ABSTRACT

Existing explanation models generate only text for recommendations but still struggle to produce diverse contents. In this paper, to further enrich explanations, we propose a new task named personalized showcases, in which we provide both textual and visual information to explain our recommendations. Specifically, we first select a personalized image set that is the most relevant to a user’s interest toward a recommended item. Then, natural language explanations are generated accordingly given our selected images. For this new task, we collect a large-scale dataset from Google Maps and construct a high-quality subset for generating multi-modal explanations. We propose a personalized multi-modal framework which can generate diverse and visually-aligned explanations via contrastive learning. Experiments show that our framework benefits from different modalities as inputs, and is able to produce more diverse and expressive explanations compared to previous methods on a variety of evaluation metrics.  

CSCS CONCEPTS

• Computing methodologies → Natural language generation.

KEYWORDS

Datasets, Text Generation, Multi-Modality, Contrastive Learning

1 INTRODUCTION

Personalized explanation generation models have the potential to increase the transparency and reliability of recommendations. Previous works considered generating textual explanations from users’ historical reviews [26], tips [12] or justifications [13]. However, these methods still struggle to provide diverse explanations since a large amount of general sentences (e.g., ‘food is good!’) exist in outputs and models lack grounding information (e.g., images) to guide generation. To further diversify and enrich explanations for recommendations, we propose a new task named personalized showcases (shown in Figure 1) to explain recommendations via both textual and visual information.

To this end, the first challenge is to build a dataset with multi-modal personalized information. We construct a large-scale dataset, namely Gest, from Google Local Restaurants including review text and corresponding pictures. Then, to improve the quality of Gest, we annotate a small subset to find highly matched image-sentence pairs, and train a CLIP-based classifier to extract visually-aware explanations from the full dataset.

For the new task, we design a multi-modal explanation generation framework. We first select images from historical photos of the business that the user is most interested in. Then, we take the displayed images and users’ historical reviews as inputs and learns to generate textual explanations with a multi-modal decoder. But generating expressive and engaging text to capture users’ interest...
remains a challenging problem. First, different from previous tasks, the alignment between multiple images and generated text poses higher requirements for information extraction and multi-modal learning. Second, a typical encoder-decoder model with a cross-entropy loss can easily lead to repetitive and dull sentences that occur frequently in the training corpus [8] (e.g., “food is great”).

To tackle these challenges, we propose a Personalized Cross-Modal Contrastive Learning (PCCL) framework by contrasting input modalities with output sequences. We first design a cross-modal loss to enforce the alignment between images and output explanations, by constructing hard negative samples with randomly replaced entities in the output. Motivated by the observation that users with similar historical reviews share similar interests, we further design a personalized loss to reweight the negative samples based on their history similarities. Experiments on both automatic and human evaluation show that our model is able to generate more expressive, diverse and visually-aligned explanations compared to a variety of baselines. Overall, our contributions are:

- To generate more informative explanations for recommendations, we present a new task: personalized showcases which can provide both textual and visual explanations.

- For this new task, we collect a large-scale dataset from Google Local (i.e., maps), and extracted high-quality samples with pre-processing and filtering.

- We propose a novel multi-modal framework for personalized showcases which applies contrastive learning to improve diversity and visual alignment of generated text.

2 TASK DEFINITION

In the personalized showcases task, given user $u$, business $b$ and the image candidate set $I_b = \{i^b_1, i^b_2, \ldots, i^b_K\}$ from $b$, we first select a set of images as visual explanations $I$ from $I_b$ which user $u$ will be interested in, based on $u$’s profile (i.e., historical reviews $X_u = \{x^u_1, x^u_2, \ldots, x^u_K\}$) and images $I_u = \{i^u_1, i^u_2, \ldots, i^u_K\}$. Then, we use the historical reviews $X_u$ and selected images $I_u$ to generate textual explanations $S$. $S_u$ and $I_u$ are personalized to explain why $b$ is recommended to $u$.

To better study each modality and provide baselines for future work, in this paper, we decompose the task into two steps as shown in Figure 2: (1) Selecting an image set as visual explanations related to a user’s interest; (2) Generating textual explanations given selected images and user’s historical reviews.

We consider following aspects to evaluate the task: (1) Accuracy: We aim to predict the target images accurately, and the generated text is expected to be relevant to the business. (2) Diversity: Both visual and textual explanations should be diverse and expressive. (3) Alignment: In our visually-aware setting, the model should be able to accurately describe the content in the given images.

3 METHODOLOGY

3.1 Personalized Image Set Selection

The first step of our framework is to select an image set as a visual explanation that is relevant to a user’s interests, and is diverse. We formulate this selection step as diverse recommendation with multi-modal inputs.

Multi-Modal Encoder. We use CLIP [16], a state-of-the-art pre-trained cross-modal retrieval model as both textual- and visual-encoders. CLIP encodes raw images as image features, and encodes user textual- and visual-profiles as user profile features.

Image Selection Model. We use a Determinantal Point Process (DPP) [9] to select the image subset, which has recently been used for different diverse recommendation tasks [1, 21]. Compared with other algorithms for individual item recommendation, DPP-based models are suitable for multiple image selection. Given user $u$ and business $b$, we predict the image set $I_{u,b}$ as follows:

$$\hat{I}_{u,b} = \text{DPP}(I_b, u),$$

where $I_b$ is the image set belonging to business $b$. In our design, we calculate user-image relevance using the CLIP-based user’s profile features and image features. More details of the model are in [21].

3.2 Visually-Aware Explanation Generation

After obtaining an image set, we aim to generate personalized explanations given a set of images and a user’s historical reviews. Specifically, we build a multi-modal encoder-decoder model with GPT-2 [18] as the backbone.

Multi-Modal Encoder. Given a set of user $u$’s historical reviews $X = \{x_1, x_2, \ldots, x_K\}$, we use the text encoder of CLIP to extract the review features $R = \{r_1, r_2, \ldots, r_K\}$. Similar operations are applied to the input images $I = \{i_1, i_2, \ldots, i_L\}$, where we use the CLIP visual encoder to extract visual features $V = \{v_1, v_2, \ldots, v_N\}$. Those features are then projected into a latent space:

$$Z^V_i = W^V v_i, Z^R_i = W^R r_i,$$

where $W^V$ and $W^R$ are two learnable projection matrices. Then we use a multi-modal attention (MMA) module with stacked self-attention layers [19] to encode the input features:

$$[H^V; H^R] = \text{MMA}(Z^V; Z^R),$$

where each $H^V$, $H^R$ aggregate features from two modalities and $|$ denotes concatenation. This flexible design allows for variable lengths of each modality and enables interactions between modalities via co-attentions.

Multi-Modal Decoder. Inspired by recent advances of pre-trained language models, we leverage GPT-2 as the decoder for generating explanations. To efficiently adapt the linguistic knowledge from GPT-2, we insert the encoder-decoder attention module into the pre-trained model with a similar architecture in [2]. With this multi-modal GPT-2, given a target explanation $Y = \{y_1, y_2, \ldots, y_N\}$, the decoding process at each time step $i$ can be formalized as:

$$\hat{y}_i = \text{Decoder}(H^V; H^R), y_1, \ldots, y_{i-1}.$$

We use a cross-entropy (CE) loss to maximize the conditional log likelihood $\log p_\theta(Y|X, I)$ for $N$ training samples $(X^{(i)}, I^{(i)}, Y^{(i)})_{i=1}^N$ as follows:

$$\mathcal{L}_{CE} = - \sum_{i=1}^N \log p_\theta(Y^{(i)}|X^{(i)}, I^{(i)}).$$

We use ground truth images from the user for training and images from our image-selection model for inference.

\[^{2252}\text{We omit the subscript } u \text{ below for simplicity}\]
3.3 Personalized Cross-Modal Contrastive Learning

Unlike image captioning tasks [22, 25] which mainly describe images, our task use multiple images as “visual prompts” to express personal feelings. To encourage expressive and visual-aligned generations, we propose Personalized Cross-Modal Contrastive Learning (PCCL). We first project the embeddings of images \( H^V \), historical reviews \( H^R \), and the target sequence \( H^Y \) into a latent space:

\[
\begin{align*}
\tilde{H}^V &= \phi_V(H^V),
\tilde{H}^R &= \phi_R(H^R),
\tilde{H}^Y &= \phi_Y(H^Y)
\end{align*}
\]  

(6)

where \( \phi_V, \phi_R, \) and \( \phi_Y \) consist of two fully connected layers with ReLU activation and average pooling over the hidden states \( H^V, H^R \) and \( H^Y \) from the last self-attention layers. With the InfoNCE loss [3, 14], we then maximize the similarity between the pair of source modality and target sequence, while minimizing the similarity between the negative pairs as follows:

\[
\mathcal{L}_{CL} = - \sum_{i=1}^{N} \log \frac{\exp(s_{i,Y})}{\exp(s_{i,Y}) + \sum_{j \neq k} \exp(s_{i,Y})}.
\]

(7)

where \( s_{i,Y} = \text{sim}(\tilde{H}^V(i), \tilde{H}^Y(j))/\tau \), \( \text{sim} \) is the cosine similarity between two vectors, \( \tau \) is the temperature parameter, \( (i) \) and \( (j) \) are two samples in the mini-batch, \( K \) is the set of negative samples for sample \( (i) \).

One challenge of this task is the model is asked to describe multiple objects in a set of images [20]. To ensure the visual grounding between multiple image features and output text, we design a novel cross-modal contrastive loss. Specifically, given a target explanation \( Y = \{y_1, y_2, ..., y_L\} \), we randomly replace the entities \(^3\) in the text with other entities presented in the dataset to construct a hard-negative sample \( Y_{ent} = \{y_{ent1}, y_2, ..., y_{entL} \} \) (i.e., “I like the sushi” to “I like the burger”), such that during training, the model is exposed to samples with incorrect entities regarding the images, which are non-trivial to distinguish from the original target sequence. Thus, we add the hidden representation of \( Y_{ent} \) as an additional negative sample \( ent \) to formulate the cross-modal contrastive loss:

\[
\mathcal{L}_{CCL} = - \sum_{i=1}^{N} \log \frac{\exp(s_{i,Y})}{\exp(s_{i,Y}) + \sum_{j \in K \cup \text{ent}} \exp(s_{i,Y})},
\]

(8)

\(^3\)We extract entities using spaCy noun chunks (https://spacy.io/).

On the other hand, to enhance the personalization [23] of explanations, we re-weight negative pairs according to user personalities. The intuition is that users with more distinct personalities are more likely to generate different explanations. Motivated by this, we propose a weighted personalized contrastive loss:

\[
\mathcal{L}_{PCL} = - \sum_{i=1}^{N} \log \frac{\exp(s_{i,Y}^{R})}{\exp(s_{i,Y}^{R}) + \sum_{j \neq k} \exp(s_{i,Y}^{R})}.
\]

(9)

where negative pairs in a mini-batch are re-weighted based on user personality similarity function \( f \). In our framework, user personalities are represented by their historical reviews. Specifically, we define \( f \) function as:

\[
f(i, j) = \alpha(1 - \text{sim}(\tilde{R}(i), \tilde{R}(j)))
\]

(10)

i.e., we reduce the weights of negative pairs with similar histories, and increase those with distinct histories. \( \alpha (\alpha > 1) \) is a hyperparameter that weighs the negative samples, \( \text{sim} \) is the cosine similarity, \( \tilde{R}(i) \) and \( \tilde{R}(j) \) are the average features of two users’ input historical reviews.

Overall, the model is optimized with a mixture of a cross-entropy loss and the two contrastive losses:

\[
\mathcal{L}_{loss} = \mathcal{L}_{CE} + \lambda_1 \mathcal{L}_{CCL} + \lambda_2 \mathcal{L}_{PCL}.
\]

(11)

where \( \lambda_1 \) and \( \lambda_2 \) are hyperparameters that weigh the two losses.

4 EXPERIMENTS

4.1 Experimental Setting

Baselines. We compare our model with popular baselines from different tasks, including image captioning, report generation and explanation generation:

- \( ST \) [22] is a classic CNN+LSTM model for image captioning.
- \( R2Gen \) [4, 24] is a state-of-the-art memory-driven transformer specialized at generating long text with visual inputs.
- \( Ref2Seq \) [13] is a popular reference-based seq2seq model.
- \( Peter \) [11] is a recent transformer-based model which uses the user and item IDs to generate explanation.

\( \text{img} \) and \( \text{text} \) refer to image and text features respectively.

Evaluation Metrics. For image selection, we report Precision@K, Recall@K and F1@K to measure the ranking quality. To evaluate
with much worse ranking accuracy, thus indicating the importance with ViT-B/32 and GPT-2 small as our backbone models.

we report

Dinstinct

Text Generation.

Results for text explanation generation are pre-

Peter

and

Ref2Seq

rics compared to other models. Though text-based models

Second,

geneneration is able to generate accurate and diverse explanations.

fits of incorporating visual features. The setting of visually-aware

and image-input models on diversity metrics validates the bene-

sented in Table 1. First, the clear gap between text-input models

to incorporate personalization into this task.

hand, a random selection model can achieve higher diversity but

can improve the performance over a single-modal model in terms

of ranking performance (e.g., precision and recall). On the other

other models.

Table 1: Performance comparison. Results are reported in percentage (%). GT is the ground truth.

| Model   | Input          | N-Gram Metrics         | Diversity Metrics | Embedding Metrics |
|---------|----------------|------------------------|-------------------|------------------|
|         |                | BLEU-1  | BLEU-4  | METEOR | NIST | DISTINCT-1 | DISTINCT-2 | CLIP-Score | BERT-Score |
| GT      |                | -       | -       | -      | -    | 6.06      | 43.23      | 28.41      | -            |
| ST      | img            | 8.24    | 0.28    | 3.41   | 28.08| 2.74      | 17.41      | 24.31      | 85.20        |
| R2Gen   | img            | 6.47    | 0.22    | 3.10   | 36.55| 3.23      | 22.45      | 24.28      | 85.89        |
| Ref2Seq | text           | 7.09    | 0.67    | 3.80   | 30.78| 0.92      | 5.89       | 23.83      | 84.71        |
| Peter   | text           | 8.89    | 0.44    | 3.28   | 34.45| 0.38      | 1.27       | 23.27      | 86.94        |
| Ours    | img            | 9.92    | 0.32    | 3.64   | 37.35| 3.37      | 26.37      | 24.68      | 88.03        |
|         | img+text       | 10.40   | 0.36    | 3.83   | 50.64| 3.58      | 28.58      | 24.50      | 88.23        |

Table 2: Performance for personalized image selection (%).

| Method  | Prec@3 | Recall@3 | F1@3 | Div@3 |
|---------|--------|----------|------|-------|
| random  | 4.87   | 6.14     | 5.43 | 30.24 |
| img     | 25.21  | 34.05    | 28.97| 17.12 |
| text    | 15.28  | 20.58    | 17.54| 18.68 |
| img+text| 25.21  | 34.37    | 29.09| 17.07 |

diversity, we introduce the truncated $div@K (K = 3)$. Formally, given $K$ images $\{i_1, \ldots, i_K\}$, $div@K$ is defined as:

$$div@K = \sum_{1 \leq m < n \leq K} \frac{dist(i_m, i_n)}{K(K - 1)/2}.$$  (12)

For textual explanations, we first evaluate with n-gram metrics: BLEU (n=1,4) [15], METEOR [5] and NIST (n=4) [6]. For diversity, we report DISTINCT [10], CLIP-Score [7], BERT-Score [27] are two embedding metrics for visual alignment and semantic quality.

Dataset & Implementation Details. We collected 1.77M google local reviews in the US, and filtered 108k high-quality samples. Data is split with 0.8/0.1/0.1 ratio for train/val/test. We use CLIP [17] with ViT-B/32 and GPT-2 small as our backbone models.

4.2 Framework Performance

Image set selection. Table 2 shows the performance of personalized image set selection. Using both user historical images and text can improve the performance over a single-modal model in terms of ranking performance (e.g., precision and recall). On the other hand, a random selection model can achieve higher diversity but with much worse ranking accuracy, thus indicating the importance to incorporate personalization into this task.

Text Generation. Results for text explanation generation are presented in Table 1. First, the clear gap between text-input models and image-input models on diversity metrics validates the benefits of incorporating visual features. The setting of visually-aware generation is able to generate accurate and diverse explanations. Second, $PC^2L$ shows substantial improvement on most of the metrics compared to other models. Though text-based models Ref2Seq and Peter achieve competitive results with our method on some n-gram metrics such as BLEU, their performance is much worse on diversity and embedding metrics.

4.3 Effectiveness of Contrastive Learning

We conduct ablation studies on different variations of our contrastive loss, as shown in Table 3. Specifically, CCL contributes more to the visual grounding by enforcing the model to distinguish random entities from the correct ones, and improves CLIP-Score compared to the vanilla contrastive framework [3]. PCL improves more on diversity by encouraging the model to focus on users with dissimilar interest. Overall, the model can benefit from both losses.

4.4 Human Evaluation

To fully evaluate our model, we randomly sample 500 test examples and conduct human evaluation on Amazon Mechanical Turk. Each example is scored by three random human judges with a 5-point Likert scale. We instruct annotators to consider expressiveness (coherent, diverse, no repetition) and visual alignment. $PC^2L$ significantly outperforms Ref2Seq (4.25 vs 3.72), which is consistent with the automatic metrics.

5 CONCLUSION

In this paper, to generate explanations with rich information for recommendations, we introduce a new task, namely personalized showcases and collected a large-scale dataset. We further design a personalized cross-modal contrastive learning framework to learn visual and textual explanations from user reviews. We hope our dataset and framework would benefit the community for future research on multi-modalities and recommendations.
REFERENCES

[1] Jinze Bai, Chang Zhou, Junshuai Song, Xiaoou Qu, Weiting An, Zhao Li, and Jun Gao. 2019. Personalized Bundle List Recommendation. The World Wide Web Conference (2019).

[2] Jun Chen, Han Gan, Kai Yi, Boyang Li, and Mohamed Elhoseiny. 2021. VisualGPT: Data-efficient Adaptation of Pretrained Language Models for Image Captioning.

[3] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In International conference on machine learning. PMLR, 1597–1607.

[4] Zhichong Chen, Yan Song, Tsung-Hui Chang, and Xiang Wan. 2020. Generating Radiology Reports via Memory-driven Transformer. arXiv preprint arXiv:2010.16056 (2020).

[5] Michael Denkowski and Alon Lavie. 2011. Meteor 1.3: Automatic metric for reliable optimization and evaluation of machine translation systems. In Proceedings of the sixth workshop on statistical machine translation. 85–91.

[6] George Doddington. 2002. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In Proceedings of the second international conference on Human Language Technology Research. 138–145.

[7] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. 2021. CLIPScore: A Reference-free Evaluation Metric for Image Captioning. arXiv preprint arXiv:2104.08718 (2021).

[8] Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2019. The curious case of neural text degeneration. arXiv preprint arXiv:1904.09751 (2019).

[9] Alex Kulesza and Ben Taskar. 2012. Determinantal Point Processes for Machine Learning. Found. Trends Mach. Learn. 5 (2012), 123–286.

[10] Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2015. A diversity-promoting objective function for neural conversation models. arXiv preprint arXiv:1510.03055 (2015).

[11] Lei Li, Yongfeng Zhang, and Li Chen. 2021. Personalized Transformer for Explainable Recommendation. In ACL/IJCNLP.

[12] Pij Li, Zhaohui Wang, Zhaochun Ren, Lishong Bing, and Wai Lam. 2017. Neural Rating Regression with Abstractive Tips Generation for Recommendation. Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval (2017).

[13] Jiannuo Ni, Jincheng Li, and Julian McAuley. 2019. Justifying recommendations using distantly-labeled reviews and fine-grained aspects. 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). 188–197.

[14] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. arXiv preprint arXiv:1807.03748 (2018).

[15] 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. 311–318.

[16] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. arXiv preprint arXiv:2103.00029 (2021).

[17] 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. In ICML.

[18] Tiejun Sun, Fei Wang, Ruwen Chen, David Lu, Dario Amodei, and Ilya Sutskever. 2019. Language Models are Unsupervised Multitask Learners.

[19] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In Advances in neural information processing systems. 5998–6008.

[20] Shaokun Wang, Tian Gan, Yuan Liu, Jianlong Wu, Yuan Cheng, and Liqiang Nie. 2022. Micro-influencer Recommendation by Multi-perspective Account Representation Learning. IEEE Transactions on Multimedia (2022).

[21] Mark Wilhelm, Ajith Ramanathan, Alexander Bronomo, Sagar Jain, Ed H. Chi, and Jennifer Gillenwater. 2018. Practical Diversified Recommendations on YouTube with Determinantal Point Processes. Proceedings of the 27th ACM International Conference on Information and Knowledge Management (2018).

[22] Ke Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron C. Courville, Ruslan Salakhutdinov, Richard S. Zemel, and Yoshua Bengio. 2015. Show, Attend and Tell: Neural Image Caption Generation with Visual Attention. In ICML.

[23] An Yan, Chaosheng Dong, Yan Gao, Jinmiao Fu, Tong Zhao, Yi Sun, and Jun Gao. 2019. Tell: Neural Image Caption Generation with Visual Attention. In ICML.

[24] An Yan, Zexue He, Xing Lu, Jiang Du, Eric Chang, Amilcare Gentili, Julian McAuley, and Jimmy Ba. 2018. VisualGPT: Simple framework for contrastive learning of visual representations. In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics. 311–318.

[25] Ke Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhutdinov, Richard S. Zemel, and Yoshua Bengio. 2015. Show, Attend and Tell: Neural Image Caption Generation with Visual Attention. In ICML.

[26] Hongyu Zang and Xiaojun Wan. 2017. Towards Automatic Generation of Product Reviews from Aspect-Sentiment Scores. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. 146–151.

[27] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. arXiv preprint arXiv:1904.09675 (2019).