Visual Question Answering for advertisement dataset by ViLT

Yohei Nishimura
ynishimura@wisc.edu

Zhikang Meng
zmeng34@wisc.edu

Duohan Zhang
dzhang357@wisc.edu

1. Introduction

Does a modern Vision-and-Language model identify the essence of the advertisement? We investigate how accurately computers recognize the images used in advertisements. The images for advertisements are entirely different from normal pictures; they have their own unique objects to encourage consumers to purchase the products. Furthermore, images in advertisements contain a great deal of information; they convey more than just images to humans due to the multi-modal information such as ad copies and the composition of product images, celebrities, backgrounds, logos, and copy. Also, advertisements encourage potential customers to purchase the products or services.

Recent multi-modal machine learning algorithms change the game in marketing; if computers capture the multiple meanings of advertisements, they can enhance the marketing area. For example, (1) they automatically extract insights (sentiment and actions) other than those assumed by the marketers and reflect these insights in new advertising plans, (2) they contribute to more accurate prediction of ad effects before placing ads, and (3) they improve matching between user preferences and advertisements. Advertisement images, however, have such ambiguous meanings that the computers are difficult to interpret what is the substance of them.

In order for computers to train the advertisement context, we select Visual Question Answering (VQA) task [1] for the advertisement-specific dataset [6], since VQA requires multi-modal information such as ad images and the text annotations per image: sentiments(what emotion the advertisement intend to evoke), symbols(what is the concept of the advertisement), and strategies(how to convey the messages of the ad to the customers), introducing the computer to the various perspective of the meanings per image.

To solve the VQA task with an ad-specific dataset, we focus on a multi-modal Vision-and-Language approach; we measure how much the algorithm can enhance the results. Specifically, we use Vision-and-Language Transformer (ViLT) [8], a simple model consisting of the transformer blocks jointly for the visual images and natural languages published in 2021. As it stems from the Vision-and-Language Pre-training models such as [12], ViLT improves the training speed to adopt the simple transformer blocks in the model. We fine-tune the pre-trained ViLT model to the advertisement dataset and experiment with how the model improves the accuracy for the inference of the VQA task of the dataset.

Our code is in public on the Github repository. Note that the required OS of the repo is Ubuntu 20.04.

2. Related Work

2.1. Marketing with Machine Learning

Research on marketing with advertisement images using machine learning is diverse. One central area is the prediction of advertising effectiveness; for example, [4] uses CNN algorithms to create predictive models of which online ads perform better. There are also many applications of computer vision research using deep learning; [19] builds the classification model of ad images using a CNN-based neural network.

In the field of ad image VQA task, [6] is the first research to understand content in advertisement images in terms of computer vision. After releasing this dataset annotated manually by cloudworkers, [16] conducts research on how to recommend the combination of the images and texts for advertisement systematically.

2.2. Visual Question Answering

VQA systems try to correctly answer questions for an image input by combining imaged-based models with Natural Language Processing (NLP) models. Compared with other vision-language tasks such as image captioning, VQA is more challenging because: (1) The questions are not predetermined, (2) The supporting visual information is high-dimensional, and (3) VQA necessitates solving many CV sub-tasks. Usually, a VQA algorithm contains three phases [14]: (1) image featurization and question featurization, (2) joint comprehension, and (3) answer generation. Many state-of-the-art VQA models utilize CNNs with their last layer removed, sometimes followed by a normalization [7] and dimensionality reduction to represent visual content. Word embeddings are used for question representation, such as count-based methods and prediction-based methods [15]. After the image and the question are processed independently to obtain separate vector representations, these features are mapped to a joint space, then combined and fed to the answer generation stage. Simple methods for consolidating image and question features include concatenation [21], but they ignore semantic relationships, and state-of-the-art techniques include joint attention models [13]. More recent works in VQA utilize transformer layers to align input text, and input image with self-attention [9]. In this project, we focused on a Vision-and-Language Transformer(ViLT) that is easy to use.

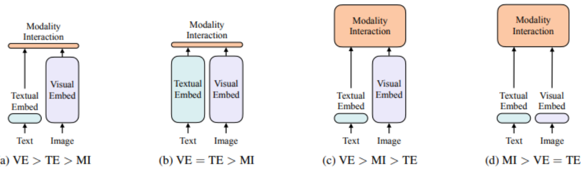
2.3. Vision-and-Language Model

Many existing vision and language models with complex text and image embedders exist with the taxonomy shown in Figure 1. For example, the CLIP model [17] deploys a complex image encoder ViT-L/14 and text encoder Transformer. However, its multi-model transformer is light-weighted and only calculates the cosine similarity of image and text embeddings. This leads to high computation costs due to imbalanced embedders and modality interactions. The Vilbert [11] uses complex co-attention transformer layers for image embeddings and uses Bert text encoders. This is also imbalanced with modality interaction and embedding. We will use the ViLT model [8], which will be introduced in the next section with straightforward architecture. ViLT is balanced between embedders and modality interactions and has high computation efficiency.

3. Methodology

Vision-and-Language Transformer(ViLT) [8] has a straightforward architecture for multimodal tasks, as shown in Figure 2. ViLT uses no regional feature proposals and no deep convolutional image while maintaining comparable performance to state-of-the-art multimodal models. Its simple design for both image and text encoders results in high parameter and time efficiency. Whole Word Masking and image data augmentation techniques are also employed by ViLT, which benefits downstream task performance. This model can be described as follows:

$$
\begin{aligned}
\bar{text} &= [text_{\text{class}}\,; text_1 T; \cdots ; text_L T] + T^{\text{pos}} \\
\bar{img} &= [img_{\text{class}}\,; img_1 I; \cdots ; img_N I] + I^{\text{pos}} \\
out^0 &= \left[\bar{text} + text^{\text{type}}; \bar{img} + img^{\text{type}}\right] \\
\hat{out}^d &= \text{MSA}\left(\text{LN}\left(z^{d-1}\right)\right) + out^{d-1}, \;\; d = 1 \ldots D \\
out^d &= \text{MLP}\left(\text{LN}\left(\hat{out}^d\right)\right) + \hat{out}^d, \;\; d = 1 \ldots D \\
pool &= \tanh\left(out_0^D W_{\text{pool}}\right)
\end{aligned}
$$



(a) VE > TE > MI

(b) VE = TE > MI

(c) VE > MI > TE

(d) MI > VE = TE

Figure 1. VE, TE, and MI are short for visual embedder, textual embedder, and modality interaction, respectively [8]. Vilbert is in category (b); CLIP is in category (c); ViLT is in category (d).

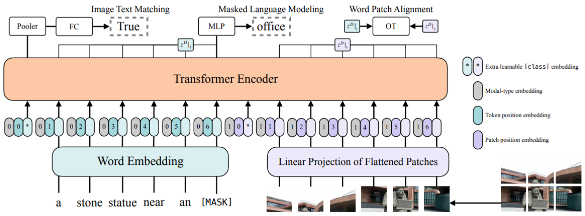


Figure 2. ViLT Model Overview [8]

Text input $text$ will pass through a linear projection layer T along with position encoding matrix T^{pos} to get text embeddings. Image input img will be sliced into a line of patches and then pass through the linear projection layer I long with position encoding matrix I^{pos} to get image embeddings. Text and image embeddings will be summed with their associating multi-modal embedder $text^{\text{type}}$, $image^{\text{type}}$ and then pass through the transformer encoder to get the final multi-modal output out. The transformer encoder composes of stacked blocks. Each block contains a multi-headed self-attention (MSA) layer and an MLP layer with layer normalization (LN) associated with each layer.

There are three objectives to be optimized in the ViLT model as follows:

- The Image text Matching (ITM) loss measures the capacity of the model to tell if the image input and text input matched or not, with fifty percent of image input replaced randomly during training. The ITM loss is calculated as the negative log-likelihood loss where the pooled output feature $pool$ is projected through an ITM head to logits.

- The Word Patch alignment (WPA) loss will measure the similarity between the text input distributions $out^D|_{text}$ and image input distributions $out^D|_{img}$ by using the IPOT [20] technique.

- The Masked Language Modeling (MLM) loss measures the capacity of the model to reconstruct the masked text input with fifteen percent of $text$ being masked randomly during training. The ground truth label is predicted from the masked text tokens $text_{\text{masked}}$ given its vector $out^D_{\text{masked}} \mid text$ in context.

The whole word masking technique in Chinese Bert [3] is deployed during pre-training with a probability of fifteen percent to force the model to learn information from all modalities. Since the ViLT model uses no regional proposal technique, it can deploy an image augmentation technique. The RandAugment [2] techniques are deployed during the pre-training except for the color inversion and cutout techniques to keep the color information and important objects information in images.

4. Experiment

4.1. Overview

We evaluate the ViLT model to compare with the baseline accuracy of ad dataset [6]. In this paper, the authors experiment with two tasks: (1) VQA on the action/reason and (2) classification of the topics and sentiments. The baseline of the original paper for the first task with LSTM and VGGnet-based architecture is 11.48%. The baseline for the second task is 60.34% of topics and 27.92% of the sentiments, respectively, with the architecture using 152-Layer ResNets.

4.2. Dataset

We use the images and text for the advertisement to fine-tune the ViLT model. The original image advertisement data can be found at here. Table 1 reports the summary of the dataset that consists of 64,832 image ads. For example, in the ”sentiment” subset, a sample question is ”What is the sentiment of this advertisement?” for the sample image 3, and possible answers are ”creative,” ”eager,” ”inspired,” ”persuaded,” and ”youthful.” Our task is to find the optimal answer given the image and question.

Type	Count	Example
Topic	204,340	Electronics
Sentiment	102,340	Cheerful
Q+A	202,090	I should bike because it’s healthy.
Symbol	64,131	Danger (+ bounding box)
Strategy	20,000	Contrast
Slogan	11,130	Save the planet... save you.

Table 1. Summary of the image dataset [6]

4.3. Implementation Details

We fine-tune the ViLT pre-trained model by the training set of the dataset, and we verify our results by the validation set of the dataset. We split the dataset into training and validation sets by 90 percent and 10 percent, respectively, because the dataset does not have a separation between training and validation data.

Assumptions of the experiment we use AdamW optimizer [10] along with the original paper of the ViLT model with a learning rate of $5e^{-5}$. The number of the training epoch is ten for all experiments (topics, sentiments, reasons, and actions). The batch size is 50, while the training and validation datasets ratio is 0.9.

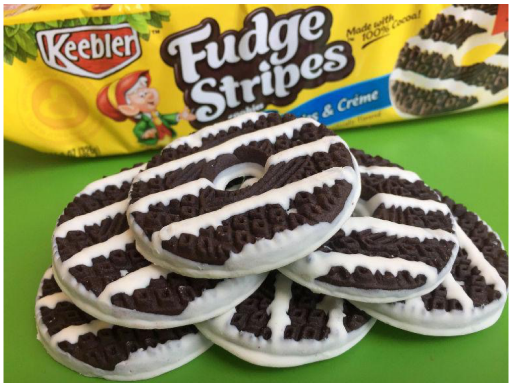


Figure 3. A sample image [6]

4.3.1 Topics and Sentiments

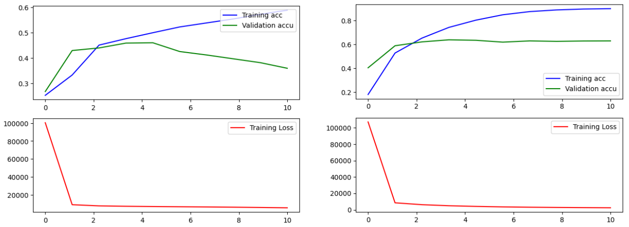
For the prediction of the topics and sentiments, we select the question ’what is the topic of the advertisement?’ for the topics, and that ’what is the sentiment of the advertisement?’ for the sentiments. Then, as in the standard VQA task setting [1] and [5], we prepare 3,129 candidate answers, adding the candidates’ words if the words in the training set are not included in the standard answers, obtaining the whole set of the candidate answers. Then, we obtain the embedding expressions from the sequence of the text data with Q&A and the image to put them into the transformer encoder in the ViLT.

Based on the assumptions above, we train the model using one RTX 3090 GPU.

4.3.2 Reasons and Actions

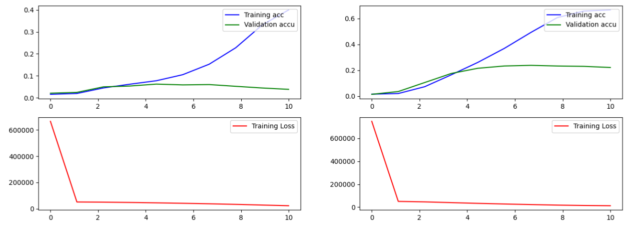
In order to experiment with the reason/action VQA task, first of all, we need to pre-process the annotation data because the dataset includes the ’sentence’ for action/reason, such as ’I should buy the bike because it’s healthy.’ Along with the original paper [6], we calculate the Term Frequency–Inverse Document Frequency (TFIDF) scores [18] on all action/reason answers to find the single word representing the answers with the highest TFIDF scores. Also, we create pairs of image and text between questions and one-word answers. This process makes sure that the single-word representation is most contentful. Then, we utilize the pair data to fine-tune the ViLT model.

We simplify the experiment by creating the candidate answer list from scratch (without 3,129 standard answers). It is because the calculated answers with the highest TFIDF



(a) Sentiments (b) Topics

Figure 4. The training and validation accuracy and the training loss of the classification. The x axis means the number of epochs.



(a) Reasons (b) Actions

Figure 5. The training and validation accuracy and the training loss of the classification. The x axis means the number of epochs.

scores rarely correspond to the standard answers, so in terms of the efficiency of the calculation, we remove the standard answers and build the candidate list only based on the results by TF-IDF. Except for this process, the same assumption is used in the experiments of topics and sentiments.

4.4. Results

The validation accuracy for the prediction of the topics after five epochs is 63.50%, which improves 3.16% from the baseline. Regarding the sentiments, the validation accuracy after five epochs is 46.01%, 18.09% better than the baseline. Based on the current experiments, ViLT improves the accuracy of the inference for both the topics and sentiments.

It is more challenging for computers to recognize the reasons and actions of the dataset; the validation accuracy of the reasons is 6.22%, while that of the actions is 23.87%. The baseline of the paper [6] is 11.48% by averaging the results from the reasons and actions; therefore, we can not compare them apple to apple. However, in the validation results, 350 is the number of the correct results in reasons, while 1,347 is the correct results in actions, calculating the summation of them in the total accurate number: 1,697. Then, the number of the total validation data between reasons and actions is 11,286, so we can average the validation accuracy by $\frac{1697}{11286} = 0.1504$, which is 15.04% increasing by 3.56% from the baseline.

One of the causes of the result that the reasons' accuracy is lower than the actions' might stem from the diversity of the answers. The answers to the reasons have more variety since annotators can write the answer without constraint, leading to more numbers of the reasons' classification labels than the actions in training/inference.

5. Conclusion

In this paper, we prove that the ViLT benefits the VQA task on the advertising dataset; we improve the baseline in all four classifications. Furthermore, our experiment proves the advantage of the transformer-based model for marketing data.

In practice, if we use a more specific, actual dataset, the model may contribute to the analysis of the consumers' insight. Because the advertising dataset in this experiment is experimental, it includes a mixture of data from multiple channels, such as posters, newspaper ads, package design, and digital image ads. In order to put this model to practical use, the model may be trained with the data concentrating on the specific channel and industry; the trained model may objectively extract what the ads identify, allowing the objective analysis of advertisements receiving good responses from consumers.

5.1. Contribution by each members

- Yohei Nishimura
  - Implemented codes for pre-processing dataset, training sentiments/topics and reasons/actions except for the calculation of TF-IDF (with Zhikang Meng)
  - Ran the code and visualize the results
  - Wrote the report on Section 1, 2.1, 4.1, 4.3, 4.4 and 5.
  - Created the slides of the presentation.
- Zhikang Meng
  - Literature review;
  - Presentation preparation;
  - Run codes for testing;
  - Write the report on Section 2.3 and Section 3;
  - Code for training reasons and actions (with Yohei Nishimura).
- Duohan Zhang
  - Literature review;
  - Write the report on Section 2.2 and 4.2;
  - Preparing and writing PPT for presentation.

References

[1] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. VQA: visual question answering. *CoRR*, abs/1505.00468, 2015. 1, 3

[2] Ekin D Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V Le. Randaugment: Practical automated data augmentation with a reduced search space. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 702–703, 2020. 3

[3] Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, and Ziqing Yang. Pre-training with whole word masking for chinese bert. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3504–3514, 2021. 3

[4] Michael Fire and Jonathan Schler. Exploring online ad images using a deep convolutional neural network approach, 2015. 1

[5] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering, 2016. 3

[6] Zaeem Hussain, Mingda Zhang, Xiaozhong Zhang, Keren Ye, Christopher Thomas, Zuha Agha, Nathan Ong, and Adriana Kovashka. Automatic understanding of image and video advertisements. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1705–1715, 2017. 1, 3, 4

[7] Kushal Kafle and Christopher Kanan. Answer-type prediction for visual question answering. pages 4976–4984, 06 2016. 1

[8] Wonjae Kim, Bokyung Son, and Ildoo Kim. Vilt: Vision-and-language transformer without convolution or region supervision, 2021. 1, 2

[9] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*, 2019. 2

[10] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization, 2017. 3

[11] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, 32, 2019. 2

[12] Jiasen Lu, Vedanuj Goswami, Marcus Rohrbach, Devi Parikh, and Stefan Lee. 12-in-1: Multi-task vision and language representation learning, 2019. 1

[13] Mateusz Malinowski, Carl Doersch, Adam Santoro, and Peter Battaglia. Learning visual question answering by bootstrapping hard attention. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 3–20, 2018. 2

[14] Sruthy Manmadhan and Binsu C Kovoor. Visual question answering: a state-of-the-art review. *Artificial Intelligence Review*, 53(8):5705–5745, 2020. 1

[15] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013. 2

[16] Shaunak Mishra, Manisha Verma, Yichao Zhou, Kapil Thadani, and Wei Wang. 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*, pages 2653–2660, 2020. 1

[17] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 2

[18] Gerard Salton and Christopher Buckley. Term-weighting approaches in automatic text retrieval. *Information Processing Management*, 24(5):513–523, 1988. 3

[19] An Tien Vo, Hai Son Tran, and Thai Hoang Le. Advertisement image classification using convolutional neural network. In *2017 9th International Conference on Knowledge and Systems Engineering (KSE)*, pages 197–202, 2017. 1

[20] Yujia Xie, Xiangfeng Wang, Ruijia Wang, and Hongyuan Zha. A fast proximal point method for computing exact wasserstein distance. In *Uncertainty in artificial intelligence*, pages 433–453. PMLR, 2020. 2

[21] Bolei Zhou, Yuandong Tian, Sainbayar Sukhbaatar, Arthur Szlam, and Rob Fergus. Simple baseline for visual question answering. *arXiv preprint arXiv:1512.02167*, 2015. 2