Contrastive Multimodal Text Generation for E-Commerce Brand Advertising

Nikhil Madaan*
nmadaan@andrew.cmu.edu
Bloomberg

Krishna Reddy Kesari
kkesari@amazon.com
Amazon

Manisha Verma
mvr@amazon.com
Amazon

Shaunak Mishra
shaunakm@amazon.com
Amazon

Tor Steiner
tssteine@amazon.com
Amazon

ABSTRACT

E-commerce platforms enable brands to connect with relevant online shoppers. While major brands are easily identifiable by shoppers, smaller and emerging brands often lean on advertising campaigns in e-commerce platforms to reach a wide audience. For such advertising campaigns, brands need to come up with a leading ad creative which may be shown together with their listed products. Designing such creatives requires domain expertise in marketing; it is time-intensive as well as expensive for small businesses in particular. To assist brands with the leading ad text which goes together with the title and image of their listed products, we propose a multimodal text generation model. The multimodality stems from using both the textual and visual components of multiple product listings from a brand to generate the ad text. In addition, we introduce a brand-contrastive loss while training the multimodal text generation model. This is done to provide shoppers with an experience which is unique to a brand, while learning from data collected from multiple brands across product categories. Our experiments demonstrate the benefits of multimodal inputs for ad text generation; images are useful especially when textual information is limited. We also demonstrate how our brand contrastive loss enables unique brand advertising experiences at scale by promoting diversity in the generated ad text across brands.

KEYWORDS

ad text generation, transformers, contrastive loss, multimodal

ACM Reference Format:
Nikhil Madaan, Krishna Reddy Kesari, Manisha Verma, Shaunak Mishra, and Tor Steiner. 2023. Contrastive Multimodal Text Generation for E-Commerce Brand Advertising. In *Proceedings of International Workshop on Multimodal Learning, KDD '23 (Multimodal KDD '23).* Long Beach, CA, USA, 6 pages.

1 INTRODUCTION

At the beginning of their shopping journey, online shoppers may not be aware of all the brands relevant to their search. While established brands are more likely to come to a shopper's mind, shoppers may not know about smaller yet emerging brands which can fulfil their shopping mission. To address such awareness gaps, brands often leverage brand advertising campaigns in e-commerce platforms (*e.g.*, sponsored brands in Amazon[1]). Brands may already have their product titles and images in the e-commerce platform's catalog; this enables them to show up as organic results in the platform.

However, such product titles are typically not suitable as brand advertising text or the title may not be catchy enough to attract shopper attention. Due to this, brands have to rely on creative strategists to design enticing ad text which can go together with the brand's product images and titles in the ad campaign. This approach to create ad text is common in the ads industry [15] but it is an expensive and time taking process. This also does not scale well with the numbers of products a brand has, *e.g.*, a subset of the brand's products may be shown for a shopper query and it is desirable to have an ad text customized for the subset.

Recent progress in text generation have opened up the possibility of generating ad creatives at scale. Specifically for text, researchers

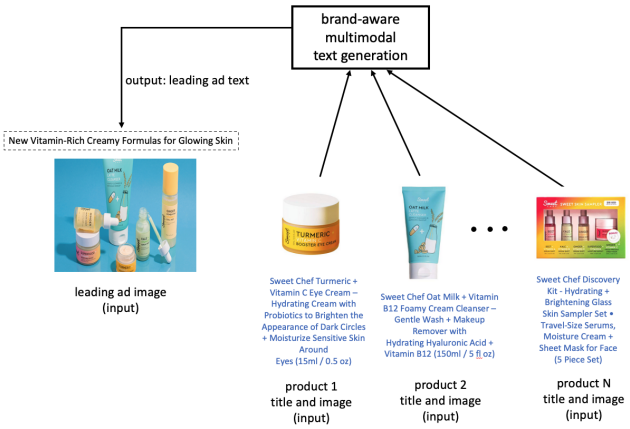


Figure 1: Illustrative example of multimodal ad text generation for a brand using multiple products (with their titles and images) from the brand as well as the leading ad image.

have proposed several kinds of models [8, 15, 21, 27, 28] that can ingest product data to generate ad text. However, existing work has mostly focused on ingesting textual inputs from a product (*e.g.*, title and description) in generating ad text and have not considered product images. It has been shown [4, 13] that images are instrumental in attracting attention, driving up user interaction and conversions. In brand advertising, images play a crucial role in establishing a brand's identity and appeal. Specifically, product images can be critical in distinguishing between two brands that sell the same type of product since the images can highlight product level differences such as make, material, color and brand logo. Visual information can also aid text generation where textual

*Work done during internship at Amazon
[1]https://advertising.amazon.com/solutions/products/sponsored-brands

data is limited, incomplete or extremely short; this is common for low-resource advertisers and locales that do not have resources to compile very thorough product attributes, descriptions or websites.

Apart from the possibility of using visual information for ad text generation, there is limited work on promoting diversity of generated ad text across brands. Given the large volume of advertisers and products, the onus is on the ad text generation model to ensure that brand-level characteristics are preserved and no two advertisers selling similar products get similar (sometimes identical) generated text as suggestions from the model.

In this paper, we build on the above gaps and possibilities (*i.e.*, using visual information for ad text generation, and promoting diversity of generated text across brands). Our proposed approach uses both product image and text (for multiple products from the brand), and uses a brand-contrastive loss to promote diverse generated ad text across brands. Specifically, via our proposed approach, we study the following research questions (RQ).

- **RQ1:** Do multimodal embeddings help generate better-quality ad text for brands?
- **RQ2:** Is our proposed model able to generate good quality ad text in the absence of sufficient textual information in the brand's products and in the cold-start setting, *e.g.*, where the brand's product and ad text are entirely missing from training data. The latter happens often as new brands onboard and advertise their products.
- **RQ3:** When multiple products are ingested for a brand's ad text generation, does generation quality stay stable? In other words, does the quality of results remain at par with the ones from using a single product for a brand's ad text generation?
- **RQ4:** Does brand-contrastive loss improve diversity of generated ad text across brands?

Through large scale experiments and several ablation studies presented in this paper for our proposed ad text generation approach, we demonstrate positive answers to the above questions. The remainder of this paper is organized as follows. We describe the proposed method in Section 3. The baselines and evaluation metrics are reported in Section 4. We elaborate on the empirical findings in section 4.3 and concluding with future work in Section 5.

2 RELATED WORK

Our work spans multiple areas of research such as advertising, text-generation and multi-modal learning summarized below.

Text Generation in advertising: Advertisers can use different platforms such as Amazon or JD.com [17, 21] to launch campaigns that show ads about different products to users. Advertisers can learn which creative elements work with their targeted audience with help of exploratory A/B tests on a large pool of creatives [12]. However, automatically understanding ad creatives (multi-modal in nature due to the presence of text and an image) and leveraging this understanding to generate creatives for new products is emerging as an active area of research. Previous work has explored generation of advertisements [6] using an advertiser's webpage. They calculate the importance of generated text with CTR data. While, we do not compute ad copy level CTR, it is also not straight forward to counterfactually estimate ad-component (text and image) level CTR from noisy user data which limits the use of this work. Another related work [8] presents a self-critical model for generating ad text,

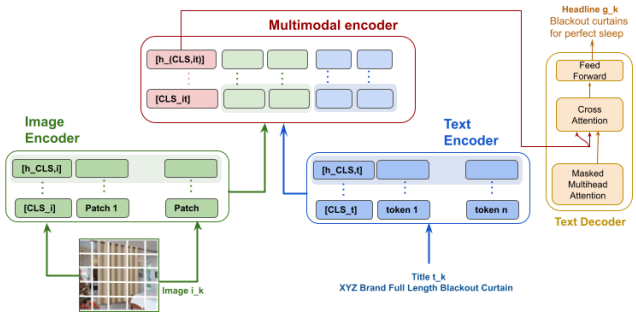


Figure 2: Brand contrastive creative text generation model

where the authors rely on two methods of generating ad text from BERT, however, their proposed architecture does not support integration of images directly which is our primary contribution. There is also work that aims to generate multi-product advertisements [27], however the authors do not use images to generate these ads. They also rely on product attributes extracted using rules which is not a scalable approach especially as new brands are constantly added to the inventory. Existing work [5] also explores product keywords and search terms to generate sponsored search advertisements. However, we use product attributes across brands to generate ad-text as search queries may not exist.

Multimodal Learning. There are several other works that use both images and text for other tasks such as visual question and answering (VQA) [22, 25], caption generation [3, 16] and keyword ranking [29] which are not aligned for ad text generation. Caption generation models only describe the detected objects in the image which isnt the goal of our work. A caption generation model, for instance would generate 'a table in a room' for a dining table advertised in a drawing room which is not as appealing as 'luxurious table for your drawing room' for the same product. Our problem is also different from VQA since the aim is to utilize the image in highlighting product features without list of predefined questions. Similarly, product image description/summarization [2, 9, 10, 23] models often describe product attributes such as color, texture or shape without any creative or marketing appeal. Advertisement text, however, is very different in terms of vocabulary and word use from product summaries or descriptions. It is neither verbose nor very descriptive. Given the nature of ad-text, we compare some of these works as baselines and show that merely training a multimodal models with advertisement text is not enough, and brand or category specific information is needed to achieve higher performance.

3 CONTRASTIVE MULTIMODAL CREATIVE TEXT GENERATION

We aim to leverage both product images and textual information to generate higher quality ad-text while preserving brand identity and encouraging diversity in generated text across brands that sell similar products. Formally, given a pair of image and text (i_k, t_k) for product a_k, the goal is to generate advertisement text (g_k). We extend the multimodal architecture in [22] with contrastive learning to generate text as shown in Figure 2. The multimodal

encoder consists of an image, text and fusion transformer that takes both product image (h_i) and product title embedding (h_t) as input to generate multimodal hidden states (h_{it}). Text encoder hidden states (h_t) are same size as that obtained from the image encoder (h_i). The fusion transformer generates multimodal hidden states (h_{it}) by linearly projecting image and text embeddings, allowing cross-attention between the image and text representations and fusing the two modalities as shown in Figure 2.

If the input advertisement contains multiple products, we use concatenated multimodal hidden states. For instance, for an advertisement with three products, multimodal hidden representations of all three product image, text pairs are concatenated $[h_{it1}, h_{it2}, h_{it3}]$, and used as input to the decoder. The decoder consists of transformer blocks, where multimodal hidden states of the encoder constitute keys (K) and values (V) and generated sequence's previous token is used as a query (Q) to compute the $crossattn(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d}})V$ and compute the probability distribution for the subsequent token. The decoder is trained using teacher forcing i.e. the objective is to maximize the likelihood of the next ground-truth token conditioned on the previous ground-truth tokens and multimodal context.

3.1 Loss Computation

We aim to generate relevant *but* diverse advertisement copy given an input product title and image. We rely on a linear combination of cross-entropy and contrastive loss to achieve that diversity. The cross entropy loss ensures that generated text does not deviate from word distribution observed at the time of training and contrastive loss ensures that embeddings of generated ad text for different brands are far apart. If model weights are represented as θ, the language modeling objective is to minimize cross-entropy loss.

$$\mathcal{L}_{CE}(\theta) = -\sum_{t=1}^{T} logP_\theta(y_t|y_{<t}, h_{(it)}) \tag{1}$$

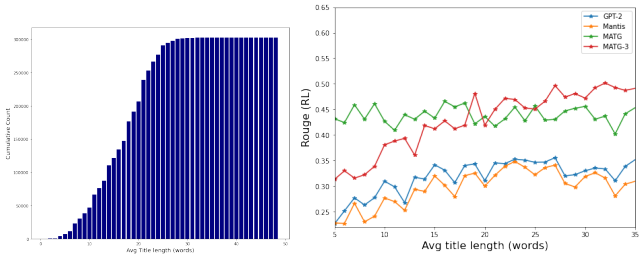
where $P_\theta(y_t|y_{<t}, h_{(it)})$ is the probability of observing token t given tokens in ground truth reference ad copy for the product and multimodal representation of product image and text.

We use contrastive loss, specifically pairwise margin loss, computed on pairs of generated ad text for different brands.

$$\mathcal{L}(y^+, y^-) = max(0, cos(h_{g^+}, h_{y^-}) - cos(h_{g^+}, h_{y^+}) + 1),$$
$$\mathcal{L}_{CL} = \sum_{(y^+,y^-)\in P} \mathcal{L}(y^+, y^-) \tag{2}$$

Given a pair of reference ad text from two different brands $((y^+, y^-))$ from a set of randomly sampled $P \in C_k^2$ pairs, we rely on the cosine-similarity between hidden states of generated text (determined by greedy search) and the reference ad text. Contrastive loss above is only computed when the entire sequence is generated. This ensures that entire generated sentence is used to compare against the reference ad text. The hidden representation h_{g^+} of the last token in generated sequence is used to compare with the last token representation of positive h_{y^+} and negative reference h_{y^-} ad copy respectively. Finally we linearly combine the cross-entropy and contrastive loss as follows.

$$\mathcal{L} = \lambda\mathcal{L}_{CE} + (1 - \lambda)\mathcal{L}_{CL} \tag{3}$$



(a) Title Length **(b) Model Performance**

Figure 3: Title length and model performance

It is worth noting that the pairwise margin loss can be replaced with noise contrastive estimation loss too, only greedy decoding of several sequences becomes prohibitively expensive.

4 EXPERIMENTS

We review the dataset, baselines and the evaluation metrics in following sub-sections.

4.1 Dataset

In this work we leverage an internal dataset of an e-commerce platform, consisting of 335K advertisements from 42K brands. Each advertisement can contain upto three products. For our experiments, we use either one or all three product images and titles for ad copy generation. The average product title 22.2 (± 7.9) words long where as the advertisement copy is 6.5 (± 1.5) words. As shown in Figure 3a, 20% of product titles are shorter than 10 words which results in text-only ad generation models to perform poorly in practice for such products. The contrastive model is trained on 600K ad-text pairs sampled randomly across different brands from the same category. We sample 300K advertisements from brands such that there is no overlap between cold-start train, validation and test sets.

4.2 Baseline approaches

We consider several baseline methods (listed below) to compare with the performance of our proposed model.

Text Only models: We train several text-only models that take into account only product title to generate ad text. We specifically compare our work with previous methods such as SCMLM [8], COBART [7], GPT2 [1] and BART [11] and T5 [20] respectively finetuned on the data described in Section 4.1. More details about each model, training hyperparameters are Section A.1.

Image only models: Given that our work combines image and textual data, we also compare against image only baselines, where we feed only the product images to the model. In this work, we compare against CLIPCap [16] finetuned on product images as input and advertisement text as the output of the caption generation model. The aim of the experiments is to evaluate whether such models can be finetuned to generate creative text.

Image and text models: As discussed previously, existing multimodal literature is for tasks such as caption generation or visual Q&A but not for creative text generation such as advertisement text. Thus, we compare our work with ManTis [23] which generates

Model	BLEU	R1	R2	RL	BertS	SBLEU
BART[11]	-23.61	-7.98	-9.64	-9.09	-3.41	-4.19
T5[20]	-13.89	-1.84	-5.42	-9.74	-2.27	-3.52
ClipCap[16]	-71.53	-54.29	-69.28	-53.57	-2.27	-164.32
ManTis[23]	-7.64	-4.29	-6.63	-4.87	-0.68	20.70
GPT-2[1]	-	-	-	-	-	-
SCMLM[8]	52.78	21.60	53.61	23.70	1.59	29.52
COBART[7]	80.56	29.14	63.25	32.14	1.14	36.12
MATG ($\lambda = 1$)	76.39	34.66	69.28	37.34	2.73	40.53
MATG	84.72	38.04	75.90	**40.91**	2.95	44.93
MATG-3 ($\lambda = 1$)	88.89	36.50	78.31	39.61	2.95	40.31
MATG-3	**90.28**	**38.96**	**86.75**	40.26	**3.41**	**41.63**
Cold Start evaluation						
Model	BLEU	R1	R2	RL	BScore	SBLEU
ManTis	-18.75	-3.37	-1.47	-3.26	0.24	-11.65
COBART	-	-	-	-	-	-
MATG-3	**57.81**	**38.55**	**89.71**	**39.13**	**2.24**	**21.12**
Short Title evaluation (len < 10 words)						
Model	BLEU	R1	R2	RL	BScore	SBLEU
ManTis	-8.20	-3.90	-8.16	-6.21	-0.57	21.43
GPT2	-	-	-	-	-	-
MATG	**95.90**	**36.69**	**80.95**	**40.00**	**2.85**	**42.86**
MATG-3	90.98	30.52	75.51	34.14	2.62	37.62

Table 1: Percentage improvements in BLEU, ROUGE (R1, R2, RL), BertScore (BertS) and SBLEU over baselines (first row in each subtable) on warm-start, brand cold-start and short titles

model	unigram	bigram	embed
ground-truth	**0.98 +/- 0.05**	**0.99 +/- 0.03**	**0.74 +/- 0.16**
COBART	0.96 +/- 0.07	0.98 +/- 0.06	0.68 +/- 0.12
ManTis	0.94 +/- 0.09	0.98 +/- 0.05	0.65 +/- 0.15
MATG-3	**0.97 +/- 0.07**	**0.99 +/- 0.04**	**0.70 +/- 0.18**

Table 2: Token level and embedding diversity across brands for 850 categories in test set and model generated outputs

product descriptions. Since no open-source checkpoints were available, we implemented ManTis [23] and trained it on our dataset. Our contrastive model is trained on one product (MATG) and all three product titles (MATG-3) to generate ad text. We initialize the text, image and multimodal encoder weights using the checkpoint released for FLAVA [22] which has been trained on several multimodal representation learning tasks. GPT2 [1] has been used to initialize the text decoder weights.

4.3 Results and Discussion

We evaluate the quality of generated ad text across different methods using three metrics: Rouge (R1, R2 and RL) [14], BLEU [18] and BertScore (BScore) [26]. As shown in Table 1, the text-only baselines (GPT-2, BART, T5) lag behind on all evaluation metrics compared to models that utilize both images and text (MATG, MATG-3). Image only models (ClipCap, ManTis) are the worst performing models, indicating that product title is important in generating better ad text. Simple concatenation of image and text embeddings as done in ManTis [23], is also not sufficient to achieve good performance, and even falls behind text-only models such as GPT-2, BART. This addresses RQ1 indicating that image representations combined with text can aid creative text generation.

We also note that using multiple products (both images and text) in the advertisements leads to better ad text. Given that product data is diverse, their titles can also vary significantly in length and density of information. We posit that images will be useful when titles are short and do not contain enough product information to generate good quality ad text. Thus, first we evaluate whether our model is able to generate good quality ad text when product tiles are short. Given that 20% of titles contain less than 10 tokens (as shown in Figure 3a), we compare the distribution of ROUGE scores with respect to title length in Figure 3b and Table 1. While multimodal fusion models MATG and MATG-3 significantly outperform all models for different title lengths, the gains are higher especially when the titles are short. However, MATG has better performance than MATG-3, because generating a creative advertisement copy for multiple products is more challenging with short titles.

Given that new brands regularly advertise products on Amazon, it is imperative to evaluate the generation quality for unseen brands i.e. cold-start brand evaluation. To answer our second research question (RQ2). We evaluated the best performing model (MATG-3 (CS)) in Table 1 for brand cold-start and found the performance is lower than when brand information is absent during training. This shows that cold start is a harder problem to solve [24] and more effort is required to improve model performance on out-of-distribution data.

We observe that adding more products as input does not hurt the quality of generated ad text. MATG-3 performs comparably (RL improvement of 40.26% over text-only GPT-2) to single product model MATG (RL improvement of 40.91%) as shown in Table 1. To answer our third research question (RQ3), we find that existing baselines do not generate higher quality ad text when multiple products are given as inputs. Some examples of generated ad text along with product information and advertiser submitted text are shown in Table 4. It is interesting to note that the multimodal model can capture nuances in the image that are absent from the text to formulate a good ad text. For instance, in case of product about Confetti, our model is able to determine that it can be used for parties based on the product image. Similarly, Table 3 contains ad text generated using multiple products. Overall, our findings suggest that product images, along with product information can improve creative text generation. The performance of our models on advertisements with multiple products is on par with those with single product, thus showing that our contrastive multimodal model is capable of fusing product information for a brand. We also observed improved performance for products where textual information was limited i.e. input titles were short supporting our hypothesis that images can be instrumental in creative text generation.

We address our fourth research question (RQ4) by computing several diversity metrics. We report token-level diversity and embedding level diversity across advertiser submitted and model generated ad text for test-set in Table 2. We define token-level diversity (unigram and bigram) as $token - diversity(h_{ik}, h_{jk}) = 1 - (\frac{t(h_{ik}) \cap t(h_{ik})}{t(h_{ik}) \cup t(h_{ik})})$ where h_{ik} and h_{ik} are randomly sampled ad text from different advertisers i and j in category k respectively. $t(\cdot)$

	Product 1		Product 2		Product 3	Reference	Generated
	Sam Leather Crossbody Bag		Rowan Bucket Bag in Teal		Cassie Convertible Crossbody Bag	Shop Anabaglish Handmade Quality Leather Bags	Sturdy & Stylish Crossbody Phone Bag
	80pcs Waterproof Nature Dinosaur Stickers for Laptop Water Bottle Scrapbook Sticker Pack		80pcs Waterproof Neon Light Vinyl Stickers for Laptop Water Bottle		80pcs Waterproof Easter Stickers for Water Bottle Envelopes Cards	Best Stickers for Kids Teens	Cute Waterproof Stickers for Kids and Teens
	PURPLE LEAF Outdoor Dining Chair Coffee		PURPLE LEAF Outdoor Dining Chair Grey		PURPLE LEAF Outdoor Dining Chair Dark Blue	Elaborate chairs through complete handiwork	Exquisite dining chairs for everyone
	Quiver Time 80+ Deck Blocks with 2 Dividers - Set of 5 Boxes - White, Black & Green		Quiver Time Red Portable Game Card Carrying Case		Black Bolt Quiver Card Case for Carrying Trading Card Games Like Pokemon	Searching for a Better Deck Box?	Looking for a Better Way to Carry Your Cards?

Table 3: Generated ad text with multiple input products

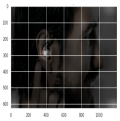
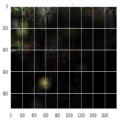
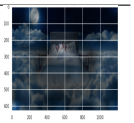
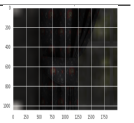
Product Image	Product Title	Reference	Generated	Attention Map
	Shure AONIC 215 True Wireless Sound Isolating Earbuds, Premium Audio Sound with Deep Bass, Bluetooth 5, Secure Fit Over-the-Ear	Decades of Experience Supporting Music Legends	The Perfect Earbuds For Your Life	
	Ultimate Confetti Bright Multicolor Biodegradable Tissue Confetti Circles- 1" 30,000 Pieces (1lb) - Confetti Balloons	The Ultimate Confetti Superstore!	Premium Quality Sturdy Party Decorations	
	Wake-Up Single Size Pocket Queen Mattress (72x66x10-inch)	Shop For Single 72x66 Size Mattress	The Most Comfortable Mattress for you	
	Casableu Polyester Blackout Printed Set of 2 Curtains - Silo Orange (Door 7 Feet)	Home that reflects you.	Elegant Blackout Curtains for Bedroom	

Table 4: Generated ad text for single product

is unigram and bigram generators for ad text. We ignore common stop words[2] for computing these metrics. Embedding based diversity is calculated using $embedding - diversity(h_{ik}, h_{jk}) = 1 - \cos(f(h_{ik}), f(h_{jk}))$ where $f(\cdot)$ is sentence embeddings[3] based cosine distance. We compute the three metrics for all pairs of ad text across advertisers in a category and report the mean and standard deviation in Table 2. While advertisers have a relatively high diversity across categories, our model achieves good performance.

We also use Self-BLEU (SBLEU) [30] (lower the better) to measure the diversity of generated ad text across brands for all the baselines in Table 1. MATG-3 achieves highest diversity across all models.

[2]https://gist.github.com/sebleier/554280#file-nltk-s-list-of-english-stopwords
[3]https://www.sbert.net/

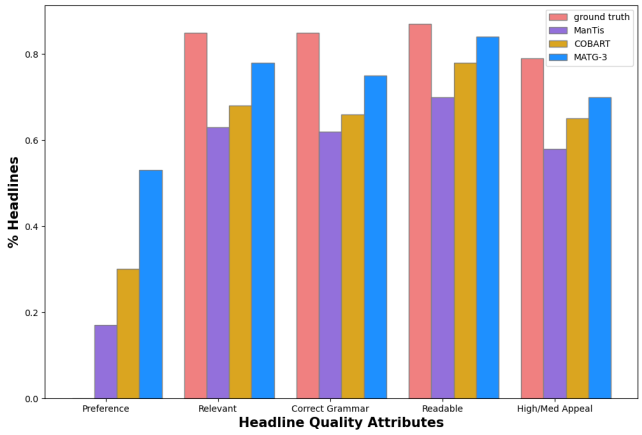


Figure 4: Manual Evaluation of generated headines across different quality dimensions

4.4 Manual Evaluation

Given the creative and non-deterministic nature of the task, we also performed manual evaluation wherein annotators were asked to annotate the quality of generated text along several dimensions.

We first asked annotators to annotate each generated image for four aspects, relevance, grammatical correctness, readability and overall advertising appeal. We sampled 500 suggestions from the test set to send annotators. Each suggestion was annotated by 3 annotators and overall agreement rate, measured using Fleiss kappa is 0.43 which indicates overall good agreement.

The aggregated results are shown in Figure 4. Annotators tend to prefer contrastive model MATG-3 output over other suggestions. While most model outputs get relatively good scores on grammar and readability, our model's generations get higher score for relevance and advertising appeal. One interesting thing to note is that advertiser submitted ad-text does not get perfect scores, indicating that our training data itself may have some relevance issues. It also

indicates that we could use some pre-processing filters to improve the quality of the training data.

5 CONCLUSION

With this work, we aim to incorporate visual cues to improve ad-text generation. Existing work uses only textual information to generate advertisements which is limiting given the visual nature of advertising. To this effect, we leverage multimodal product information for creative advertisement text generation. We propose a contrastive multimodal network that exploits both product image and title to generate a creative advertisement headline. We demonstrate that images aid in generating better ad text, especially in cases where product titles are short or highly similar. There is room for improvement however, by using image-level modeling for different brands by incorporating other auxilary generation tasks.

REFERENCES

[1] 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.

[2] 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*. ACM. https://doi.org/10.1145/3292500.3330725

[3] Jaemin Cho, Seunghyun Yoon, Ajinkya Kale, Franck Dernoncourt, Trung Bui, and Mohit Bansal. 2022. Fine-grained Image Captioning with CLIP Reward. https://doi.org/10.48550/ARXIV.2205.13115

[4] Yung Kyun Choi, Sukki Yoon, Kacy Kim, and Yeonshin Kim. 2019. Text versus pictures in advertising: effects of psychological distance and product type. *International Journal of Advertising* 38, 4 (2019), 528–543. https://doi.org/10.1080/02650487.2019.1607649 arXiv:https://doi.org/10.1080/02650487.2019.1607649

[5] Siyu Duan, Wei Li, Cai Jing, Yancheng He, Yunfang Wu, and Xu Sun. 2020. Query-Variant Advertisement Text Generation with Association Knowledge. https://doi.org/10.48550/ARXIV.2004.06438

[6] J Weston Hughes, Kenghao Chang, and Ruofei Zhang. 2019. Generating Better Search Engine Text Advertisements with Deep Reinforcement Learning. KDD.

[7] Yashal Shakti Kanungo, Gyanendra Das, A Pooja, and Sumit Negi. 2022. COBART: Controlled, optimized, bidirectional and auto-regressive transformer for ad headline generation. (2022).

[8] 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

[9] Shashank Kedia, Aditya Mantha, Sneha Gupta, Stephen Guo, and Kannan Achan. 2021. Generating Rich Product Descriptions for Conversational E-commerce Systems. In *Companion Proceedings of the Web Conference 2021*. ACM. https://doi.org/10.1145/3442442.3451893

[10] Fajri Koto, Jey Han Lau, and Timothy Baldwin. 2022. Can Pretrained Language Models Generate Persuasive, Faithful, and Informative Ad Text for Product Descriptions?. In *Proceedings of The Fifth Workshop on e-Commerce and NLP (ECNLP 5)*. Association for Computational Linguistics, Dublin, Ireland, 234–243. https://doi.org/10.18653/v1/2022.ecnlp-1.27

[11] 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 preprint arXiv:1910.13461* (2019).

[12] Wei Li, Xuerui Wang, Ruofei Zhang, Ying Cui, Jianchang Mao, and Rong Jin. 2010. Exploitation and Exploration in a Performance Based Contextual Advertising System. In *KDD 2010*.

[13] Yiyi Li and Ying Xie. 2020. Is a Picture Worth a Thousand Words? An Empirical Study of Image Content and Social Media Engagement. *Journal of Marketing Research* 57, 1 (2020), 1–19. https://doi.org/10.1177/0022243719881113 arXiv:https://doi.org/10.1177/0022243719881113

[14] Chin-Yew Lin. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out*. ACL.

[15] Shaunak Mishra, Manisha Verma, Yichao Zhou, Kapil Thadani, and Wei Wang. 2020. Learning to create better ads: Generation and ranking approaches for ad creative refinement. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*. 2653–2660.

[16] Ron Mokady, Amir Hertz, and Amit H. Bermano. 2021. ClipCap: CLIP Prefix for Image Captioning. https://doi.org/10.48550/ARXIV.2111.09734

[17] Lara O'Reilly and Laura Stevens. 2018. Amazon, With Little Fanfare, Emerges as an Advertising Giant. *The Wall Street Journal* (2018).

[18] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: A Method for Automatic Evaluation of Machine Translation. ACL.

[19] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning Transferable Visual Models From Natural Language Supervision. https://doi.org/10.48550/ARXIV.2103.00020

[20] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2019. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer.

[21] Huajie Shao, Jun Wang, Haohong Lin, Xuezhou Zhang, Aston Zhang, Heng Ji, and Tarek Abdelzaher. 2021. Controllable and Diverse Text Generation in E-Commerce. In *Proceedings of the Web Conference 2021* (Ljubljana, Slovenia) *(WWW '21)*. Association for Computing Machinery, New York, NY, USA, 2392–2401. https://doi.org/10.1145/3442381.3449838

[22] Amanpreet Singh, Ronghang Hu, Vedanuj Goswami, Guillaume Couairon, Wojciech Galuba, Marcus Rohrbach, and Douwe Kiela. 2021. FLAVA: A Foundational Language And Vision Alignment Model. https://doi.org/10.48550/ARXIV.2112.04482

[23] Michael Sollami and Aashish Jain. 2021. Multimodal Conditionality for Natural Language Generation. https://doi.org/10.48550/ARXIV.2109.01229

[24] Maksims Volkovs, Guangwei Yu, and Tomi Poutanen. 2017. Dropoutnet: Addressing cold start in recommender systems. *Advances in neural information processing systems* 30 (2017).

[25] Ziyi Yang, Yuwei Fang, Chenguang Zhu, Reid Pryzant, Dongdong Chen, Yu Shi, Yichong Xu, Yao Qian, Mei Gao, Yi-Ling Chen, Liyang Lu, Yujia Xie, Robert Gmyr, Noel Codella, Naoyuki Kanda, Bin Xiao, Lu Yuan, Takuya Yoshioka, Michael Zeng, and Xuedong Huang. 2022. i-Code: An Integrative and Composable Multimodal Learning Framework. https://doi.org/10.48550/ARXIV.2205.01818

[26] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2019. BERTScore: Evaluating Text Generation with BERT.

[27] Xueying Zhang, Kai Shen, Chi Zhang, Xiaochuan Fan, Yun Xiao, Zhen He, Bo Long, and Lingfei Wu. 2022. Scenario-based Multi-product Advertising Copywriting Generation for E-Commerce. https://doi.org/10.48550/ARXIV.2205.10530

[28] Xueying Zhang, Yanyan Zou, Hainan Zhang, Jing Zhou, Shiliang Diao, Jiajia Chen, Zhuoye Ding, Zhen He, Xueqi He, Yun Xiao, Bo Long, Han Yu, and Lingfei Wu. 2022. Automatic Product Copywriting for E-commerce. *Proceedings of the AAAI Conference on Artificial Intelligence* 36, 11 (Jun. 2022), 12423–12431. https://doi.org/10.1609/aaai.v36i11.21508

[29] Yichao Zhou, Shaunak Mishra, Manisha Verma, Narayan Bhamidipati, and Wei Wang. 2020. Recommending Themes for Ad Creative Design via Visual-Linguistic Representations. In *Proceedings of The Web Conference 2020* (Taipei, Taiwan) *(WWW '20)*. Association for Computing Machinery, New York, NY, USA, 2521–2527. https://doi.org/10.1145/3366423.3380001

[30] Yaoming Zhu, Sidi Lu, Lei Zheng, Jiaxian Guo, Weinan Zhang, Jun Wang, and Yong Yu. 2018. Texygen: A Benchmarking Platform for Text Generation Models. https://doi.org/10.48550/ARXIV.1802.01886

A SUPPLEMENTARY MATERIAL

A.1 Hyperparameter settings

Multimodal models MATG and MATG-3, have three transformer blocks. The image, text and multimodal encoder has 12, 12 and 6 hidden layers respectively with 768 dimensions. The text decoder has 12 hidden layers and hidden states of 768 dimension. ManTis uses a CLIP [19] checkpoint with 12 hidden layers in the text and image encoder with 768 dimensions. For all the headline generation models, we use the AdamW optimizer ($lr = 5e-6, weight\_decay = 0.01, betas = (0.9, 0.999)$) and ReduceLROnPlateau schedular ($patience = 2, factor = 0.5, min_lr = 1e-8$). We train all the models using 8*Tesla V100 GPUs. We use a batch size of 64 for ClipCap, GPT-2, BART, ManTis, MATG, and 32 for MATG-3. We train the models until validation loss increases or saturates.