Can Pretrained Language Models Generate Persuasive, Faithful, and Informative Ad Text for Product Descriptions?

Fajri Koto[1] **Jey Han Lau**[1] **Timothy Baldwin**[1,2]
[1]The University of Melbourne
[2]MBZUAI
ffajri@student.unimelb.edu.au, jeyhan.lau@gmail.com, tb@ldwin.net

Abstract

For any e-commerce service, persuasive, faithful, and informative product descriptions can attract shoppers and improve sales. While not all sellers are capable of providing such interesting descriptions, a language generation system can be a source of such descriptions at scale, and potentially assist sellers to improve their product descriptions. Most previous work has addressed this task based on statistical approaches (Wang et al., 2017), limited attributes such as titles (Chen et al., 2019; Chan et al., 2020), and focused on only one product type (Wang et al., 2017; Munigala et al., 2018; Hong et al., 2021). In this paper, we jointly train image features and 10 text attributes across 23 diverse product types, with two different target text types with different writing styles: bullet points and paragraph descriptions. Our findings suggest that multimodal training with modern pretrained language models can generate fluent and persuasive advertisements, but are less faithful and informative, especially out of domain.

TITLE: Lantern Press Philadelphia - Skyline with Comcast Tower
PART NUMBER: LANT-99520-10x15W
MODEL NUMBER: LANT-99520-10x15W
CATEGORY: HOME
COLOR: Multi
BRAND: Lantern Press
SIZE: 10 x 15 Wood Sign
CLASSIFICATION: base_product
WEIGHT: 13.0
KEYWORD: Decor wooden sign wall decoration

BULLET POINTS

- Philadelphia, Pennsylvania - Skyline with Comcast Tower - Abstract (Cream) 99520 (10x15 Wood Wall Sign, Wall Decor Ready to Hang)
- Measures 10x15 inches
- Holes in corners, ready for hanging, Printed in the USA, sustainable birch
- Perfect for your home, office, or a gift

PARAGRAPH DESCRIPTION

This original high-quality wood print from Lantern Press boasts sharp detail and vivid imagery of Philadelphia, Pennsylvania - Skyline with Comcast Tower - Abstract (Cream) 99520 (10x15 Wood Wall Sign, Wall Decor Ready to Hang). Product measures 10 x 15 inches 100% Printed in America, "Grade A" sustainable birch Holes in corners, ready for tacker sign to be hung Wood print will ship in a sturdy box, protected in a water-proof sleeve. Lantern Press is a dynamic art company that specializes in the world's leading imagery.

Figure 1: **Top**: Input consisting of an image and textual attributes of a product. **Bottom**: Two target texts: bullet points and a paragraph description

1 Introduction

Generative pretrained language models such as GPT-2 (Radford et al., 2019), T5 (Raffel et al., 2020), and BART (Lewis et al., 2020a) have led to impressive gains in language generation applications beyond machine translation, such as story geneneration (Fan et al., 2018; Goldfarb-Tarrant et al., 2020), summarization (Zhang et al., 2020; Qi et al., 2020), and dialogue systems (Ham et al., 2020). Although such transformer-based language models (Vaswani et al., 2017) are capable of generating fluent texts through a sequence-to-sequence framework, they still suffer from unfaithfulness and factuality issues (Maynez et al., 2020; Wang et al., 2020; Moradi et al., 2021).

In this paper, we comprehensively discuss the utility of modern pretrained language models over an ad text generation task for product descriptions, with a focus on faithfulness, persuasiveness, and informativeness. While previous work has been limited to short ad generation tasks conditioned on titles (Chen et al., 2019; Chan et al., 2020), and used traditional neural models (Munigala et al., 2018; Zhang et al., 2019a) or statistical approaches (Wang et al., 2017), we focus on a data-to-text generation approach to product description generation for an English e-commerce service. Specifically, we explore various textual attributes and images as the input, and generate two types of product descriptions: (1) bullet points, and (2) paragraph descriptions (see Figure 1). Bullet points provide a list of key information regarding a product, while paragraph descriptions are made up of sentences structured into a coherent narrative.

We argue there are two underlying motivations for the ad text generation task, especially for product descriptions. Application-wise, the utility is to improve the seller experience for e-commerce services when registering a new product. The generated descriptions can reduce the need for manual data entry, and potentially improve sales due to better descriptions (in terms of attractiveness, structure, and persuasiveness). Research-wise, ad

Proceedings of The Fifth Workshop on e-Commerce and NLP (ECNLP 5), pages 234 - 243
May 26, 2022 ©2022 Association for Computational Linguistics

text generation is an under-studied task, and arguably a good proxy for persuasive text generation (Wei et al., 2016; Rehbein, 2019; Luu et al., 2019; El Baff et al., 2020).

While previous work has discussed ad text generation of e-commerce service for a few product types such as fashion (Munigala et al., 2018), computers (Wang et al., 2017), and house decor (Hong et al., 2021), in this work, we use twenty diverse product types and an additional three product types for out-of-domain prediction. With this setting, we aim to study model generalization and robustness over in-domain and out-of-domain test sets.

To summarize our contributions: (1) we study the application of modern pretrained language models based on data-to-text generation for product description in an e-commerce service; (2) we explore multimodal training by incorporating image features for ad generation and perform automatic and manual evaluation; (3) we study model robustness for out-of-domain prediction; and (4) we conduct analysis of attributes that significantly contribute to ad text generation.

2 Related Work

Data-to-text generation is the task of translating a semi-structured table to natural text, and has been applied in different real-world scenarios, such as weather forecasting reports (Liang et al., 2009), sport (Puduppully et al., 2019), health-care descriptions (Hasan and Farri, 2019), and biographies (Wang et al., 2020). While the goal of most previous tasks is to generate descriptive text, there are few studies (Wang et al., 2017) on data-to-text generation for the advertisement domain, and the work that has been done has tended to focus exclusively on the product type of *computer* and be based on pre-neural statistical approaches and template-based techniques.

Previous work has mostly used titles of e-commerce products to generate short ads in Chinese (Chen et al., 2019; Chan et al., 2020) and English (Munigala et al., 2018; Kanungo et al., 2021). Similarly, Zhang et al. (2019a) generate a product description for Chinese e-commerce, conditioned on the title and a small number of attributes (with an average length of six words).[1] In this work, we comprehensively study product description generation in English based on ten diverse attributes (*à la* a data-to-text scheme, with the average number of concatenated attributes being 64 words in Table 1) that incorporates joint training over images of the product.

Attributes	**Coverage**	**#words**			**|Vocab|**
	(%)	**max**	μ	σ	
TITLE	100	95	15.7	6.22	193,649
PRODUCT TYPE	100	1	1	0	20
CLASSIFICATION	100	1	1	0	3
BRAND	99.49	17	1.58	0.88	46, 552
KEYWORD	92.17	958	32.32	55.72	292,372
COLOR	80.19	32	1.44	1.01	18,839
SIZE	69.96	16	1.82	1.44	15,187
MODEL NUMBER	33.75	9	1.15	0.52	67,215
PART NUMBER	47.64	12	1.08	0.41	91,084
WEIGHT	20.76	1	1	0	1,786
BULLET POINTS	100	766	86.8	67.9	225,784
PARAGRAPH DESC.	100	516	90.9	72.9	472,711

Table 1: Statistics of attributes. For BULLET POINTS, the average number of bullets in the overall dataset is 5.

Component		**% of novel *n*-grams**			
A	**B**	**1**	**2**	**3**	**4**
10 attr.	BP	86.7	96.3	98.1	98.7
10 attr.	PD	85.1	93.7	95.2	95.9
BP	PD	66.2	86.9	90.9	92.7

Table 2: Abstractiveness of BULLET POINTS (BP) and PARAGRAPH DESCRIPTIONS (DP) based on novel *n*-gram overlap. "10 attr." means the concatenatenation of all attributes, and values in the table are calculated relative to component B.

3 Data Construction

We use 200,000 e-commerce products spanning 20 different product types as described in Figure 2. For copyright reasons we are not able to release this data to the public. This dataset is randomly split into 180K/10K/10K training, development and test instances, respectively. We also create an Xtreme test set (4,266 samples) in which we filter out test samples that have overlapping descriptions with the training data. Lastly, we additionally use three different product types as an out-of-domain test set, comprising 1,000 products of each of the three produce types: SAREE, COMPUTER, and CELLULAR_PHONE. In total, there are three different test sets: (1) main; (2) Xtreme; and (3) out-of-domain.

In Table 1, we show the overall statistics of ten product atttributes and two target texts: BULLET POINTS and PARAGRAPH

[1] These attributes are not clearly described in the paper.

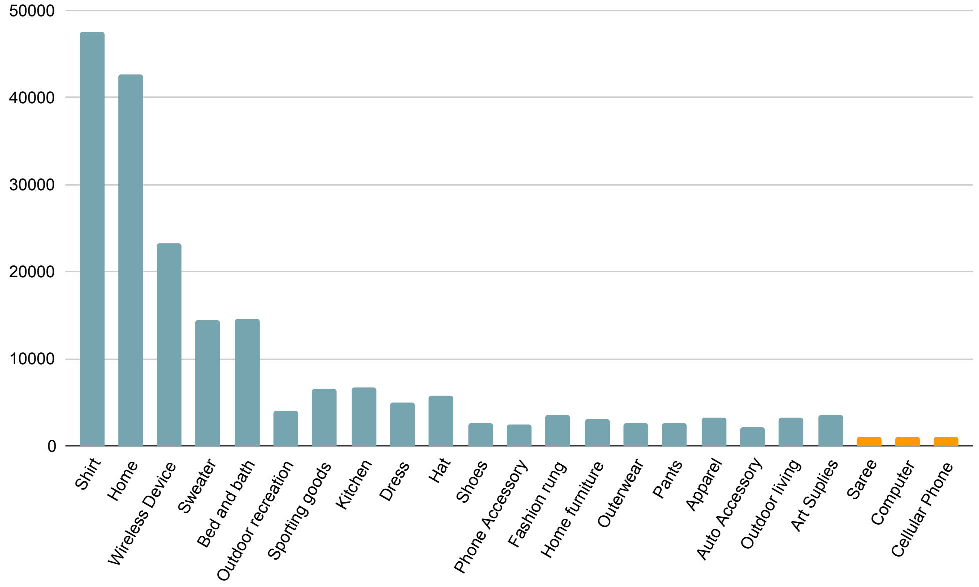

Figure 2: Distribution of 20 product types in the main dataset and 3 additional product types from the out-of-domain test set. The main and additional data is in English, and gathered from different regions (countries).

DESCRIPTIONS. The selection of product attributes is based on a minimum coverage of 20% in the dataset. Overall, the five attributes with the highest coverage are TITLE, PRODUCT TYPE, CLASSIFICATION, BRAND, and KEYWORD.[2] The average length of BULLET POINTS and PARAGRAPH DESCRIPTIONS is 87 and 91, respectively, significantly longer than most previous work except Wang et al. (2017) who focused on the product type of *computer* and tested only pre-neural statistical approaches (see Table 3).

To understand the abstractiveness of our dataset, in Table 2 we show the percentage of novel *n*-grams in BULLET POINTS and PARAGRAPH DESCRIPTIONS. Overall, we observe that the two target texts are highly abstractive, with more than 85% of novel *n*-grams, computed relative to the concatenated attributes. We also found that there is a high proportion of novel *n*-grams between the two target texts.[3] We suspect, though, that the low lexical overlap between the two text types in this task might not be attributed to paraphrasing or lexical choice, but rather to content selection.

Work	Lang.	Product Types	#words of source (μ)	#words of target (μ)
Zhang et al. (2019a)	ZH	N/A	18	25
Chan et al. (2020)	ZH	N/A	18	22
Hong et al. (2021)	ZH	1	N/A	76
Wang et al. (2017)	EN	1	N/A	117
Munigala et al. (2018)	EN	1	6	18
Kanungo et al. (2021)	EN	1	19	6
This work	EN	23	64	87 & 91

Table 3: Dataset comparison between our work and previous work

4 Model

Problem Formulation. As discussed in Section 3, a product in our dataset consists of up to ten attributes $\{a_1, a_2, a_3, ..., a_{10}\}$, one image I, and two target texts $\{t_1, t_2\}$. The goal of this work is to learn a function that estimates the probabilities $P(t_1|a_1, a_2, a_3, ..., a_{10}, I)$ and $P(t_2|a_1, a_2, a_3, ..., a_{10}, I)$.

Architecture. This work relies on pretrained language models such as BERT (Devlin et al., 2019), T5 (Raffel et al., 2020), and BART (Lewis et al., 2020a). To perform data-to-text generation, we formulate a structured input based on special tokens that are randomly initialized before the fine-tuning. The textual input is the concatenation of each attribute preceded by each corresponding special token (see Figure 3).

To accommodate multimodal training, we fol-

[2] CLASSIFICATION means other categories such as base product or different variation.

[3] We also note that no previous work reported on the abstractiveness of their data.

low Xing et al. (2021) in extracting n **R**egions **o**f **I**nterest (RoIs) (i.e. bounding boxes) of the image using `detectron2`, a pretrained masked R-CNN (He et al., 2017).[4] Formally, an Image I is chunked by `detectron2` into $\{\text{RoI}_1, \text{RoI}_2, ..., \text{RoI}_n\}$. We obtained a fixed-size latent representation of each RoI based on intermediate features of `detectron2` (ResNet-101 (He et al., 2016)). To align the embedding size with pretrained language models we use a linear layer. Similar to the textual input, we also introduce a special token `[IMAGE]` that is concatenated at the beginning of the input.

For the target texts, we introduce special tokens `[BULLET POINTS]` and `[DESCRIPTION]` as the start token. Specifically, for bullet points, we concatenate all points with token <q> as the separator. Finally, for the encoder-decoder, we use BERT-base with raw decoder following (Liu and Lapata, 2019), BART-base, and T5-base, and train the model with standard cross-entropy loss.

5 Experiments

5.1 Set-Up

We experiment in three settings: (1) training with the text input only; (2) training with the image features only; and (3) multimodal training incorporating both text and image features, as depicted in Figure 3. For the text features, we encode the text using the three pretrained LMs of BERT, BART, and T5, while for the other two we only experimented with BART because of its higher performance in the first experiment. For image feature extraction, we experimented with $\{10, 20, 30, 40, 50\}$ RoIs, and tuned based on the development set. We report results of 50 and 20 RoIs for the second and third experiment, respectively.

For `TITLE`, `KEYWORD`, and other attributes, we set the maximum token length to 30, 100, and 10 based on the statistics in Table 1. This results in a maximum token length of 220 for the source text (including the special tokens). For the two target texts, we set the maximum token length to 250, and train them separately. Our preliminary experiments show that performing multi-task training (i.e. using both target texts at the same time) performs worse than single-task training.

We use the huggingface PyTorch framework (Wolf et al., 2020) for our experiments with three pretrained language models: BERT-base[5] (Devlin et al., 2019), T5-base[6] (Raffel et al., 2020), and BART-base[7] (Lewis et al., 2020a). All experiments are run on 4×V100 16GB GPUs.

For the BERT model, we follow Liu and Lapata (2019) in adding a randomly-initialized transformer decoder (layers = 6, hidden size = 768, feed-forward = 2,048, and heads = 8) on top of BERT, and train it for 200K steps. We use the Adam optimizer and learning rate $lr = 2e^{-3} \times \min(\text{step}^{-0.5}, \text{step} \times 20,000^{-1.5})$ and $0.1 \times \min(\text{step}^{-0.5}, \text{step} \times 10,000^{-1.5})$ for BERT and the transformer decoder, respectively. We use a warmup of 20,000, a dropout of 0.2, a batch size total of 200 (10 × 4 GPUs × gradient accumulation of 5), and save checkpoints every 10,000 steps. We compute ROUGE scores (R1) to pick the best checkpoint based on the development set.

For T5 and BART, we train them for 30 epochs (around 20K steps) with an initial learning rate of $1e^{-4}$ (Adam optimizer). We use a total batch size of 300 (15 × 4 GPUs × gradient accumulation of 5), a warmup of 10% of total steps, and save checkpoints for every 1,000 steps. We also compute ROUGE scores (R1) to pick the best checkpoint based on the development set.

5.2 Evaluation

As discussed in Section 3, we use three different test sets: main, Xtreme, and out-of-domain. For automatic evaluation, we use ROUGE-1/2/L (Lin, 2004), BLEU-4 (Papineni et al., 2002), METEOR (Banerjee and Lavie, 2005), and BERTScore (Zhang et al., 2019b). For BERTScore we compute the F1 score using `roberta-large` (layer 17) as recommended by Zhang et al. (2019b).

For manual evaluation, we first obtain 50 random samples for each of the three test sets, ensuring there is no overlap between the main and Xtreme test sets. We hire four expert workers with Master degree qualifications to annotate four descriptions for each product: (1) gold; (2) BART; (3) BART+image; and (4) image only. The total number of annotations is 2 workers × 4 models × 150 samples × 2 descriptions = 2,400 annotations. One worker was asked to work on either bullet points or paragraph descriptions, and was paid $50.

There are five aspects that are manually evaluated by our workers: (1) Fluency: the description is fluent and grammatically correct; (2) At-

[4] https://github.com/facebookresearch/detectron2

[5] `bart-base-uncased`

[6] `t5-base`

[7] `facebook/bart-base`

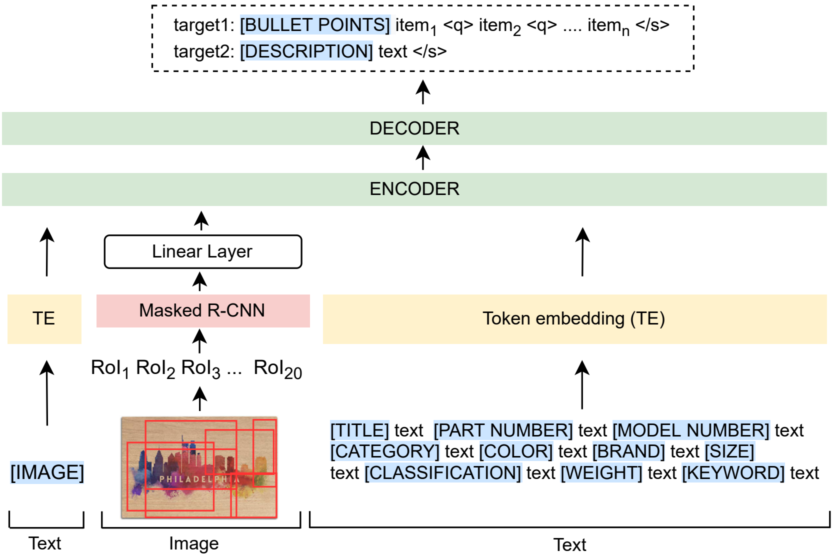
target1: [BULLET POINTS] item$_1$ <q> item$_2$ <q> .... item$_n$ </s>
target2: [DESCRIPTION] text </s>
DECODER
ENCODER
Linear Layer
TE
Masked R-CNN
Token embedding (TE)
RoI$_1$ RoI$_2$ RoI$_3$ ... RoI$_{20}$
[IMAGE]
PHILADELPHIA
[TITLE] text [PART NUMBER] text [MODEL NUMBER] text [CATEGORY] text [COLOR] text [BRAND] text [SIZE] text [CLASSIFICATION] text [WEIGHT] text [KEYWORD] text
Text
Image
Text

Figure 3: Model architecture used in this work.

Model	**Main Test**						**Xtreme Test**						**Out-of-domain Test**						**Avg.**
	R-1	**R-2**	**R-L**	**B-4**	**M**	**BS**	**R-1**	**R-2**	**R-L**	**B-4**	**M**	**BS**	**R-1**	**R-2**	**R-L**	**B-4**	**M**	**BS**	
BULLET POINTS																			
BERT	51.8	40.7	50.6	32.6	25.0	88.7	35.1	20.7	33.5	15.6	14.1	85.5	11.9	2.4	11.1	1.5	3.3	80.5	33.6
T5	45.4	34.7	44.1	30.6	28.9	87.4	28.2	30	31.1	13.4	13.3	83.4	12.6	4.2	10.8	2.3	4.4	79.2	32.4
BART	58.9	48.5	57.7	43.5	38.5	90.7	39.8	24.8	38.1	20.8	19.6	86.5	**17.5**	**6.1**	**16.5**	**3.0**	**5.0**	**81.5**	38.7
Image only	43.4	30.9	32.3	27.5	25.3	87.3	27.6	13.5	26.1	11.4	11.6	83.8	9.9	0.7	9.0	0.8	2.9	79.4	29.1
BART+Image	**59.3**	**48.9**	**58.1**	**43.7**	**38.6**	**90.8**	**40.1**	**24.9**	**38.4**	**21.0**	**19.7**	**86.6**	**17.5**	5.9	16.4	2.8	4.9	81.4	**38.8**
PARAGRAPH DESCRIPTIONS																			
BERT	41.0	30.1	35.5	24.2	19.4	86.4	27.5	15	21.4	10.9	10.5	83.3	10.9	1.5	7.2	0.9	2.5	79.6	28.2
T5	40.7	31.9	36.7	28.6	29.3	85.8	23.8	14.2	19.8	11.5	12.6	81.2	10.8	4.4	9.1	2.2	4.6	77.8	29.2
BART	54.8	45.1	50.1	40.2	37.7	90.1	**36.1**	**22.6**	**29.5**	**18.3**	**18.2**	**85.9**	16.5	**5.6**	12.0	**2.8**	5.5	81.1	36.2
Image only	41.1	29.4	35.4	26.6	25.6	86.9	24.7	10.9	18.2	9.1	10.1	83.1	12.2	0.8	7.4	0.8	3.3	79.9	28.1
BART+Image	**54.9**	**45.3**	**50.3**	**40.4**	**37.9**	**90.2**	35.8	22.4	29.3	**18.3**	18.0	85.8	**17.2**	**5.6**	**12.3**	2.7	**5.6**	**81.5**	**36.3**

Table 4: Main experimental results of automatic metrics. R-1, R-2, R-L, B-4, M, and BS are ROUGE-1, ROUGE-2, and ROUGE-3, BLEU-4, METEOR, and BERTScore, respectively.

tractiveness: the description is interesting and eye-catching; (3) Persuasive words: the description uses persuasive words or phrases; (4) Faithfulness: information in the description is captured by the image and the attributes; and (5) Informativeness: the description is informative and complete relative to the available attributes. Except for the third aspect which is binary (yes/no), we use a slider scale with values between 0–100 for all aspects.

In manual evaluation, workers were presented the product image and list of text attributes with four different descriptions. The four descriptions are shuffled, so the model information of each description is not apparent to the worker. Workers were asked to carefully read each description, and then asked to put the evaluation scores in the available field.

5.3 Results

Table 4 shows the experimental results based on the automatic metrics. Overall, we observe similar trends for both `BULLET POINTS` and `PARAGRAPH DESCRIPTIONS`, namely that BART is substantially better than T5 and BERT across the three test sets. Using only image features for generating both ad text types yields a comparable score to T5, but tends to be lower for almost all test sets and metrics. The multimodal training (i.e. “BART+image”) slightly improves BART performance for the main test set, but achieves mixed results for the Xtreme and out-of-domain test sets with both `BULLET POINTS` and `PARAGRAPH DESCRIPTIONS`. We also observe that Xtreme and the out-of-domain test sets are harder, with high performance gaps, relative to the main test set.

Model	Main Test					Xtreme Test					Out-of-domain Test					Avg.
	Flu.	Att.	Per.	Fa.	Inf.	Flu.	Att.	Per.	Fa.	Inf.	Flu.	Att.	Per.	Fa.	Inf.	
BULLET POINTS																
Gold	0.62	0.59	0.77	0.54	0.52	0.58	0.56	0.74	0.57	0.55	0.62	0.58	0.51	0.59	0.60	0.59
Image only	0.63	0.61	**0.82**	0.43	0.43	0.65	0.61	**0.87**	0.43	0.44	**0.64**	**0.56**	**0.74**	0.13	0.27	0.53
BART	0.63	0.59	0.78	0.56	0.53	0.61	0.58	0.64	0.58	**0.56**	0.48	0.42	0.27	0.53	0.52	0.55
BART+image	**0.66**	**0.63**	0.79	**0.58**	**0.56**	**0.64**	**0.62**	0.72	**0.57**	0.54	0.44	0.42	0.30	**0.55**	**0.54**	**0.57**
PARAGRAPH DESCRIPTIONS																
Gold	0.82	0.52	0.29	0.60	0.48	0.74	0.46	0.31	0.53	0.47	0.84	0.49	0.18	0.52	0.48	0.44
Image only	0.77	0.43	0.29	0.42	0.38	**0.81**	0.43	**0.34**	0.43	0.41	**0.74**	0.25	**0.20**	0.17	0.16	0.33
BART	**0.82**	**0.53**	0.31	**0.62**	0.50	0.74	0.50	0.21	**0.60**	**0.53**	0.69	0.39	0.16	0.53	**0.42**	0.44
BART+image	0.81	**0.53**	**0.33**	0.60	**0.51**	0.76	**0.52**	0.23	0.59	**0.53**	0.71	**0.41**	0.15	**0.56**	0.41	**0.45**

Table 5: The primary experimental results for manual evaluation. Flu., Att., Per., Fa., and Inf. denote Fluency, Attractiveness, Persuasiveness, Faithfulness, and Informativeness, respectively. The presented scores are the average of two annotations. Entries in bold refer to the best overall score (excluding Gold texts).

Aspects	BULLET POINTS	DESCRIPTION
Fluency	0.51	0.50
Attractiveness	0.50	0.42
Persuasiveness	0.39	0.32
Faithfulness	0.51	0.41
Informativeness	0.34	0.45

Table 6: Pearson correlation scores between two annotators in manual evaluation. For persuasiveness we present the Kappa score.

For example, in BULLET POINTS, ROUGE-1 of BART drops substantially by -19.1 and -41.4 in the Xtreme and out-of-domain test sets, resp., implying that the model does not generalize well to different test sets.

In Table 6 we show the inter-annotator agreement of manual evaluation in the form of Pearson correlation for fluency, attractiveness, faithfulness, and informativeness; and the Kappa score for persuasiveness. Overall, we found that annotators have moderate correlation and agreement. In Table 5, scores of the Gold text can be interpreted as the upper bound of the manual evaluation. Note that for faithfulness and informativeness, these aspects are only evaluated based on the ten selected attributes.

For the main and Xtreme test sets in Table 5, most models generate fluent, attractive, persuasive, faithful, and informative texts for BULLET POINTS and PARAGRAPH DESCRIPTIONS, *relative to the performance of the gold texts*. When using only image features (the "image only" model), the model's faithfulness and informativeness decrease markedly, indicating the importance of textual attributes for this task. BART and BART+image models yield comparable results with the gold texts, with slightly better faithfulness and informativeness.[8]

For the out-of-domain test set, we observe that the human evaluation performance over the three models (Image only, BART, and BART+image) is generally lower than the gold text. Interestingly, we find that the "image only" model generates fluent and persuasive texts, but with substantially low faithfulness and informativeness. It is also worth mentioning that the BART model's performance is not as good as for the main test set, which indicates the out-of-domain challenge in applying models in real world scenarios.

In addition, we calculated the average performance of the manual evaluation, and found that the BART+image model performs best for both target texts. These results are in line with the averaged automatic evaluation scores in Table 4.

Based on the manual evaluation results in Table 5, the relatively low faithfulness scores for the gold texts (around 0.5–0.6) suggests that they contain new information that is not found in the input attributes. Although this means the gold texts are not faithful, they are likely to be still *factually correct*, as they are written by the product sellers (Maynez et al., 2020). Taking the faithfulness scores of the gold texts as the upper bound, we could conclude that the BART models are performing as well as they could (seeing that they are trained on not very faithful target texts in the first place). Ultimately, our results in this task high-

[8] These results are to be expected in the manual evaluation, since both aspects are only examined based on the ten selected attributes.

BULLET POINTS

TITLE: Yosoo Digital Clock Portable Electronic Bell with Backlight LCD Screen Display Alarm Clock Car Desk Table Decoration Clock(Pink)
PART NUMBER: Yosoo6wkf82orh7
CATEGORY: HOME
BRAND: Yosoo
CLASSIFICATION: base_product
WEIGHT: 43.0
KEYWORD: LCD Digital Clock, Digital Alarm Clock, Portable Digital Clock,Table Desk Digital Clock, Digital Clock

Gold:

- Large LCD screen for an easy and accurate reading of time Display the current time, week clearly, with alarm function and snooze function
- Press the light button at the back of the clock, the backlight will bright 3 seconds or so, convenient for you to read time at night
- Built with a clip and stand, you can put it on your desk or table
- Light weight and compact size, multi-colors are available. A good choice to improve concept of punctuality

BART:

- Delivery Time 10-15 Days
- Return Range 30 Days
- If you have any questions, please feel free to contact us and we will get back to you within 24 hours.
- The clock is made of high-quality materials, durable and easy to clean.

BART + image:

- This digital clock is made of high quality materials, durable and easy to clean.
- The clock is equipped with backlight LCD screen display, which is very convenient to read.
- It can be used in bedroom, living room, kitchen, office, etc. It is also a good gift for your family and friends.
- If you have any questions, please feel free to contact us and we will reply you within 24 hours.

PARAGRAPH DESCRIPTION

TITLE: LJSWG Men's Oklahoma Flag Oklahoma National T-shirt Asphalt L
CATEGORY: SHIRT
COLOR: Asphalt
BRAND: LJSWG
SIZE: Large
CLASSIFICATION: base_product
KEYWORD: Personalised Cheap Men Tee Casual Cotton, Girls Games Plus Size Clothing Captain America: Civil War AKON, Oklahoma City

Gold:

Men's Oklahoma Flag Oklahoma National Custom Text, ID, Name Or Message On High Quality Hanes Cotton T-Shirts. 100% Preshrunk Cotton Takes The Worry Away From Shrunk-age. Seamless Rib At Neck And Collar. Double-needle Stitching For Dorabiltiy. This Classic Crew Neckline T-shirtis Is Great For Every Occasion And Situation.The Design Is Printed With Advanced Printing Technology. It Is Printed With A Water-soluble And Eco-friendly Ink.Trendy, Brightly Colored Graphics. A Unique Gift Idea For A Friend Or Family Member.

BART:

100% Soft Cotton Fabric With A Soft Touch And Quality Printing Techniques.It Will Never Fade, Peel Or Crack And Can Be Machine Washed & Ironed.This Customized T Shirt Will Be The Best And Sincere Gift For Your Family,friends And Team.

BART + image:

Oklahoma Flag Oklahoma National T-shirt. Art Heat Press Print On Front. Wash Inside Out In Cold Water, Hand Dry Recommended. Most Of Our Designs Are Available In Men`s Sizes.Please Check Our Store For All Other Varieties.

Figure 4: Example of generated BULLET POINTS and PARAGRAPH DESCRIPTIONS.

lighted the fact that our current human faithfulness evaluation does not always capture factuality, prompting further questions on how we can assess this dimension, which we leave for future work.

Figure 4 depicts some example outputs of the BART models for BULLET POINTS and PARAGRAPH DESCRIPTIONS. The first example shows that the prediction of the BART+image model contains better content than the BART text-only model, with a description of the LCD screen and usage examples. Similarly in the second example, the BART+image model generates more specific content for the t-shirt product by mentioning *Flag Oklahoma National*.

6 Analysis

Which attributes contribute to ad generation? To answer this question, we performed an ablation study using the BART models. We decode both BULLET POINTS and PARAGRAPH DESCRIPTIONS using different numbers of attributes as context, and report the average automatic performance in Table 7.

We observe there are three prominent attributes for this task — TITLE, BRAND, and KEYWORD — for both BULLET POINTS and PARAGRAPH DESCRIPTIONS. Interestingly, using only TITLE can produce 32.98 and 29.93 average performance, and adding KEYWORD to the input boosts performance by 11.05 and 10.57, for BULLET POINTS and PARAGRAPH DESCRIPTIONS, respectively.

7 Discussion and Conclusion

In this work, we described the first attempt at multimodal training for ad generation by incorporating image representations and text embeddings as input. We found that multimodal training yields the best performance in terms of overall scores in the both automatic and manual evaluation. We observe that modern pretrained language models can generate fluent advertisements, but are less faithful and

Attributes (#Attr)	BULLET POINTS		PARAGRAPH DESCRIPTIONS	
	Avg.	Δ	Avg.	Δ
TITLE (1)	32.98	**32.98**	29.93	**29.93**
prev. + PRODUCT TYPE (2)	34.53	1.55	31.38	1.45
prev. + CLASSIFICATION (3)	34.53	0.00	31.38	0.00
prev. + BRAND (4)	39.75	**5.22**	39.33	**7.95**
prev. + KEYWORD (5)	50.80	**11.05**	49.90	**10.57**
prev. + COLOR (6)	51.63	0.83	50.15	0.25
prev. + SIZE (7)	53.15	1.52	51.03	0.88
prev. + PART NUMBER (8)	55.12	1.97	52.32	1.28
prev. + MODEL NUMBER (9)	56.30	1.18	52.77	0.45
prev. + WEIGHT (10)	56.30	0.00	52.77	0.00

Table 7: Ablation study on the main test set using BART by incrementally adding different attributes. Avg means the average score of ROUGE-1, ROUGE-2, ROUGE-L, BLEU-4, METEOR, and BERTScore. Δ means the difference score between the given and previous row. The bold entries are the top-3 highest Δ scores.

informative, especially in out-of-domain settings.

Can pretrained language models generate persuasive, faithful, and informative ad text for product descriptions? The answer to this question is *yes to a certain extent*, particularly for in-domain scenarios. And although the BART models have similar human faithfulness performance to the gold texts, we believe that it does not necessarily imply that they are factually correct and further validation is necessary. One way forward may be to allow human judges to have access to some external knowledge (e.g. search engines or product catalogues), which will help them assess the factuality of the generated texts.

Furthermore, since the product descriptions in our e-commerce dataset might introduce new information, retrieval augmented generation (Lewis et al., 2020b; Kim et al., 2020; Shuster et al., 2021) is one potential direction for future work. This is because information on some products is likely to be available on the Internet, and incorporating it into the generation model could potentially improve the resulting ad text.

References

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Zhangming Chan, Yuchi Zhang, Xiuying Chen, Shen Gao, Zhiqiang Zhang, Dongyan Zhao, and Rui Yan. 2020. Selection and generation: Learning towards multi-product advertisement post generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3818–3829, Online. Association for Computational Linguistics.

Qibin Chen, Junyang Lin, Yichang Zhang, Hongxia Yang, Jingren Zhou, and Jie Tang. 2019. Towards knowledge-based personalized product description generation in e-commerce. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 3040–3050.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 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)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Roxanne El Baff, Henning Wachsmuth, Khalid Al Khatib, and Benno Stein. 2020. Analyzing the Persuasive Effect of Style in News Editorial Argumentation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3154–3160, Online. Association for Computational Linguistics.

Angela Fan, Mike Lewis, and Yann Dauphin. 2018. Hierarchical neural story generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 889–898, Melbourne, Australia. Association for Computational Linguistics.

Seraphina Goldfarb-Tarrant, Tuhin Chakrabarty, Ralph Weischedel, and Nanyun Peng. 2020. Content planning for neural story generation with aristotelian rescoring. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4319–4338, Online. Association for Computational Linguistics.

Donghoon Ham, Jeong-Gwan Lee, Youngsoo Jang, and Kee-Eung Kim. 2020. End-to-end neural pipeline for goal-oriented dialogue systems using GPT-2. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 583–592, Online. Association for Computational Linguistics.

Sadid A Hasan and Oladimeji Farri. 2019. Clinical natural language processing with deep learning. In *Data Science for Healthcare*, pages 147–171. Springer.

Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. 2017. Mask R-CNN. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2961–2969.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778.

Yunsen Hong, Hui Li, Yanghua Xiao, Ryan McBride, and Chen Lin. 2021. SILVER: Generating persuasive Chinese product pitch. In *PAKDD (2)*, pages 652–663. Springer.

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*, pages 263–271, Online. Association for Computational Linguistics.

Jihyeok Kim, Seungtaek Choi, Reinald Kim Amplayo, and Seung-won Hwang. 2020. Retrieval-augmented controllable review generation. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 2284–2295, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020a. BART: Denoising sequence-to-sequence pretraining for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020b. Retrieval-augmented generation for knowledge-intensive NLP tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.

Percy Liang, Michael Jordan, and Dan Klein. 2009. Learning semantic correspondences with less supervision. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 91–99, Suntec, Singapore. Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Yang Liu and Mirella Lapata. 2019. Text summarization with pretrained encoders. 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)*, pages 3730–3740, Hong Kong, China. Association for Computational Linguistics.

Kelvin Luu, Chenhao Tan, and Noah A. Smith. 2019. Measuring online debaters' persuasive skill from text over time. *Transactions of the Association for Computational Linguistics*, 7:537–550.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1906–1919, Online. Association for Computational Linguistics.

Pooya Moradi, Nishant Kambhatla, and Anoop Sarkar. 2021. Measuring and improving faithfulness of attention in neural machine translation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2791–2802, Online. Association for Computational Linguistics.

Vitobha Munigala, Abhijit Mishra, Srikanth G Tamilselvam, Shreya Khare, Riddhiman Dasgupta, and Anush Sankaran. 2018. Persuaide! an adaptive persuasive text generation system for fashion domain. In *Companion Proceedings of the The Web Conference 2018*, pages 335–342.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Ratish Puduppully, Li Dong, and Mirella Lapata. 2019. Data-to-text generation with entity modeling. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2023–2035, Florence, Italy. Association for Computational Linguistics.

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-SequencePre-training. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2401–2410, Online. Association for Computational Linguistics.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

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. *Journal of Machine Learning Research*, 21:1–67.

Ines Rehbein. 2019. On the role of discourse relations in persuasive texts. In *Proceedings of the 13th Linguistic Annotation Workshop*, pages 144–154, Florence, Italy. Association for Computational Linguistics.

Kurt Shuster, Spencer Poff, Moya Chen, Douwe Kiela, and Jason Weston. 2021. Retrieval augmentation reduces hallucination in conversation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3784–3803, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008.

Jinpeng Wang, Yutai Hou, Jing Liu, Yunbo Cao, and Chin-Yew Lin. 2017. A statistical framework for product description generation. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 187–192, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Zhenyi Wang, Xiaoyang Wang, Bang An, Dong Yu, and Changyou Chen. 2020. Towards faithful neural table-to-text generation with content-matching constraints. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1072–1086, Online. Association for Computational Linguistics.

Zhongyu Wei, Yang Liu, and Yi Li. 2016. Is this post persuasive? ranking argumentative comments in online forum. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 195–200, Berlin, Germany. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi 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 Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Yiran Xing, Zai Shi, Zhao Meng, Gerhard Lakemeyer, Yunpu Ma, and Roger Wattenhofer. 2021. KM-BART: Knowledge enhanced multimodal BART for visual commonsense generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 525–535, Online. Association for Computational Linguistics.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter Liu. 2020. PEGASUS: Pre-training with extracted gap-sentences for abstractive summarization. In *ICML 2020: 37th International Conference on Machine Learning*, volume 1, pages 11328–11339.

Tao Zhang, Jin Zhang, Chengfu Huo, and Weijun Ren. 2019a. Automatic generation of pattern-controlled product description in e-commerce. In *The World Wide Web Conference*, pages 2355–2365.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019b. BERTScore: Evaluating text generation with BERT. In *International Conference on Learning Representations*.