Faithful Multimodal Explanation for Visual Question Answering

**نویسنده اول: Jialin Wu**

**سال انتشار: 2019**

**محل چاپ: ACL**

**لینک به صفحه دانلود مقاله: https://arxiv.org/abs/1809.02805**

## در یک جمله بنویسید مقاله چه کار کرده

اومده همانند مقاله‌ی "Multimodal Explanations; Justifying Decisions and Pointing to the Evidence" یک مدل برای multimodal explanation بر روی مسئله VQA طراحی کرده ولی ادعا کرده که توضیحات تولید شده Faithful هستند یعنی دید واقعی تری از فرایندی که مدل بر اساس آن تصمیم گرفته است، میدهد.

## چکیده مقاله

AI systems’ ability to explain their reasoning is critical to their utility and trustworthiness. Deep neural networks have enabled significant progress on many challenging problems such as visual question answering (VQA). However, most of them are opaque black boxes with limited explanatory capability. This paper presents a novel approach to developing a high-performing VQA system that can elucidate its answers with integrated textual and visual explanations that faithfully reflect important aspects of its underlying reasoning process while capturing the style of comprehensible human explanations. Extensive experimental evaluation demonstrates the advantages of this approach compared to competing methods using both automated metrics and human evaluation.

## کارای قبلی که مقاله ذکر کرده چی بودند (background literature, previous work)

One approach to explainable VQA is to generate visual explanations, which highlight image regions that most contributed to the system’s answer, as determined by attention mechanisms (Lu Question: What sport is pictured? Explanation: Because the man is riding a wave on a surfboard. Answer: Surfing Figure 1: Example of our multimodal explanation. It highlights relevant image regions together with a textual explanation with corresponding words in the same color. et al., 2016) or gradient analysis (Selvaraju et al., 2017). However, such simple visualizations do not explain how these regions support the answer. An alternate approach is to generate a textual explanation, a natural-language sentence that provides reasons for the answer. Some recent work has generated textual explanations for VQA by training a recurrent neural network (RNN) to directly mimic examples of human explanations (Hendricks et al., 2016; Park et al., 2018). A multimodal approach that integrates both a visual and textual explanation provides the advantages of both. Words and phrases in the text can point to relevant regions in the image.

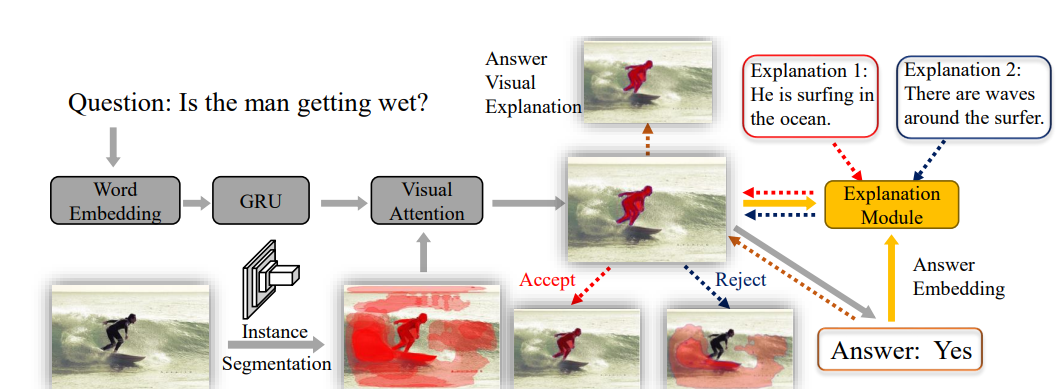
همانطور که از متن بالا مشخص است، مقالاتی را که روی تولید توضیحات برای تصمیم شبکه کار کردن، یا unimodal بودند یا multimodal و بهترین نتایج هم برای مقاله ای بود که روی حالت multimodal کار کرد.

## به طور خلاصه در یک پاراگراف بگید که مقاله، چالش، ایراد یا ضعف کارهای قبلی رو چگونه ارزیابی کرده

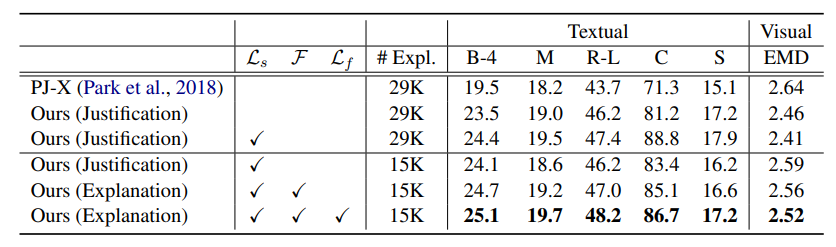
توی این مقاله، نویسنده ها ادعا میکنن که نتایج و توضیحاتی که مدل state of the art تولید میکنه، توضیح واقعی از فرایند سپری شده درون مدل نیست و میگه باید توضیحات شرط Faithful بودن رو رعایت کنن.

## ایده مقاله برای حل چالشها، ایرادها یا ضعف های پاراگراف قبل چی بوده (خلاصه)

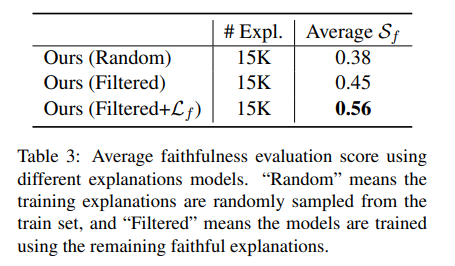
طراحی یک مدل جدید به شکل زیر:

نقاطی از تصویر که مدل به اونجاها توجه کرده تا جواب رو تولید کنه، توسط مدل مشخص میشه و توضیحات تولید شده که با این نواحی مطابقت ندارند، ریجکت میشوند. در واقع توضیحات هم روی یک تصویر، نواحی ای را مشخصی میکنند، اگر این نواحی با نواحی ای که در طی تولید جواب به آن توجه شده، مطابقت نداشته باشد، آن جمله و توضیح، جمله و توضیح مناسبی نیست و رد میشود.

## مقاله تا چه حد تونسته با روش پیشنهادی خودش به اون چالشها، ایرادها و ضعف ها پاسخ بده (metrics,evaluations,results)



همانطور که مشخص است، این مدل علاوه بر اینکه ادعا میکنه faithful هست، نتایج بهتری نیز روی دیتاست VQA-X نسبت به State of art تولید کرده است. اما برای سنجش میزان faithfulness نیز از معیارهایی استفاده کردند.



## 

## شما برای حل چالشهای گفته شده چه ایده ای دارین؟ (خلاصه)

تولید توضیحات بر اساس نواحی مورد توجه واقع شده با استفاده از تکنیک reinforcement learning

## شما برای بهبود این مقاله چه ایده ای دارین؟ (خلاصه)

استفاده از تکنیک های جدید در شبکه های عصبی و تست کردن مقادیر hyper parameters