commit crime" would be "dealt with robustly". Appealing for calm, family spokesman Stafford Scott said: "We understand your frustration, we understand your anger - don't feel that the family doesn't feel the anger and the frustration too. "But what the family knows is taking it to the streets doesn't give you justice. "Burning down your own homes, burning down your neighbourhood is not going to give you justice." Mr Scott, who runs race advocacy group Tottenham Rights...
Ground Truth:	The family of a black man who died after being apprehended by police has appealed for peace after violent protests in the wake of his death.
CoNT:	The family of a man who died after he was attacked by anti-racism protesters have appealed for calm.
MLE:	nll: Thousands of black people have protested against the death of a man who was killed in a street attack.

E Societal Impact

Our proposed model can greatly improve conventional auto-regressive generation models allowing a lightweight model to achieve on-par performance with large model. We believe this is generally a benefit of generation models in real-world applications while it could also have negative societal implications. When these technologies are successfully deployed on a large scale, it may lead to job losses. Generative technologies may also lead to cheaper production of fake news.