COBART: Controlled, Optimized, Bidirectional and Auto-Regressive Transformer for Ad Headline Generation

Yashal Shakti Kanungo*
Amazon
United States of America

Gyanendra Das*†
Amazon
India

Pooja A
Amazon
India

Sumit Negi
Amazon
India

ABSTRACT

Online ads are essential to all businesses and ad headlines are one of their core creative component. Existing methods can generate headlines automatically and also optimize their click-through-rate (CTR) and quality. However, evolving ad formats and changing creative requirements make it difficult to generate optimized & customized headlines. We propose a novel method that uses prefix control tokens along with BART [16] fine-tuning. It yields the highest CTR and also allows users to control the length of generated headlines for use across different ad formats. The method is also flexible and can easily be adapted to other architectures, creative requirements and optimization criteria. Our experiments demonstrate a 25.82% increment in Rouge-L and a 5.82% increment in estimated CTR over previously published strong ad headline generation baseline.

CCS CONCEPTS

• **Computing methodologies** → **Natural language generation**;
• **Information systems** → **Online advertising**.

KEYWORDS

ad generation; controlled generation; sponsored advertising; ad optimization; headline generation; E-commerce

ACM Reference Format:
Yashal Shakti Kanungo, Gyanendra Das, Pooja A, and Sumit Negi. 2022. COBART: Controlled, Optimized, Bidirectional and Auto-Regressive Transformer for Ad Headline Generation. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '22), August 14–18, 2022, Washington, DC, USA.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3534678.3539069

Input (Single ad with multiple products; two shown)			
Product 1	Boyd's Cosmetics NYC- No Lines Temporary Wrinkle Remover with our Brush It Away Medium ...		
Product 2	No Lines Temporary Wrinkle Remover ...		
Generated Headlines			
UniLM	*Boyd's New York City Cosmetics*		
BART	*New & improved anti-aging with our brush*		
COBART	**Length →**		
CTR ↓	Short	Medium	Long
Low (Bucket 1/15)	*Boyd's Cosmetics NYC- No Lines*	*Save Now on New York City Collections*	*No Lines Temporary Wrinkle Remover with Brush It Away*
Medium (Bucket 7/15)	*Shop Boyd's NYC, No Lines Remover*	*Tackle Fine Lines & Wrinkles with Boyd's NYC*	*Boyd's Cosmetics NYC - No Lines Temporary Wrinkle Remover*
High (Bucket 15/15)	*Shop New York City's #1 Makeup Brand*	*New & improved anti-aging with our brush - Boyd's NYC*	*A New Generation of Makeup That Targets Fine Lines with a New Look*

Table 1: Ad headline generation for a multi-product ad. Our method allows to *control* and *optimize* the length and click-through-rate (CTR) of headlines during inference.

1 INTRODUCTION

Online advertising is an evolving domain that is now integral to all businesses. Organizations across the world use ads to improve discovery and reach of their businesses. Sellers on E-commerce platforms use online ads to build brand awareness and connect shoppers to their products.

One of the key component of ads is their 'headline'. Headlines are a part of all ad formats including text, image, and video ads as illustrated in Figure 1. They summarize the key properties of the underlying products and promote the customers to engage. They may also be the first preview of the product/brand for shoppers. Thus having the right style and appeal is important to attract shopper attention. Style and content are not the only factors for generating headlines. Length is equally important for ads across different formats and target device screen sizes. Long headlines are used in text ads, medium length headlines are shown alongside images and shorter headlines are typically overlaid on images and

*Equal contribution. Correspondance to yashalk@amazon.com.
†Work done during internship at Amazon

KDD '22, August 14–18, 2022, Washington, DC, USA

ACM ISBN 978-1-4503-9385-0/22/08.
https://doi.org/10.1145/3534678.3539069

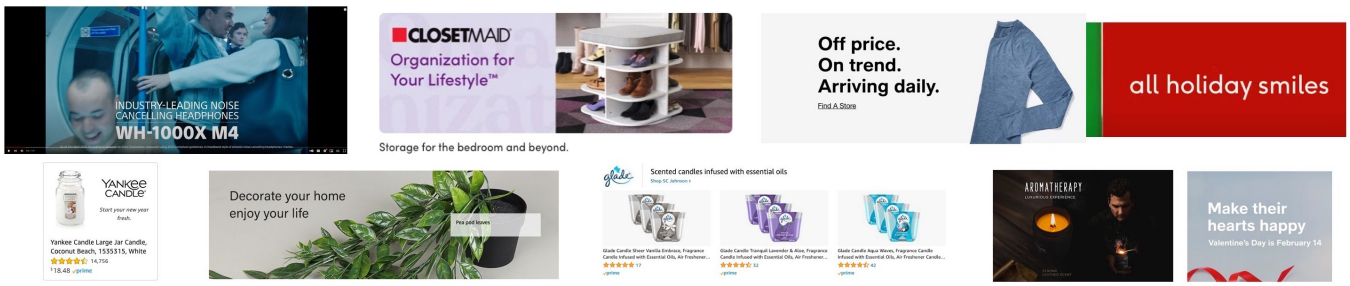

Figure 1: Ads from across the internet with different headlines. These include text-only ads, image+text ads, or screengrabs from video ads with headlines included within the video. The headlines have varying lengths, messaging, and advertising intent. Some headlines are informative with no adjectives, some are short, and some others tend to be much longer.

videos. Mobile-only ad formats also encourage text ads with shorter headlines.

Language barriers, varying shopper interests, market trends, seasonal trends, varying inventory of products, changing style preferences etc. add to the complexity and scale of the problem. With rapidly evolving product catalogs with sometimes billions of products, manually writing such headlines is infeasible and will require extensive human effort. The content present within the product catalog may also have a high variance in quality. This makes it difficult to use it as reference, leading to keyword stuffing and poor headlines. It is thus difficult to write these headlines for different products, ad formats and requirements at scale.

It is not sufficient to just automate the headline generation process. One of the key metrics that is used to measure headline attractiveness and engagement is the click-through-rate (CTR). It is the ratio of number of clicks a headline receives to the number of times the headline was shown. Thus, ideal headlines need to be automatically written, meet platform and policy requirements, meet creative requirements (length, style etc.) and yield high CTR. Moreover, any automatic solution that solves this problem practically needs to be highly scalable and extendable to meet a wide range of requirements.

We propose a novel method to fine-tune a pre-trained Bidirectional and Auto-Regressive Transformer (BART) [16] model to generate ad headlines for multiple input products. It jointly learns to control and optimize any desired characteristic of the headline such as CTR and length using special control tokens. These control tokens are provided as input to the bidirectional encoder and allow the encoder and decoder's Attention to be conditioned to control CTR and length. We compare our proposed method to other state-of-the-art controlled generation methods [9], other ad headline generation method [12], and multiple other baselines and ablations. An example ad with headlines generated by baselines and our method is illustrated in the Table 1.

Our key contributions are:

- Our method allows users to control and optimize desired characteristics of generation at inference time.
- Our method allows to mix-and-match multiple characteristics. We also show that it can be combined with other optimization techniques to further improve the performance.
- To validate our proposal, we build extensive baselines, try multiple variations & ablations, and demonstrate significant improvement in Rouge-L and CTR. The results show that it is possible to effectively use control tokens to fine-tune language models that were pre-trained without them.
- Thus, our proposal solves a large-scale real-world problem using a novel method with multiple practical applications. It also does not have any negative impact on model inference latency and can replace existing methods with minimal changes.

2 RELATED WORK

Natural Language Generation. Transformer[26] based methods for Natural Language Generation (NLG) have shown great potential in the recent years. They span all possible combinations of the Transformer encoder and decoder, including encoder-only NLG [5], decoder-only NLG [3], encoder-decoder NLG [16, 19, 20, 24], combining differently pre-trained models [22], and joint pre-training [24]. Methods [1, 20] have shown that it is possible to use task specific prefixes to train a single model for multiple tasks. All of these methods outperform earlier LSTM + Attention [23] based methods by implicitly generating better headlines. These methods can be used for headline generation but they do not perform any additional optimization to improve CTR or control the generation.

Controlled NLG. In [9], authors propose to use LSTM based VAEs in a generator-discriminator setting to condition the generator output on the VAE and an additional semantic style code to control the generation. However, they target each feature to be controlled independently. CTRL [13] proposes to condition the output of a decoder Language Model (LM) on specific control codes in order to augment the generation. Decoder only models offer lesser flexibility for conditional text generation compared to encoder-decoder models such as [16]. PPLM [4] combines Transformer based LM with bag-of-words from different topics of choice to change style. FUDGE [31] models the style conditional LMs using Bayesian factorization instead of modeling the conditional generation directly. MuCoCo [15] combines pretrained LMs with differentiable constraints to frame the generation as a constrained optimization problem and control the generation. However, it is relatively slow for practical deployment.

Headline Generation. Headline Generation is akin to other text-to-text generation tasks such as summarization. The discussed NLG methods can be applied for headline generation, and various methods have proposed the use of Universal Transformers [6], encoding structure of the text for better headlines [32], or using Reinforcement Learning (RL) based techniques along with Transformer based architecture [12] for improving the quality of generation. As discussed earlier, the attractiveness of headlines may be judged by the CTR they yield. Optimized headlines yield higher CTR. Reinforcement Learning (RL) based techniques such as Actor-Critic, Self-critical Sequence Training (SCST) [21] etc. have been shown to improve ad headline generation CTR when used with LSTMs [10, 25, 30]. [11] proposes to use style dependent Layer Norm and Transformer-Query transformation to transform headlines. However, these methods either implicitly improve generation quality & do not allow for explicit control during inference or do not take advantage of in-domain datapoints and observed CTR.

In our method, we extend the BART encoder-decoder architecture [16] with characteristic control tokens (COBART: Section 3.1). To evaluate our method, we also create two more alternative extensions to BART that optimize CTR. The first (SCBART: Section 3.2) uses Self-critical training [21]. The second extension (VBART: Section 3.3) uses variational Transformer encoder with a generator and a discriminator [9] .

3 METHODS

Every ad consists of a set of one or more products P such that the i^{th} product is represented by the tokens in its title $x^i = (x^i_1, x^i_2 ... x^i_{|x^i|})$.

During training, we have access to the original ad headline tokens $H = (h_1, h_2 ... h_{|h|})$. We also have a set of observed or computed characteristics of each headline $\Phi = \{\phi_1, \phi_2 ... \phi_{|\Phi|}\}$. Examples of these characteristics include the observed continuous CTR value, the discreet number of words present in the headline, categorical season in which the headline was advertised, binary variables such as presence of brand names etc.

We primarily experiment with controlling and optimizing the observed CTR of the headlines (ϕ_{ctr}) and also include results of using the number of words present in the headline (ϕ_{Length}) without any loss of generality. These two aspects define the control over attractiveness and semantics (ϕ_{ctr}) and the structure of the headlines (ϕ_{Length}).

During inference, we are required to generate the headline $\hat{H} = (\hat{h}_1, \hat{h}_2 ... \hat{h}_{|\hat{h}|})$ given the input products P and the desired level of attractiveness and length.

3.1 Controlled and Optimized BART

The Bidirectional and Auto-Regressive Transformer (BART) model uses the Transformer [26] architecture as a denoising autoencoder for pretraining. It is pretrained using a novel text infilling scheme combined with sentence permutation. For an original sentence $(x_1 x_2 x_3 . x_4 x_5 .)$ BART encodes $(x_4 x_5 . x_1 \_ .)$ using a bi-directional Transformer encoder. It then passes the encoder's output to an auto-regressive left-to-right Transformer decoder that regenerates the original sentence $(x_1 x_2 x_3 . x_4 x_5 .)$ by infilling (x_2, x_3) and permuting the sentences to the correct order.

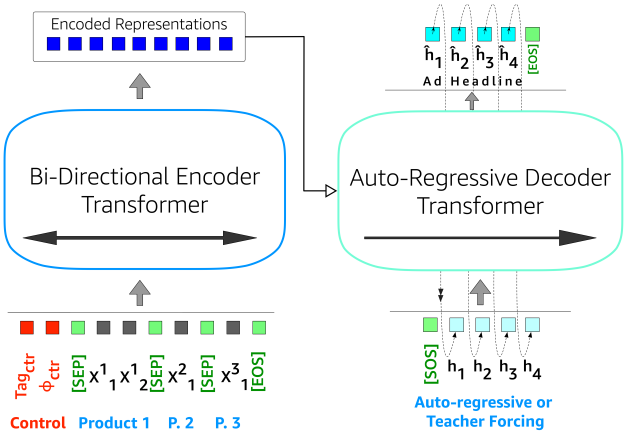

Figure 2: The COBART model takes computed characteristic control tokens (Φ) based prefixes as input to the bi-directional encoder during training. During inference, these are replaced by user desired characteristics and passed on to the decoder for controlled and optimized generation.

Headline generation without any controllable characteristics. We fine-tune a pretrained BART model [16] following the input strategy used in [12]. We input the product title tokens $(x^1_1, x^1_2 ... x^1_{|x^1|})$, with titles from different products concatenated using an otherwise unused separator token. An example input would be:

$$X = (x^1_1, x^1_2, [SEP], x^2_1, [SEP], x^3_1) \tag{1}$$

We thus minimize the loss:

$$\mathcal{L}_{BART} = -\log \prod_{t=1}^{|H|} p(h_t | h_{1:t-1}, X) \tag{2}$$

Headline generation with controllable characteristics. We propose a simple yet highly effective addition to jointly learn headline generation along with optimization of characteristics. We propose to add a control token as prefix to the input for each characteristic $\Phi = \{\phi_1, \phi_2 ... \phi_{|\Phi|}\}$ that we want to optimize and control. We compute the characteristics Φ for each headline in the training set and add them as multiple prefixes along with corresponding tags. Figure 2 illustrates the model. For instance, to optimize headline CTR, we bucketize the observed CTR ϕ_{ctr} for each headline into multiple categorical buckets and then use it as an additional control token in the input. We perform all our experiments with ϕ_{ctr} and also include results for experiments to control the length of the generated headline using the control token ϕ_{Length}.

An example input for COBART would be:

$$X_{CO} = (Tag_{ctr}, \phi_{ctr}, [SEP], x^1_1, x^1_2, [SEP], x^2_1, [SEP], x^3_1) \tag{3}$$

This can be extended and combined with control tokens for any desired characteristic present within the training set. The Attention formulation automatically allows the encoded representation of the input to be updated based on the control tokens without any modifications to the Transformer architecture.

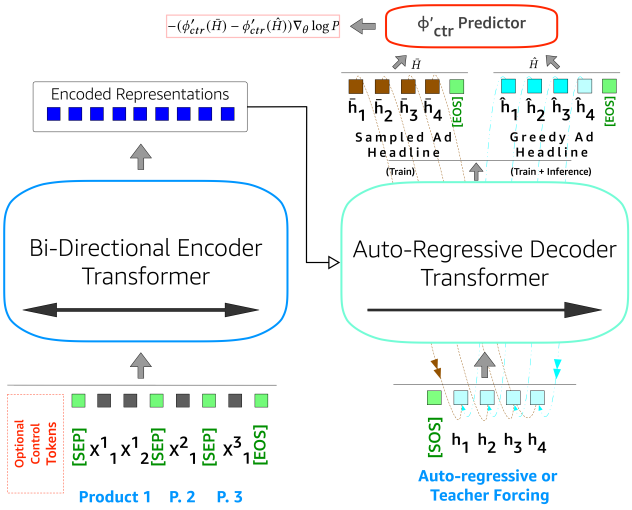

Figure 3: The Self-critical BART uses RL based reward difference between sampled and greedy headlines generated from the decoder to update the model. We also optionally use control tokens to obtain Self-critical COBART.

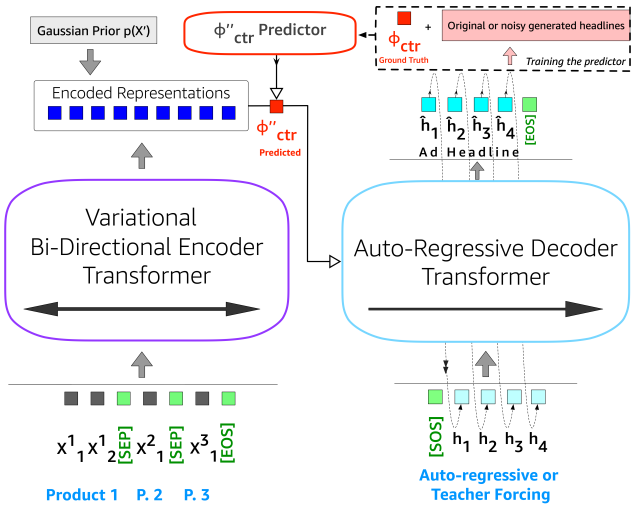

Figure 4: The VBART model conditions the generation on predicted CTR along with the encoded representation rather than using control tokens. It adds additional discriminator loss using wake-sleep procedure to train the predictor.

3.2 Self-critical BART

We extend previous Self-critical Sequence Training (SCST) [21] ad headline generation methods [10, 12] by experimenting the usage of SCST for improving CTR of headlines generated by a pretrained and fine-tuned BART model. In this method illustrated in Figure 3, instead of using the original observed CTR ϕ_{ctr} of train headlines, we use an estimate ϕ'_{ctr} of each generated headline as a reward that drives the optimization criterion. The estimate ϕ'_{ctr} is predicted by an oracle model based on the DeBERTa [7] architecture.

For parameters θ of the BART model, we follow the policy π_θ and generate the headline $\bar{H}$ by sampling. We then aim to minimize the negative expected reward $\phi'_{ctr}(\bar{H})$.

$$\mathcal{L}_{RL} = -\mathbb{E}_{\bar{H}\sim\pi_\theta}[\phi'_{ctr}(\bar{H})] \tag{4}$$

Using the REINFORCE trick [27] and SCST, we can estimate the gradient to optimize ϕ'_{ctr} as:

$$\nabla_\theta \mathcal{L}_{SC\_BART} \approx -(\phi'_{ctr}(\bar{H}) - \phi'_{ctr}(\hat{H}))\nabla_\theta \log P_{sc} \tag{5}$$

where $\hat{H}$ is the headline generated by BART using greedy strategy, X is the input from Eq 1 and

$$P_{sc} = \prod_{t=1}^{|\bar{H}|} p(h_t|h_{1:t-1}, X) \tag{6}$$

We train the model using the following combined loss with hyperparameter λ:

$$\mathcal{L}_{SC\_Total} = \lambda * \mathcal{L}_{BART} + (1-\lambda) * \mathcal{L}_{SC\_BART} \tag{7}$$

The proposed usage of characteristics control tokens is directly compatible with the Self-critical framework. Thus, we also experiment with Self-critical COBART which combines SCST with COBART.

3.3 Variational BART

We also extend previous style-transfer and controllable generation method [9] that uses variational LSTM auto-encoders. It allows conditioning of the decoder on a set of structured variables in addition to the latent output of the encoder as shown in Figure 4.

We extend the architecture to work with Transformer encoder-decoder. The pre-trained Transformer encoder probabilistically encodes the input X from eq. 1 with standard Gaussian prior $p(X')$ to X' such that;

$$X' \sim q(X'|X) \tag{8}$$

Additionally, a Transformer discriminator model is fine-tuned further during training to predict a CTR estimate ϕ''_{ctr}. It is trained by pairing the original CTR ϕ_{ctr} with the original human-written headlines and a small set of noisy generated headlines.

We pose the generation to be conditioned on the estimate ϕ''_{ctr} and X' with the distribution:

$$P_{vt} = \prod_{t=1}^{|\bar{H}|} p(h_t|h_{1:t-1}, X', \phi''_{ctr}) \tag{9}$$

Following [9], we update the encoder and the discriminator using a wake and sleep procedure. The model is updated using the loss:

$$\begin{aligned} \mathcal{L}_{vt} &= \mathbf{KL}(q(X'|X)||p(X')) \\ &-\lambda_1 \mathbb{E}_{q(X'|X)\phi''_{ctr}}[\log P_{vt}] \\ &-\lambda_2 \mathbb{E}_{p(X')\phi_{ctr}}[\log \phi''_{ctr}] \end{aligned} \tag{10}$$

4 EXPERIMENTS

We primarily experiment with all the discussed methods and baselines to generate headlines that optimize CTR. We then also experiment to study the effectiveness of our method to control the length of generation. We finally train a model that jointly allows to control and optimize both the length and the CTR.

4.1 Training

Data. We conducted our experiments on products and human-written headlines written in English. We used over 500,000 ad campaigns that were created on Amazon by sellers who have signed-up for advertising. The problem requires 10x compression of the input and over 50% abstraction.

We deduplicated all the ads with the same input-output pairs and made sure there was no overlap in the train, val and test splits. We also made sure that none of the ads from the val and test sets were used for training of any supporting models (oracle models, ϕ_{ctr} predictors or calculation etc.). We only selected the ads that comply with ad policies as verified by internal experts.

Framework. We use the HuggingFace [28] pre-trained versions of the Transformer models. We fine-tune the 'large' variants of the models using Adam [14] optimizer with early-stopping for upto 15 epochs. For early stopping, we track Rouge-L [17] on the validation set with patience of 3 epochs and minimum increment of 0.1. For our VBART implementation, we refer to the Texar [8] library. We use 'fp16' for faster training and inference latency optimization.

COBART. For training COBART that allows for controlling and optimizing CTR, we obtain the observed CTR for all the headlines in the training data and bucketize it into 15 equal sized buckets based on CTR percentile with the prefix tag 'engaging:'. For controlling the length of the headlines, we add a control tag with three possible values: 'short', 'medium', and 'long' based on whether the actual headline has ≤ 5, 6 to 8 or more than 8 words respectively. We use the prefix 'length' for these length control tokens.

We train 3 variants, first only on ϕ_{ctr} to compare Rouge-L and CTR against other methods without effects of user specified length. We then train COBART only on ϕ_{Length} to compare against the baseline length control method without effects of CTR and finally on both ϕ_{ctr} & ϕ_{Length} to study if multiple criteria can be jointly controlled and optimized.

SCBART. For the training of Self-critical variants, ϕ'_{ctr} is estimated using an oracle model that is treated as a black-box and is not updated during the training of Language Models. We follow the procedure discussed in Section 3.2 and experiment with different values of λ.

VBART. We follow the training procedure indicated in [9] using Transformer based models. The discriminator is also updated following a sleep-wake-procedure. Apart from the difference in the backbone LM architecture, another difference is that we omit the auto-encoder based loss as our model is not an auto-encoder and thus the generator output cannot be fed back into the encoder.

4.2 Inference and Deployment

During inference, the control tokens from training are replaced with the desired attribute. For instance, ϕ_{Length} is replaced with the 'short' token if short headlines are desired, or the token corresponding to the highest CTR bucket if a completely engaging headline is required. For other models which use continuous CTR, we use the highest known CTR as the input for all the test headlines.

We use Beam Search with a beam size of 5, Length Penalty of 1.5 (For ϕ_{ctr} experiments), and Repetition Penalty of 2.0 as tuned on the validation set.

All the results are reported for the setting in which the CTR input provided during inference corresponds to the highest CTR. For instance, if bucket 15 had the highest CTR during training, then during inference, all headlines are generated with bucket 15 as the desired CTR.

4.3 Additional baselines and ablations

We use the best performing method in [12] as our first baseline. It has already been shown to outperform LSTM and RNN based methods and we thus exclude them.

For a more thorough comparison, we also fine-tuned the large variants of T5 [20] and ProphetNet [19] to generate the headlines.

As another baseline, we fine-tune the BART model using historical ads that have observed CTR ϕ_{ctr} higher than the median CTR of the entire training data. The idea being that the model should learn to generate headlines with only the style of high CTR headlines.

To incorporate continuous ϕ_{ctr}, we bucketize the feature into multiple percentile based buckets. We study the effect of using fine-grained buckets by using different number of buckets in the VBART and COBART models.

In the COBART model, we propose that adding the control tokens as an input to the encoder improves performance. We thus also study the effect of concatenating the ϕ_{ctr} value to the output of the encoder rather than as input.

We also experiment with different values of λ in the total Self-critical loss function shown in equation 7.

Finally, to study the impact of length control, we compare our COBART variant that is trained to control generation length against our BART model whose length is controlled using Length Penalty [18]. We try to vary the penalty value to obtain desired lengths.

5 RESULTS

We first evaluate if the use of different techniques with the same pre-trained model and the same training data results in different generated headlines. Figure 5 illustrates that all the three methods generate more than 90% new headlines compared to baseline and human-written reference. There is a very high overlap between COBART and SC-COBART.

5.1 Overlap with human-written headlines

We measure several overlap metrics of the generated model headlines with the reference human-written headlines that were approved by internal experts. We report the Rouge-L [17] F1 metric. All the other commonly used overlap metrics such as CIDEr, BLEU-4, METEOR etc. were in complete agreement with Rouge-L and we thus omit them.

Model	Inputs	Rouge-L	CTR v. Human	CTR v. Baseline
Transformer Variants and Baselines				
UniLM SCST with Rouge-L [5, 12]	Titles	-	7.84%	-
T5 [20]	Titles	-0.72%	-1.96%	-9.09%
ProphetNet [19]	Titles	1.57%	2.16%	-5.27%
BART [16]	Titles	10.75%	-3.73%	-10.73%
BART trained on filtered high CTR data	Titles, $\phi_{ctr}{}^*$	5.18%	4.90%	-2.73%
SC-BART and ablations				
SC-BART : λ = 0.0	Titles, $\phi'_{ctr}{}^*$	-23.33%	-10.20%	-16.73%
SC-BART : λ = 0.1	Titles, $\phi'_{ctr}{}^*$	-17.43%	-5.88%	-12.73%
SC-BART : λ = 0.5	Titles, $\phi'_{ctr}{}^*$	16.87%	7.25%	-0.55%
SC-BART : λ = 0.9	Titles , $\phi'_{ctr}{}^*$	18.94%	3.92%	-3.64%
Variational BART and ablations				
VBART : 2 ϕ''_{ctr} buckets	Titles, $\phi_{ctr}{}^*$, ϕ''_{ctr}	11.40%	3.92%	-3.64%
VBART : 15 ϕ''_{ctr} buckets	Titles, $\phi_{ctr}{}^*$, ϕ''_{ctr}	14.02%	6.27%	-1.45%
VBART : Continuous ϕ''_{ctr}	Titles, $\phi_{ctr}{}^*$, ϕ''_{ctr}	19.95%	10.39%	2.36%
COBART for ϕ_{ctr} and ablations				
Concatenating ϕ_{ctr} with encoder embeddings	Titles, ϕ_{ctr}	20.58%	11.37%	3.27%
COBART : 2 ϕ_{ctr} buckets	Titles, ϕ_{ctr}	6.91%	0.15%	-7.13%
COBART : 15 ϕ_{ctr} buckets	Titles, ϕ_{ctr}	**21.89%**	**11.76%**	**3.64%**
SC-COBART	Titles, ϕ_{ctr}, $\phi'_{ctr}{}^*$	**25.82%**	**14.12%**	**5.82%**

Table 2: The % improvement over UniLM baseline in overlap with ad policy compliant human-written headlines (Rouge-L) and the estimated CTR. We also report % improvement in the estimated CTR compared to human-written headlines. The COBART model that optimizes ϕ_{ctr} outperforms other techniques. It is further boosted when combined with SCST.
(*) indicates the feature is only used during training. Other features are replaced with the highest value during inference.

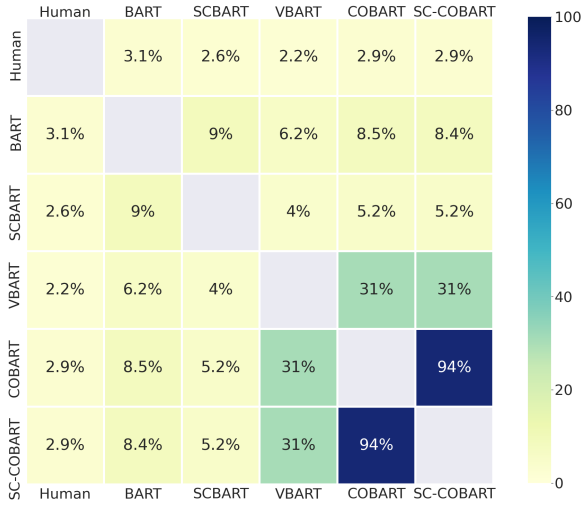

Figure 5: The % of generated headlines that are exactly the same across different sources.

Table 2 shows the Rouge-L metrics across all the experiments.

In agreement with previously published results which indicate that BART performs well on summarization related tasks [16], we observe that BART achieves better Rouge-L score than T5, ProphetNet, and UniLM that is fine-tuned with SCST for Rouge-L. It shows an improvement of 10.75% in the Rouge-L score over UniLM. In the case when the BART model is only fine-tuned on the top-50 percentile of data that has the highest CTR, we observe that the Rouge-L score drops significantly from an improvement of 10.75% to an improvement of 5.18%. This could be an indication that the model is not able to generalize to all kinds of products and brands present in the complete dataset and learn commonly associated product features.

On using the Self-critical framework to optimize CTR for the BART model, we see an improvement in Rouge-L for higher values of λ. On increasing the contribution of the CTR based loss (lower λ values), we see that the CTR improves at the expense of Rouge-L. Moreover, model does not learn when only the CTR based loss is used.

In the case of Variational BART generator, we see that the Rouge-L improves consistently across all the experiments. We see the highest improvement when the CTR is used directly as a continuous value which shows that dividing CTR into categorical buckets does lead to loss of information.

Similar to Variational BART, adding finer and more CTR buckets improves the performance of the COBART model and yields the highest Rouge-L score of any model. Increasing the number of buckets beyond 15 does not yield any benefit. On comparing the results to an ablation in which the CTR is concatenated with output Key and Value embeddings from the encoder, we see that the performance decreases compared to when it is given as input to the encoder. Combining the Self-critical framework and COBART model boosts and yields the overall highest Rouge-L score.

5.2 Estimated CTR

We use the oracle CTR model to estimate the CTR of the generated headlines. In Table 2, we also report the % improvement in the estimated CTR compared to the estimated CTR of the 1) Baseline Self-critical UniLM model 2) Human-written headlines.

Trade-off between Rouge-L and CTR. In some interesting scenarios the CTR improved but the Rouge-L score did not and vice-versa (UniLM, T5 and ProphetNet v. BART etc.). We observe that at times the models may generate more novel words (adjectives etc.) that improve CTR but those words do not align with the reference headline and specific features of the products and brands. There is thus a trade-off in such scenarios and in the ideal scenario, both the overlap scores and CTR should increase.

The best SC-BART variant outperforms BART trained without SCST in terms of CTR. It also nearly equals the CTR of the baseline SC-UniLM model at a significantly higher Rouge-L score. For VBART, using continuous ϕ''_{ctr} values as input to the generator gives the highest boost to the estimated test CTR. The model outperforms the baseline model by 2.36% and also outperforms the human-written headlines by 10.39%. Using bucketized CTR in this setting leads to worsened performance.

On using just 2 buckets for CTR with the COBART model, we see a big drop in the estimated test CTR compared to the model that uses 15 buckets. This shows that a certain level of hyper-parameter tuning is required for continuous control characteristics. On comparing the results to an ablation when the CTR is concatenated with encoder output, we see that the performance decreases compared to when it is given as input to the encoder. This highlights the advantage of conditioning the encoder output on the control tokens. COBART model outperforms baseline CTR by 3.54% and human-written headline CTR by 11.76%. When combined with SCST, we see a further boost in the CTR performance even when only 6% of all the test headlines change.

5.3 Controlling Length

Table 3 shows the median length of different types of headlines generated using the BART + Length Penalty baseline and the COBART variant trained to control length. We see that COBART is effectively able to control the length of the generated output whereas BART + Length Penalty hits an upper bound. On increasing the penalty beyond 2.0 (We tried upto 10) we see that the effective length actually decreases due to diminishing relative effects of the penalty.

81% of long and medium & 35% of short and medium headlines generated by the baseline BART were exactly the same. Whereas, only 4% of long and medium and fewer than 1% of short and medium headlines were same in COBART.

Thus, our results show that COBART is a much more reliable method to control length. It is able to generalize to all kinds of products without needing parallel training corpora with different headlines available for the same product.

Control Tag	Average num. of characters	Average num. of words
BART - Length Penalty: 0.25	31	5
BART - Length Penalty: 1	37	6
BART - Length Penalty: 1.5	40	6
BART - Length Penalty: 2.0	35	6
COBART - ϕ_{Length}: Short	30	5
COBART - ϕ_{Length}: Medium	42	7
COBART - ϕ_{Length}: Long	47	9

Table 3: Controlling headline length using COBART ϕ_{Length} and Length Penalty (BART) during inference

6 QUALITATIVE ANALYSIS AND DISCUSSION

6.1 Optimizing CTR and Rouge-L

Table 4 illustrates model generated headlines for a set of products from the test set. Across these and other samples in the test set, we notice that the models with higher CTR tend to generate headlines that are more descriptive and complete and have much rarer keyword stuffing issues. On comparing to the set of human-written headlines, we can see that the models are able to generate headlines with great quality. All the models are also able to effectively combine information from all the products in the ad. We also observe that even though the train and test splits do not have any product-title overlap, the models are able to learn phrases associated with certain brands and use them for new headlines generated for their newer products. This behavior is reflected in the increment in the Rouge-L score.

We also analyzed if there are any particular words that are used more often by one model over another. We thus compared the frequency of words across all headlines generated by different models. It is interesting to note that compared to BART, the SC-COBART model uses 'Adventure' 9x times, 'Brighten' 8x times, 'Always' 7x times, 'Everyday' 7x times and 'Powerful' 4x times. At the same time it uses 'Good' 7 times fewer, '24/7' 5.5 times fewer, 'favorite' 3 times fewer and 'Pain' 2 times fewer. The increased CTR and quality cannot be attributed to just the presence of these words but it does show a trend that the proposed model is able to use more attractive words that entice higher CTR.

6.2 Controlling Length

Table 5 illustrates headlines generated by the baseline approach and our proposed approach to control length. It is evident that COBART is able to generate headlines with different lengths much more consistently. The short headlines tend to describe the products in a few words, medium headlines add some additional words and the long headlines are consistently most descriptive.

On examining the model that controls both CTR and length, we see that the model is able to jointly control both the parameters effectively and yields results consistent with other experiments. For instance, on examining the subset of test ads with short human written headlines, we see that COBART is also able to generate

Product Title 1	Product Title 2	Human	BART	SCBART	VBART	COBART ϕ_{ctr}	SC-COBART ϕ_{ctr}
2 Pack 12 Colors Makeup Naked Eyeshadow ...	60 Colors Eyeshadow Palette, 4 in1 Color Board ...	*Vibrant Colors, Most Definitely Worth Eyeshadow*	*Best Eyeshadow Palette Set*	*BestLand Eyeshadow palette*	*Best Best Gift for Your Loved Ones*	*Shimmering Glitter Eyeshadow Palette*	*Shimmering Glitter Eyeshadow Palette Set*
[2020 Upgraded Version] ZeeHoo Wireless Car Charger,15W Qi ...	ZeeHoo Wireless Car Charger,10W Qi Fast Charging Auto-Clamping ...	*Fast Charging Auto Clamping Wireless Car Charger*	*ZeeHoo Wireless Car Charger Mount*	*ZeeHoo Wireless Car Charger*	*Auto Clamping Wireless Car Charger Mount*	*Auto Clamping Wireless Car Charger Mount by ZeeHoo*	*Auto Clamping Wireless Fast Charging Car Mount*
New! - Shush Biker - High Performance Hearing ...	NEW - Shush Acoustic - Universal-Fit ...	*New: High Fidelity Earplugs with Ceramic Filter*	*Earplugs with Ceramic Filter - Shush*	*Shush Earplugs - Hear What You've*	*Earplugs with Ceramic Filter - Superior Sound*	*Earplugs with Ceramic Filter for Live Music*	*Earplugs with Ceramic Filter for Motorcyclists*
Teamoy Sewing Machine Case, Travel Tote Bag ...	Teamoy Sewing Machine Carrying Case, Sewing ...	*Sewing Machine Carry Made Easy*	*Protect Your Sewing Machine in Style*	*Sewing Machine Tote Bag for Home*	*Easy to Carry Your Sewing Machine*	*Carry your sewing machine wherever you go*	*Carry your sewing machine wherever you go*
Happyluxe Travel Pillow | Small Pillow for Neck ...	Happyluxe Travel Pillow | Small Pillow for Neck ...	*The Perfect Pillow for Travel and Lounging*	*The Perfect Pillow for Neck Pain Relief*	*The Perfect Little Pillow For Your Little One*	*The Perfect American Made Travel Pillow for Camping*	*The Perfect American Made Travel Pillow for You*	*The Perfect American Made Travel Pillow for You*
Tria Age-Defying Smooth Beauty Laser ...	Tria Beauty Age-Defying Eye Wrinkle ...	*Anti Wrinkle, Anti Aging FDA Cleared Laser*	*Wrinkle Correcting Laser for Younger Skin*	*Age defying skincare that doesn't look expensive*	*Age Defying Beauty Laser That Actually Works*	*Age defying treatments that actually work*	*Age defying skincare laser technology*
TRX PRO3 Suspension Trainer System Design & Durability| ...	TRX GO Suspension Trainer System: Lightweight ...	*Full body training at home, outdoors or on the go*	*Portable full body workout at home or outdoors*	*Get a Full Body Workout At Home or On*	*GetGet the ultimate go-anywhere gym*	*Get Ready For Summer? Get Your Suspension System*	*Get Ready To Get Your Suspension Starter Kit*

Table 4: Generated headlines across different models for multiple input products (First 2 of all products shown)

Product Title 1	Product Title 2	BART - LP 0.25 (Short)	BART - LP 1 (Medium)	BART - LP 1.5 (Long)	COBART ϕ_{Length} (Short)	COBART ϕ_{Length} (Medium)	COBART ϕ_{Length} (Long)
Kids Easy Lazy Halloween Costume Shirt Skeleton ...	Womens Mens Easy Lazy Halloween Costume Shirt ...	*Fun Easy Lazy Halloween Shirts*	*Easy Lazy Halloween Costume Shirts*	*Easy Lazy Halloween Costume Shirts*	*Fun Halloween costume shirts*	*Fun Halloween shirts for the whole family*	*Fun easy lazy costume shirts for kids and adults*
Thompson's TH.010502-18 Waterseal Fabric Seal - Aersol	Thompson's TH.087731-42 WaterSeal Oxy Foaming ...	*Shop Thompson's WaterSeal Products*	*Shop Thompson's WaterSeal Products Today*	*Shop Thompson's WaterSeal Products Today*	*Protect Your Surfaces & Furniture*	*Protect Your Surfaces & Furniture from Water Damage*	*Protect Your Surfaces & Furniture From Water & Dirt*
Capsuline Clear Gelatin Empty Capsules 000 1000 ...	Capsuline Colored Gelatin Empty Capsules Size 0 Red ...	*Capsuline Colored Gelatin Capsules*	*Make Your Own Supplements with Capsuline Gelatin*	*Make Your Own Supplements with Capsuline Gelatin*	*Make Your Own Supplements*	*Make Your Own Supplements with Capsuline Capsules*	*Make your own Supplements at home with Capsuline*
Newport Vessels 8-Feet 10-Inch Dana Inflatable Sport ...	Newport Vessels 9-Feet 6-Inch Del Mar Inflatable Sport ...	*Inflatable Sport Tender Dinghy Boats*	*Inflatable Sport Tender Dinghy Boats*	*Inflatable Sport Tender Dinghy Boats*	*Sport Tender Dinghy Boats*	*Sport Tender Dinghy Boats from Newport Vessels*	*Sport Tender Dinghy Boats - Fun for All Ages*

Table 5: Generation with different lengths using Length Penalty (LP) with BART and COBART (First 2 of all products shown)

short headlines 99.83% times while still yielding CTR improvement of 6.7%. Table 1 illustrates this joint control over both the length and CTR (3 chosen buckets). This enables automated ad headline generation that works across all ad formats, requirements, and also optimizes CTR using a single model.

7 CONCLUSION

We propose a novel solution to the challenging and high-impact problem of ad headline generation. Our proposed method is able to control and optimize ad headline generation by using control tokens, SCST and the BART model. We compare our model to strong baselines and ablations and demonstrate its efficacy both quantitatively and qualitatively. We would continue to evaluate the model behavior over an extended period and experiment with more characteristics control tokens.

ACKNOWLEDGMENT

Thanks to *Ramaiah M S* and *Manisha Verma* for proof-reading and scrutinizing the paper.

REFERENCES

[1] Vamsi Aribandi, Yi Tay, Tal Schuster, Jinfeng Rao, Huaixiu Steven Zheng, Sanket Vaibhav Mehta, Honglei Zhuang, Vinh Q. Tran, Dara Bahri, Jianmo Ni, Jai Gupta, Kai Hui, Sebastian Ruder, and Donald Metzler. 2021. ExT5: Towards Extreme Multi-Task Scaling for Transfer Learning. *arXiv:2111.10952 [cs]* (Nov. 2021). http://arxiv.org/abs/2111.10952 arXiv: 2111.10952.

[2] AWS. [n.d.]. AWS Neuron - Amazon Web Services. https://aws.amazon.com/machine-learning/neuron/

[3] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. *arXiv:2005.14165 [cs]* (July 2020). http://arxiv.org/abs/2005.14165 arXiv: 2005.14165.

[4] Sumanth Dathathri, Andrea Madotto, Janice Lan, Jane Hung, Eric Frank, Piero Molino, Jason Yosinski, and Rosanne Liu. 2020. Plug and Play Language Models: A Simple Approach to Controlled Text Generation. *arXiv:1912.02164 [cs]* (March 2020). http://arxiv.org/abs/1912.02164 arXiv: 1912.02164 version: 4.

[5] Li Dong, Nan Yang, Wenhui Wang, Furu Wei, Xiaodong Liu, Yu Wang, Jianfeng Gao, Ming Zhou, and Hsiao-Wuen Hon. 2019. Unified Language Model Pre-training for Natural Language Understanding and Generation. *arXiv:1905.03197 [cs]* (Oct. 2019). http://arxiv.org/abs/1905.03197 arXiv: 1905.03197.

[6] Daniil Gavrilov, Pavel Kalaidin, and Valentin Malykh. 2019. Self-attentive Model for Headline Generation. In *Advances in Information Retrieval (Lecture Notes in Computer Science)*, Leif Azzopardi, Benno Stein, Norbert Fuhr, Philipp Mayr, Claudia Hauff, and Djoerd Hiemstra (Eds.). Springer International Publishing, Cham, 87–93. https://doi.org/10.1007/978-3-030-15719-7_11

[7] Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. DeBERTa: Decoding-enhanced BERT with Disentangled Attention. *arXiv:2006.03654 [cs]* (Oct. 2021). http://arxiv.org/abs/2006.03654 arXiv: 2006.03654.

[8] Zhiting Hu, Haoran Shi, Bowen Tan, Wentao Wang, Zichao Yang, Tiancheng Zhao, Junxian He, Lianhui Qin, Di Wang, Xuezhe Ma, Zhengzhong Liu, Xiaodan Liang, Wangrong Zhu, Devendra Singh Sachan, and Eric P. Xing. 2019. Texar: A Modularized, Versatile, and Extensible Toolkit for Text Generation. *arXiv:1809.00794 [cs]* (July 2019). http://arxiv.org/abs/1809.00794 arXiv: 1809.00794.

[9] Zhiting Hu, Zichao Yang, Xiaodan Liang, Ruslan Salakhutdinov, and Eric P. Xing. 2018. Toward Controlled Generation of Text. *arXiv:1703.00955 [cs, stat]* (Sept. 2018). http://arxiv.org/abs/1703.00955 arXiv: 1703.00955.

[10] J. Weston Hughes, Keng-hao Chang, and Ruofei Zhang. 2019. Generating Better Search Engine Text Advertisements with Deep Reinforcement Learning. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. ACM, Anchorage AK USA, 2269–2277. https://doi.org/10.1145/3292500.3330754

[11] Di Jin, Zhijing Jin, Joey Tianyi Zhou, Lisa Orii, and Peter Szolovits. 2020. Hooks in the Headline: Learning to Generate Headlines with Controlled Styles. *arXiv:2004.01980 [cs]* (May 2020). http://arxiv.org/abs/2004.01980 arXiv: 2004.01980 version: 3.

[12] Yashal Shakti Kanungo, Sumit Negi, and Aruna Rajan. 2021. Ad Headline Generation using Self-Critical Masked Language Model. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Industry Papers*. Association for Computational Linguistics, Online, 263–271. https://doi.org/10.18653/v1/2021.naacl-industry.33

[13] Nitish Shirish Keskar, Bryan McCann, Lav R. Varshney, Caiming Xiong, and Richard Socher. 2019. CTRL: A Conditional Transformer Language Model for Controllable Generation. *arXiv:1909.05858 [cs]* (Sept. 2019). http://arxiv.org/abs/1909.05858 arXiv: 1909.05858 version: 2.

[14] Diederik P. Kingma and Jimmy Ba. 2017. Adam: A Method for Stochastic Optimization. *arXiv:1412.6980 [cs]* (Jan. 2017). http://arxiv.org/abs/1412.6980 arXiv: 1412.6980.

[15] Sachin Kumar, Eric Malmi, Aliaksei Severyn, and Yulia Tsvetkov. 2021. Controlled Text Generation as Continuous Optimization with Multiple Constraints. *arXiv:2108.01850 [cs]* (Aug. 2021). http://arxiv.org/abs/2108.01850 arXiv: 2108.01850.

[16] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. *arXiv:1910.13461 [cs, stat]* (Oct. 2019). http://arxiv.org/abs/1910.13461 arXiv: 1910.13461.

[17] Chin-Yew Lin. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out*. Association for Computational Linguistics, Barcelona, Spain, 74–81. https://aclanthology.org/W04-1013

[18] Kenton Murray and David Chiang. 2018. Correcting Length Bias in Neural Machine Translation. *arXiv:1808.10006 [cs]* (Aug. 2018). http://arxiv.org/abs/1808.10006 arXiv: 1808.10006.

[19] Weizhen Qi, Yu Yan, Yeyun Gong, Dayiheng Liu, Nan Duan, Jiusheng Chen, Ruofei Zhang, and Ming Zhou. 2020. ProphetNet: Predicting Future N-gram for Sequence-to-Sequence Pre-training. *arXiv:2001.04063 [cs]* (Oct. 2020). http://arxiv.org/abs/2001.04063 arXiv: 2001.04063 version: 3.

[20] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *arXiv:1910.10683 [cs, stat]* (July 2020). http://arxiv.org/abs/1910.10683 arXiv: 1910.10683.

[21] Steven J. Rennie, Etienne Marcheret, Youssef Mroueh, Jarret Ross, and Vaibhava Goel. 2017. Self-critical Sequence Training for Image Captioning. *arXiv:1612.00563 [cs]* (Nov. 2017). http://arxiv.org/abs/1612.00563 arXiv: 1612.00563.

[22] Sascha Rothe, Shashi Narayan, and Aliaksei Severyn. 2020. Leveraging Pre-trained Checkpoints for Sequence Generation Tasks. *arXiv:1907.12461 [cs]* (April 2020). http://arxiv.org/abs/1907.12461 arXiv: 1907.12461.

[23] Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get To The Point: Summarization with Pointer-Generator Networks. *arXiv:1704.04368 [cs]* (April 2017). http://arxiv.org/abs/1704.04368 arXiv: 1704.04368.

[24] Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2019. MASS: Masked Sequence to Sequence Pre-training for Language Generation. *arXiv:1905.02450 [cs]* (June 2019). http://arxiv.org/abs/1905.02450 arXiv: 1905.02450 version: 5.

[25] Yun-Zhu Song, Hong-Han Shuai, Sung-Lin Yeh, Yi-Lun Wu, Lun-Wei Ku, and Wen-Chih Peng. 2020. Attractive or Faithful? Popularity-Reinforced Learning for Inspired Headline Generation. *Proceedings of the AAAI Conference on Artificial Intelligence* 34, 05 (April 2020), 8910–8917. https://doi.org/10.1609/aaai.v34i05.6421 Number: 05.

[26] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention Is All You Need. *arXiv:1706.03762 [cs]* (Dec. 2017). http://arxiv.org/abs/1706.03762 arXiv: 1706.03762.

[27] Ronald J. Williams. 1992. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning* 8, 3 (May 1992), 229–256. https://doi.org/10.1007/BF00992696

[28] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. HuggingFace's Transformers: State-of-the-art Natural Language Processing. *arXiv:1910.03771 [cs]* (July 2020). http://arxiv.org/abs/1910.03771 arXiv: 1910.03771 version: 5.

[29] Haoran Xu, Sixing Lu, Zhongkai Sun, Chengyuan Ma, and Chenlei Guo. 2021. VAE based Text Style Transfer with Pivot Words Enhancement Learning. *arXiv:2112.03154 [cs]* (Dec. 2021). http://arxiv.org/abs/2112.03154 arXiv: 2112.03154.

[30] Peng Xu, Chien-Sheng Wu, Andrea Madotto, and Pascale Fung. 2019. Clickbait? Sensational Headline Generation with Auto-tuned Reinforcement Learning. *arXiv:1909.03582 [cs]* (Sept. 2019). http://arxiv.org/abs/1909.03582 arXiv: 1909.03582 version: 1.

[31] Kevin Yang and Dan Klein. 2021. FUDGE: Controlled Text Generation With Future Discriminators. *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (2021), 3511–3535. https://doi.org/10.18653/v1/2021.naacl-main.276 arXiv: 2104.05218.

[32] Ruqing Zhang, Jiafeng Guo, Yixing Fan, Yanyan Lan, and Xueqi Cheng. 2020. Structure Learning for Headline Generation. *Proceedings of the AAAI Conference on Artificial Intelligence* 34, 05 (April 2020), 9555–9562. https://doi.org/10.1609/aaai.v34i05.6501 Number: 05.

A REPRODUCING EXPERIMENTS ON YOUR DATA

(1) Process your data and convert it to the HuggingFace Dataset format (Link).
(2) Compute observed CTR, length or any other desired control tag and obtain/train an oracle model if needed
(3) Setup Summarization / Generation script that uses HuggingFace Transformers (Link)
  (a) BART (Link)
  (b) T5 (Link)
  (c) ProphetNet (Link)
  (d) UniLM (Link)
(4) COBART
  (a) Add control tag prefixes to the input either as known tokens or extended tokens
  (b) Use the Summarization / Generation script for training
(5) SCBART
  (a) Calculate the reward function using the oracle model (Link)
  (b) Compute $\mathcal{L}_{RL}$ using headline generated using sampling.
  (c) Use convex combination of $\mathcal{L}_{RL}$ and $\mathcal{L}_{BART}$
  (d) Update the parameters
(6) VBART
  (a) Refer to the VAE auto-encoder implementation (Link)
  (b) Encode the input to X' and decode to generate the headline
  (c) Use the loss without the auto-encoder component
(7) SC-COBART
  (a) Follow COBART steps
  (b) Follow SCBART steps to use convex combination of $\mathcal{L}_{RL}$ and $\mathcal{L}_{COBART}$