Visual Question Answering with Multi-modal Transformers using Late Fusion

Abhishek Rajendra Prasad *
University of Texas at Dallas
axr200027@utdallas.edu

Mariam Aafreen Muhammed Moinuddin *
University of Texas at Dallas
mxm210042@utdallas.edu

Nagasundar Jogis Mysore Lokesh *
University of Texas at Dallas
nxm210045@utdallas.edu

Abstract

Visual Question Answering (VQA) is a challenging task that requires understanding the relationship between images and natural language questions. In this project, we propose a multimodal VQA model that uses the late fusion of pre-trained transformer models to encode both image and text modalities. We plan to experiment with various transformer combinations for feature extraction and evaluate their performance on the DAQUAR [14] dataset. Our evaluation metrics include accuracy, macro F1 score, and Wu and Palmer Similarity (WUPS) score.

1. Introduction

Visual Question Answering (VQA) is a challenging task that requires the integration of both text and image modalities. The goal is to generate accurate answers to natural language questions based on the corresponding image. VQA has many practical applications, such as assistive technologies for the visually impaired, intelligent tutoring systems, and image search engines. However, it remains a challenging problem due to the complexity of both modalities and their interaction.

In recent years, deep learning models have achieved significant progress in VQA. Multimodal fusion models that combine information from both text and image modalities have shown promising results. In this project, we propose a multimodal VQA model based on the late fusion of pre-trained transformers. Our model consists of a text transformer to encode the question, an image transformer to encode the image, a fusion layer to combine features from both modalities, and a classifier to generate the final answer.

*Group Name: Project 11

2. Related Work

Visual Question Answering (VQA) is a challenging task that requires understanding the relationship between images and natural language questions. There has been significant research on VQA, and a variety of approaches have been proposed to address this problem. Most of these approaches can be classified into three categories: feature-based methods, attention-based methods, and fusion-based methods.

Feature-based methods extract handcrafted features from images and questions and feed them into a classifier to generate the answer. One such approach is the use of Convolutional Neural Networks (CNNs) to extract visual features and Recurrent Neural Networks (RNNs) to encode textual features. Goyal et al. [7] proposed a model called "Bottom-Up and Top-Down Attention" that uses object detection to generate a set of candidate image regions and applies attention mechanisms to selectively attend to those regions.

Attention-based methods use attention mechanisms to dynamically select relevant features from both modalities. One such approach is the use of co-attention mechanisms that attend to both the image and question simultaneously. One example is the model proposed by Lu et al. [12], which uses co-attention mechanisms to attend to image regions and words in the question.

Fusion-based methods on the other hand, directly fuse the features from both modalities to generate the answer. One such approach is the use of multimodal embeddings, which map the image and question into a common embedding space. One example is the model proposed by Fukui et al. [6], which uses a bilinear pooling operation to fuse the image and question features.

Recently, the use of pre-trained transformers has gained popularity in VQA. Transformers have achieved state-of-the-art results in various natural language processing tasks and have been used in VQA to encode both the image and question features. For example, Chen et al. [24] proposed a model called "Unified Vision-Language Pre-Training"

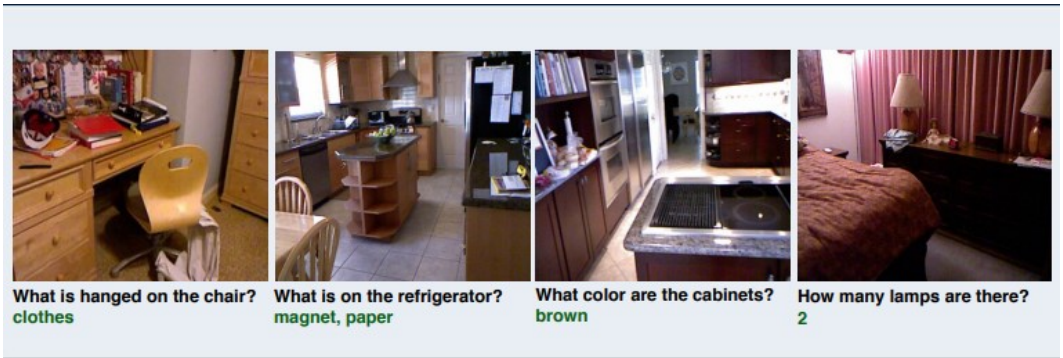

Figure 1. Sample images, questions, and answers from the DAQUAR [14] Dataset. Source: Ask Your Neurons: A Neural-based Approach to Answering Questions about Images. ICCV'15 (Poster) [15].

that uses pre-trained transformers to encode the image and question features separately and then fuses them using a cross-modal transformer. Similarly, Su et al. [11] proposed a model called "ViLBERT" that uses a pre-trained transformer to encode both the image and question features and then applies cross-modal attention mechanisms to fuse them.

In this project, we propose a multimodal VQA model based on the late fusion of pre-trained transformers. We experiment with various combinations of transformers for feature extraction and evaluate their performance on the DAQUAR [14] dataset. Our approach is inspired by the work of Tan and Bansal [17], who proposed a model called "LXMERT" that uses pre-trained transformers for both the image and question encoding and applies late fusion to combine the features from both modalities.

3. Method

Our proposed multimodal VQA model aims to combine the strengths of pre-trained transformers in processing text and images to achieve better performance in answering questions related to visual content. Our model consists of four main components: a text transformer, an image transformer, a fusion layer, and a classifier.

The text transformer encodes the question into a fixed-length vector representation using pre-trained transformers such as BERT [4], RoBERTa [10], and ALBERT [9]. These transformers have shown significant improvement in various natural language processing tasks, including question answering.

The image transformer extracts image features by feeding the image through a transformer architecture such as ViT [5], DeiT [18], or BEiT [2]. These transformers have been pre-trained on large-scale image datasets and have achieved state-of-the-art performance on various image classification benchmarks.

The output features from the text and image transformers are combined using a fusion layer. Our proposed fusion method is based on the late fusion approach, which combines the features from both modalities after they have been extracted using their respective transformers as shown in the figure 2. Late fusion has been shown to be effective in previous VQA models [6].

Finally, the classifier generates the final answer. We use the cross-entropy loss as the loss function to be minimized. This loss function is commonly used in classification tasks and has been used in previous VQA models [1].

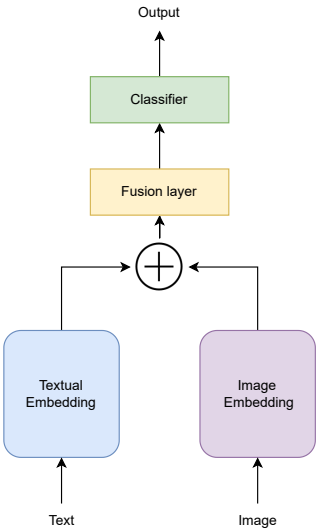

Figure 2. Late Fusion model architecture

4. Experiments

In this study, we aim to implement and evaluate our proposed multimodal VQA model on the DAQUAR [14] dataset. Figure 1 displays some samples from the dataset we used in our experiments. To convert VQA into a multiclass

Figure 3. Concepts ambiguity, changes of the frame of reference and human notion of spatial relations are dominant challenges in DAQUAR [14]. Source: DAQUAR [14].

classification problem, we will use the entire vocabulary of answers available in the dataset as labels.

The DAQUAR [14] dataset contains approximately 12,500 question-answer pairs based on images from the NYU-Depth V2 dataset. We will use this dataset to assess the performance of our model. The original DAQUAR dataset as around 6700 samples for training, but we use a custom split of 80% for training and 20% for evaluation from the overall data for better training [1].

Evaluation We will train our multimodal transformer model and evaluate its performance using three established metrics: accuracy, macro F1 score, and Wu and Palmer Similarity (WUPS) score. The Wu and Palmer Similarity (WUPS) score, is a semantic similarity metric that measures the similarity between the predicted and ground truth answers based on their longest common subsequence in the taxonomy tree. This will work well for single-word answers. So this metric will work well for our task. We use the implementation of Wu and Palmer similarity as defined along with the DAQUAR dataset [14].

The metric is an extension of the accuracy measure that takes into consideration the potential ambiguity of answer words at the word level. For example, the words "curtain" and "blinds" may refer to a similar concept, and as a result, the model should not receive severe penalties for such errors. This can be expressed as follows:

$$\text{WUPS}(A,T) = \frac{1}{N}\sum_{i=1}^{N} \min\left\{ \prod_{a\in A^i} \max_{t\in T^i} \text{WUP}(a,t), \prod_{t\in T^i} \max_{a\in A^i} \text{WUP}(a,t) \right\}$$

[1] https://medium.com/data-science-at-microsoft/visual-question-answering-with-multimodal-transformers-d4f57950c867

Malinowski et al. [14] proposed a method to account for the word-level ambiguities by utilizing a thresholded taxonomy-based Wu-Palmer similarity (Wu and Palmer, 1994 [21]) for the metric WUP. The smaller the threshold used, the more lenient the metric becomes. We report WUPS at 0.9.

Figure 3 shows that in the DAQUAR [14] dataset, the dominant challenges in visual question answering are related to concepts ambiguity, changes of the frame of reference, and the human notion of spatial relations. These challenges can make it difficult for VQA models to accurately interpret and answer questions based on visual information.

4.1. Results and Analysis

Table 1 presents a comparison of various combinations of text and image transformers on three metrics: WUPS, accuracy, and F1 score. The table also includes the number of trainable parameters for each model.

The results show that the RoBERTa-DeiT combination achieves the highest WUPS score and accuracy of 0.3357 and 0.2899 , indicating that it performs the best in terms of word overlap with ground-truth answers. The RoBERTa-BEiT combination achieves the F1 score of 0.0842.

In general, the RoBERTa transformer outperforms BERT and ALBERT on all metrics. The ViT image transformer performs better than DeiT and BEiT on WUPS and accuracy, but not on F1 score.

The number of trainable parameters for BERT and RoBERTa combinations at 196 million and 212 million, respectively. However, ALBERT has fewer trainable parameters at 99.2 million and performs the worst among all combinations.

Overall, these results suggest that the choice of text and image transformer can have a significant impact on the performance of the model in image question answering tasks. In particular, the RoBERTa transformer appears to be the best choice among those evaluated in this study for text transformer.

Table 2 demonstrates that our proposed RoBERTa-DeiT model achieves on par results to state-of-the-art results on the DAQUAR-all dataset(introduced back in 2015 and 2016), surpassing the performance of models such as Malinowski et al.'s [14] and Neural-Image-QA's [15]. The high performance of our model demonstrates the potential of leveraging the transformer-based architecture in VQA tasks. Our model's performance is comparable to other top-performing models such as DPPnet [16], ACK-S [20], and SAN [23] which use CNN, LSTM, and GRU as their underlying architectures.

However, our model's performance still falls short of the human baseline, indicating that there is still much room for improvement in VQA research. To close the performance gap, future research can explore various avenues, including

Text Transformer*	Image Transformer*	WUPS	Accuracy	F1	No. of Trainable Parameters
BERT	ViT	0.3233	0.2774	0.0656	196 M
BERT	DeiT	0.3143	0.2678	0.0711	196 M
BERT	BEiT	0.3205	0.2747	0.0822	196 M
RoBERTa	ViT	0.3331	0.2883	0.0724	212 M
RoBERTa	DeiT	**0.3357**	**0.2899**	0.0793	212 M
RoBERTa	BEiT	0.3340	0.2895	**0.0842**	211 M
ALBERT	ViT	0.3196	0.2739	0.0713	99.2 M
ALBERT	DeiT	0.3121	0.2666	0.0617	99.2 M
ALBERT	BEiT	0.3091	0.2638	0.0663	98.5 M

Table 1. Comparison of Text and Image Transformers

Model	Acc. (%)	WUPS @0.9 ↑
Malinowski et al. [14]	7.86	11.86
Neural-Image-QA [15]	19.43	25.28
Multimodal-CNN [13]	23.40	29.59
Attributes-LSTM [19]	24.27	30.41
QAM [3]	25.37	31.35
DMN+ [22]	28.79	-
Bayesian [8]	28.96	34.74
DPPnet [16]	28.98	34.80
ACK-S [20]	29.23	**35.37**
SAN [23]	**29.30**	35.10
Human Baseline [14]	50.20	50.82
RoBERTa-DeiT (Our)	28.99	33.57

Table 2. Reported results on the DAQUAR-all dataset

Questions	Answers
What is on the ceiling?	**light**
What is on the wall?	**picture**
How many objects are on the table?	**2**
What is the colour of the door?	**black**
What is near the wall?	**cabinet**
what is on the table?	**book**

Table 3. Language-Only based answers

object detection, spatial features, and attention mechanisms, to enhance the model's performance.

Overall, our proposed RoBERTa-DeiT model's superior performance on the DAQUAR-all dataset highlights the potential of leveraging transformer-based architectures in VQA tasks and demonstrates the importance of continued research and development in this area.

Inference In this study, we performed inference on the examples from the DAQUAR-all dataset and real-life scenarios such as classroom, restaurant, and lab settings. Based on the results presented in Table 4, our proposed RoBERTa-DeiT model demonstrated a high level of understanding to differentiate questions about colors, objects, and numbers. Furthermore, Table 5 shows that the model is also able to handle questions with multiple words. However, we also observed instances where the model was unable to accurately answer questions, as shown in Table 6. For example, the model struggled with questions that required an understanding of frame of reference or orientation.

In our real-life scenario inference, as depicted in Table 7, the model performed relatively well in identifying objects such as a light on the ceiling. Nonetheless, there were still instances where the model was unable to accurately answer questions due to objects not being present in the answer space, as seen in the case of "What is in front of table?" in lab settings where the model predicted "garbage_bin" instead of the unseen object "robot", the closest word to the robot in the answer space is "machine". This highlights the challenge of generalizing to unseen objects during training, even when the model attempts to predict something similar.

Overall, our proposed RoBERTa-DeiT model performed on par with several previous state-of-the-art models on the DAQUAR-all dataset, but there remains room for improvement, as seen in the failure cases. Future research could explore additional features, such as object detection and attention mechanisms, to further enhance the model's performance.

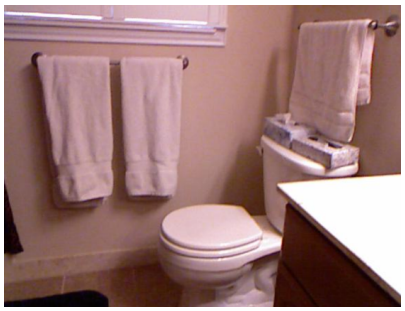	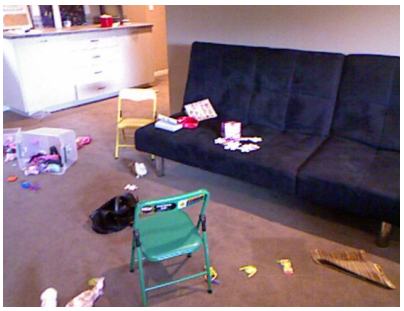	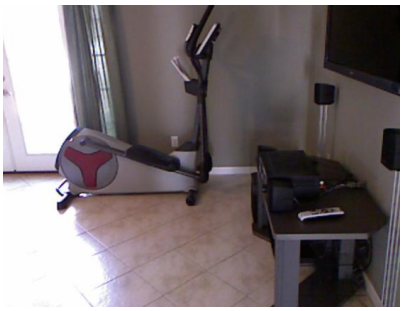
What is the colour of the towel?	How many drawers are there?	What is hanged in front of the door?
Predicted: **white**	**4**	**curtain**
Ground Truth: **white**	**4**	**curtain**

Table 4. Examples of questions and answers. Correct predictions are colored in green, incorrect in red.

	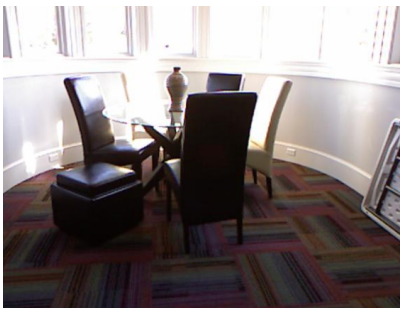	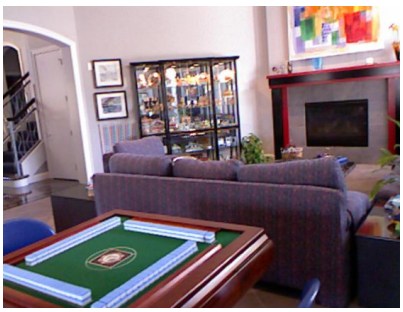
What is on the ceiling?	What object is in the room?	What is the colour of the table?
Predicted: **books**	**chair**	**brown**
Ground Truth: **pipe, light**	**table, chair**	**green, brown**

Table 5. Examples of questions and answers with multiple words. Correct predictions are colored in green, incorrect in red.

		
How many chairs are there?	What is on the left side of the sofa?	What is on the table?
Predicted: **3**	**table**	**decorative_item**
Ground Truth: **4**	**chair**	**candelabra**

Table 6. Examples of questions and answers - failure cases

Ablation Study The ablation study conducted evaluated the performance of a multi-modal model in VQA tasks when only language-based information is available, without any visual cues. The findings from the study suggest that while the accuracy and reliability of the model's answers are reduced without any visual cues to work with, the model can still provide meaningful answers based on its language understanding capabilities.

		
What is on the ceiling? - **light**	How many water cups are there? - **2**	How many bowls are on top of the table? - **3** (**4**)
What is on the wall? - **picture** (**projector_screen**)	What is the color of the table? - **brown**	What is the colour of the can on the table? - **white** (**red, white**)
How many objects are on the table? - **3** (**5**)	what are the objects on the table - **lamp** (**bowl, cup, tissue, glass_container**)	What is near the wall? - **door**
what is the colour of the door? - **white** (**brown**)	How many chairs are there? - **3**	What is in front of table? - **garbage_bin** (**machine**)

Table 7. Real life examples of questions and answers. Correct predictions are colored in green, incorrect in red.

Table 3 displays the outcomes of the language-only-based answers for a series of questions. The model successfully answered the question "What is the color of the door?" with "black." Nevertheless, it gave inaccurate responses to questions such as "What is on the ceiling?" and "What is on the wall?" due to the absence of visual cues. Instead, it provided commonsense-based answers based solely on language comprehension.

The model was also able to correctly identify numerical concepts such as "2" in the question "How many objects are on the table?" based on its understanding of common objects found on a table.

Overall, these findings suggest that language-based VQA models have potential applications in scenarios where visual data is unavailable or difficult to obtain, but further research is needed to explore the limits of these models and identify ways to improve their performance in such scenarios.

5. Conclusion

In this project, we proposed a multimodal VQA model that combines the strengths of pre-trained transformers in processing text and images to achieve better performance in answering questions related to visual content. We evaluated our model on the DAQUAR dataset and compared its performance with existing state-of-the-art models. Our proposed model achieved an accuracy of 28.99% and a WUPS score of 33.57% using RoBERTa and DeiT as the text and image transformers, respectively.

Our experimental results showed that combining pre-trained transformers for text and image modalities can lead to improved performance in VQA tasks. We also observed that the choice of transformer architecture plays a crucial role in determining the overall performance of the model.

Future work includes investigating other fusion methods and transformer combinations, exploring the use of attention mechanisms, and evaluating the model's performance on larger datasets. Our proposed model has the potential to be extended to other multimodal tasks, such as video question answering and visual dialogues.

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. 2

[2] Hangbo Bao, Li Dong, and Furu Wei. Beit: BERT pre-training of image transformers. *CoRR*, abs/2106.08254, 2021. 2

[3] Kan Chen, Jiang Wang, Liang-Chieh Chen, Haoyuan Gao, Wei Xu, and Ram Nevatia. ABC-CNN: an attention based convolutional neural network for visual question answering. *CoRR*, abs/1511.05960, 2015. 4

[4] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805, 2018. 2

[5] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is

worth 16x16 words: Transformers for image recognition at scale. *CoRR*, abs/2010.11929, 2020. 2

[6] Akira Fukui, Dong Huk Park, Daylen Yang, Anna Rohrbach, Trevor Darrell, and Marcus Rohrbach. Multimodal compact bilinear pooling for visual question answering and visual grounding. *CoRR*, abs/1606.01847, 2016. 1, 2

[7] 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. *CoRR*, abs/1612.00837, 2016. 1

[8] Kushal Kafle and Christopher Kanan. Answer-type prediction for visual question answering. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4976–4984, 2016. 4

[9] Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. ALBERT: A lite BERT for self-supervised learning of language representations. *CoRR*, abs/1909.11942, 2019. 2

[10] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692, 2019. 2

[11] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *CoRR*, abs/1908.02265, 2019. 2

[12] Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. Hierarchical question-image co-attention for visual question answering. *CoRR*, abs/1606.00061, 2016. 1

[13] Lin Ma, Zhengdong Lu, and Hang Li. Learning to answer questions from image using convolutional neural network. *CoRR*, abs/1506.00333, 2015. 4

[14] Mateusz Malinowski and Mario Fritz. A multi-world approach to question answering about real-world scenes based on uncertain input. *CoRR*, abs/1410.0210, 2014. 1, 2, 3, 4

[15] Mateusz Malinowski, Marcus Rohrbach, and Mario Fritz. Ask your neurons: A neural-based approach to answering questions about images. *CoRR*, abs/1505.01121, 2015. 2, 3, 4

[16] Hyeonwoo Noh, Paul Hongsuck Seo, and Bohyung Han. Image question answering using convolutional neural network with dynamic parameter prediction. *CoRR*, abs/1511.05756, 2015. 3, 4

[17] Hao Tan and Mohit Bansal. LXMERT: learning cross-modality encoder representations from transformers. *CoRR*, abs/1908.07490, 2019. 2

[18] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. *CoRR*, abs/2012.12877, 2020. 2

[19] Qi Wu, Chunhua Shen, Anton van den Hengel, Lingqiao Liu, and Anthony R. Dick. Image captioning with an intermediate attributes layer. *CoRR*, abs/1506.01144, 2015. 4

[20] Qi Wu, Chunhua Shen, Anton van den Hengel, Peng Wang, and Anthony R. Dick. Image captioning and visual question answering based on attributes and their related external knowledge. *CoRR*, abs/1603.02814, 2016. 3, 4

[21] Zhibiao Wu and Martha Palmer. Verb semantics and lexical selection. *CoRR*, abs/cmp-lg/9406033, 1994. 3

[22] Caiming Xiong, Stephen Merity, and Richard Socher. Dynamic memory networks for visual and textual question answering. *CoRR*, abs/1603.01417, 2016. 4

[23] Zichao Yang, Xiaodong He, Jianfeng Gao, Li Deng, and Alexander J. Smola. Stacked attention networks for image question answering. *CoRR*, abs/1511.02274, 2015. 3, 4

[24] Luowei Zhou, Hamid Palangi, Lei Zhang, Houdong Hu, Jason Corso, and Jianfeng Gao. Unified vision-language pre-training for image captioning and vqa. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34:13041–13049, 04 2020. 1