Learning Deep Representations of fine-grained visual descriptions

**نویسنده اول: Scott Reed**

**سال انتشار: 2016**

**محل چاپ: IEEE CVPR**

**لینک به صفحه دانلود مقاله: https://arxiv.org/abs/1605.05395**

## در یک جمله بنویسید مقاله چه کار کرده

Training a language model from scratch for the problem of zero-shot text-based image retrieval and image annotation

## چکیده مقاله

State-of-the-art methods for zero-shot visual recognition formulate learning as a joint embedding problem of images and side information. In these formulations the current best complement to visual features are attributes: manually encoded vectors describing shared characteristics among categories. Despite good performance, attributes have limitations: (1) finer-grained recognition requires commensurately more attributes, and (2) attributes do not provide a natural language interface. We propose to overcome these limitations by training neural language models from scratch; i.e. without pre-training and only consuming words and characters. Our proposed models train end-to-end to align with the fine-grained and category-specific content of images. Natural language provides a flexible and compact way of encoding only the salient visual aspects for distinguishing categories. By training on raw text, our model can do inference on raw text as well, providing humans a familiar mode both for annotation and retrieval. Our model achieves strong performance on zero-shot text-based image retrieval and significantly outperforms the attribute-based state-of-the-art for zero-shot classification on the Caltech UCSD Birds 200-2011 dataset.

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

کارای قبلی برای مسئله zero-shot visual recognition معمولا از joint embedding ای استفاده میشد که در اون attribute ها نقش مکمل تصاویر رو داشتن، البته روی image captioning قبلا کار شده بود ولی برای این مسئله zero-shot روی دیتاست هایی مثل Caltech USD birds با استفاده از این تکنیک کار نشده بود.

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

مشکلات و چالشهای استفاده از attribute و تشکیل یک وکتور از آنها برای این مسئله چالشهای زیر رو داشت:

1. finer-grained recognition requires commensurately more attributes,
2. and attributes do not provide a natural language interface

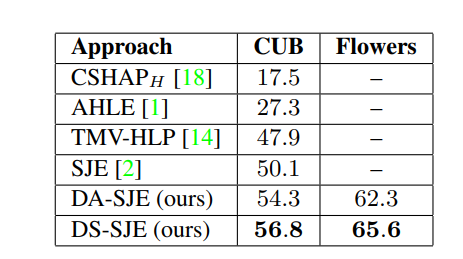
## ایده مقاله برای حل چالشها، ایرادها یا ضعف های پاراگراف قبل چی بوده (خلاصه)

اومدن از زبان طبیعی به عنوان رابط استفاده کردن و یه مدل زبانی رو از اول train کردن.

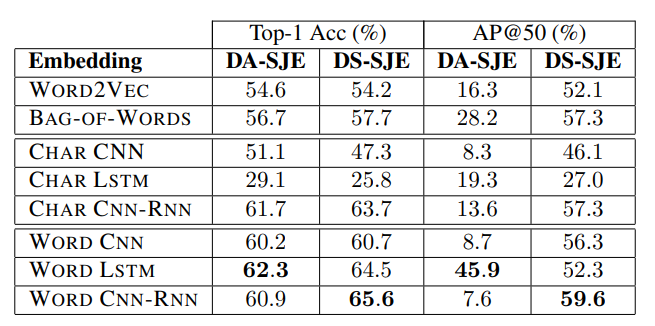
We propose to overcome these limitations by training neural language models from scratch; i.e. without pre-training and only consuming words and characters. Our proposed models train end-to-end to align with the fine-grained and category-specific content of images

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

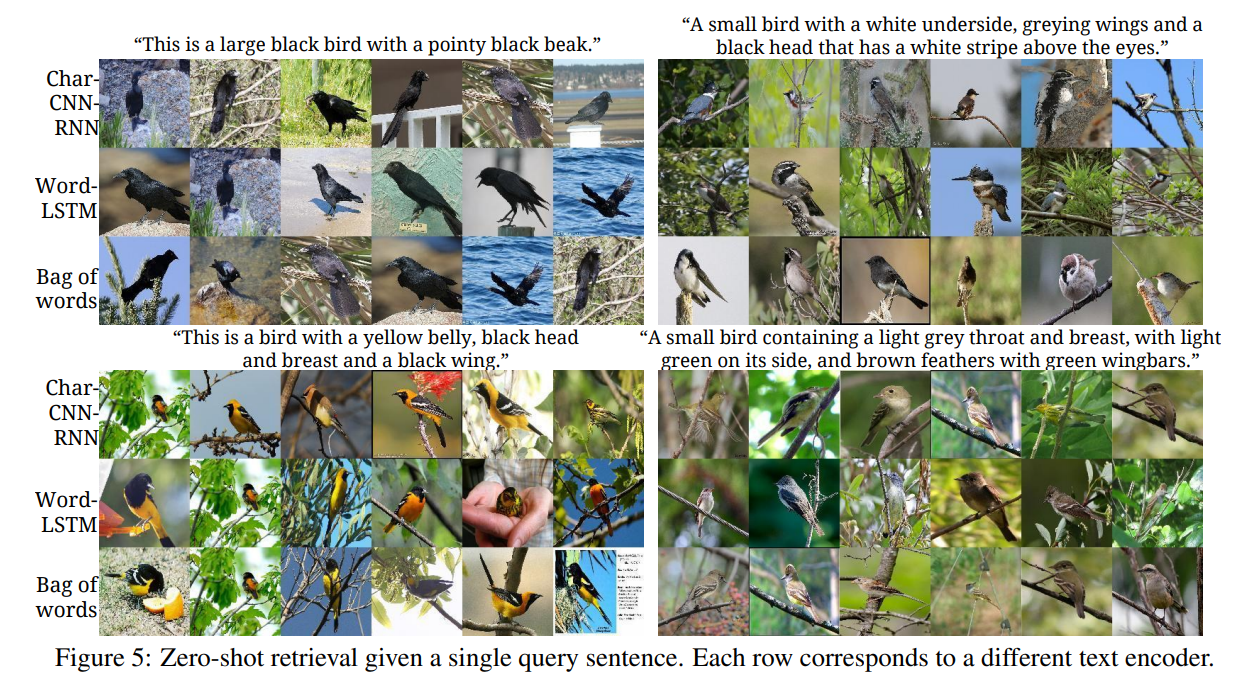
همونطور که مشخصه، نسبت به روشهای دیگر برای zero-shot visual recognition بیشترین دقت را کسب کرده است هم بر روی دیتاست CUB و هم Flowers



توی این قسمت هم اومده با مدل های زبانی متفاوتی train کرده و بهترین نتیجه بر اساس top-1 accuracy برای Word CNN-RNN بوده است.



این هم نتایج image retrieval است که ورودی آن یک کوئری به زبان طبیعی است.



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

به نظرم میشه از reinforcement learning برای مسئله zero-shot visual recognition استفاده کرد، کم کم یاد بگیره چه ویژگیهایی برای تمایز بین اشیاء مناسب است. هرچند برای جزئیاتش هیچ ایده ای ندارم و دانشش رو هم ندارم.

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

از آنجا که هنوز چیزی درباره sequence models, RNN, LSTM نمیدانم، نمیتوانم برای بهبودش چیزی بگم.