Multimodal Explanations: Justifying Decisions and Pointing to the Evidence

**نویسنده اول: Dong Huk Park**

**سال انتشار: 2018**

**محل چاپ: IEEE CVPR**

**لینک به صفحه دانلود مقاله: https://arxiv.org/abs/1802.08129**

## در یک جمله بنویسید مقاله چه کار کرده

اومده برای اولین بار توضیحات Multi Modal رو معرفی کرده. یعنی علاوه بر مشخص کردن نقاطی از عکس که در جواب نهایی تاثیر داشتن، یک متن به زبان طبیعی هم خروجی میده که دلیل این انتخاب رو توضیح بده.

## چکیده مقاله

Deep models that are both effective and explainable are desirable in many settings; prior explainable models have been unimodal, offering either image-based visualization of attention weights or text-based generation of post-hoc justifications. We propose a multimodal approach to explanation, and argue that the two modalities provide complementary explanatory strengths. We collect two new datasets to define and evaluate this task, and propose a novel model which can provide joint textual rationale generation and attention visualization. Our datasets define visual and textual justifications of a classification decision for activity recognition tasks (ACT-X) and for visual question answering tasks (VQA-X). We quantitatively show that training with the textual explanations not only yields better textual justification models, but also better localizes the evidence that supports the decision. We also qualitatively show cases where visual explanation is more insightful than textual explanation, and vice versa, supporting our thesis that multimodal explanation models offer significant benefits over unimodal approaches.

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

L. A. Hendricks, Z. Akata, M. Rohrbach, J. Donahue, B. Schiele, and T. Darrell. Generating visual explanations. In Proceedings of the European Conference on Computer Vision (ECCV), 2016. 2, 6, 7

This was one of the previous works that worked on generating a textual explanation on classification problem on a fine-grained dataset of birds and flowers.

Other works include those generating probabilities of different regions using an attention mechanism.

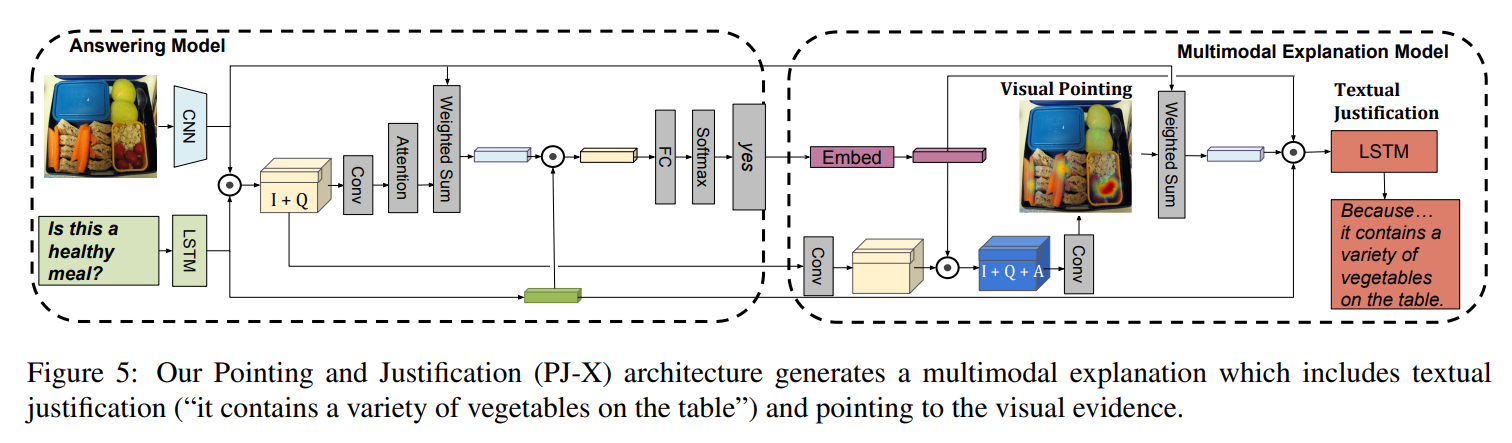
## به طور خلاصه در یک پاراگراف بگید که مقاله، چالش، ایراد یا ضعف کارهای قبلی رو چگونه ارزیابی کرده

کارهای قبلی unimodal بودند یعنی همزمان متن و مشخص کردن نقاط روی تصویر را انجام نمیدادند. همچنین این مدل دقت بهتری در هر دو زمینه از اینها کسب کرده است.

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

یک مدل جدید طراحی کردن به اسم PJ-X که مخفف pointing and justification explanations هست که دو قسمت داره، یه قسمت برای تولید جواب و قسمت بعدی که multimodal explanation model هست، برای تولید متن و مشخص کردن نقاط روی عکس با استفاده از مکانیزم attention.

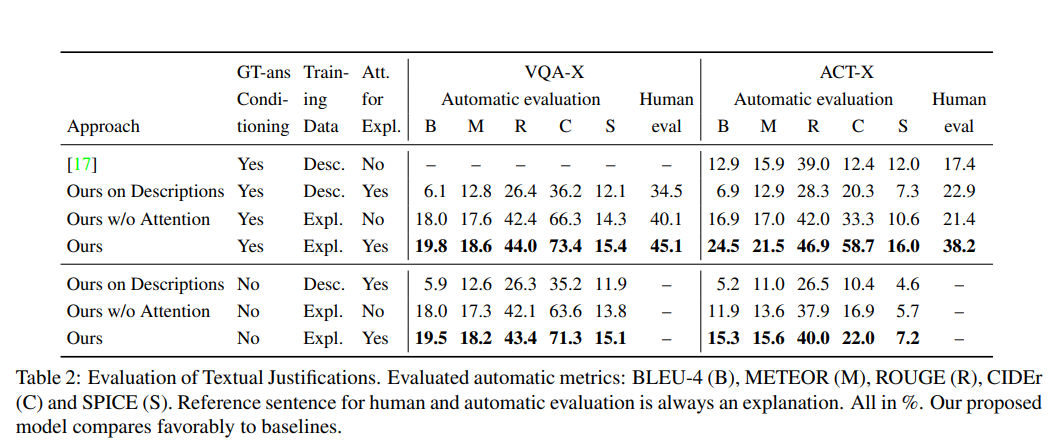
عکس زیر مدلشون هست که طراحی کردن:



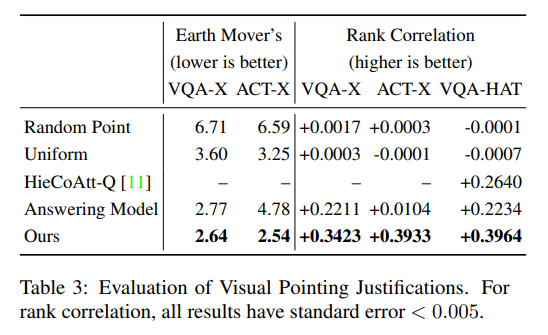
البته مدلشون شامل جزییات دیگه مثل توابع ضرر هم میشه که خیلی متوجه فرمولاش نشدم.

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

روی دو تا دیتاست که یکیش برای مسئله پرسش و پاسخ تصویری (VQA-X) و دیگری برای مسئله تشخیص فعالیت (ACT-X) مدلشون رو تست کردن. برای شباهت دو جمله از معیارهای متفاوتی استفاده کردند مثل BLEU-4, METEOR ، همچنین از انسان هم برای ارزیابی نزدیک بودن دو جمله استفاده کردن و در همه موارد به دقت بالاتری نسبت به state-of-the-art رسیدن.



برای بخش pointing to evidence هم از معیارهای Earth Mover’s, Rank Correlation استفاده کردند و بهترین نتایج را نسبت به state of the art کسب کردن.



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

Actually my idea is more or less same as this; extracting features from image using a CNN and from question (for VQA task) using a RNN and combine them. Details may be different like loss functions or hyper-parameters or other Neural Network techniques.

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

میتونن مدلشون رو روی دیتاست های دیگه و کاربردهای دیگر هم تست کنن و ببینن در چه حد generalize شده.