Multimedia Search from Composite Inputs

Adolfo Apolloni, Alex Raistrick, Omar Al-Ejel

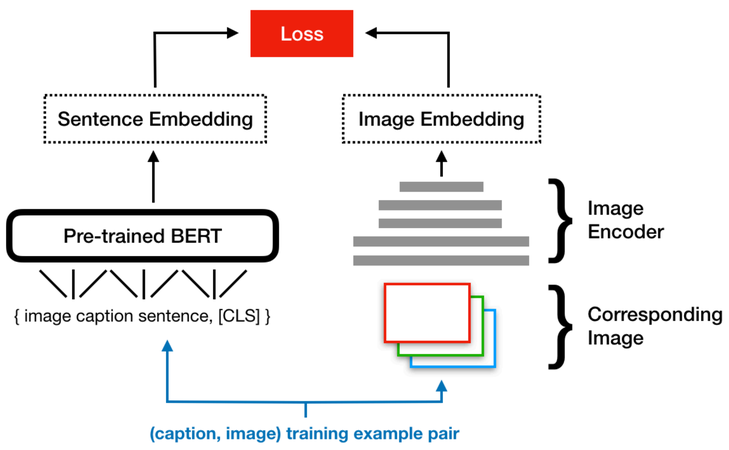
EECS 442: Computer Vision

December 10, 2019

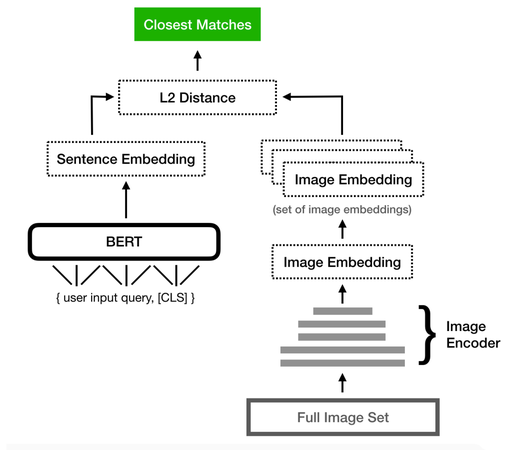
IMAGE-TEXT SEARCH

PROJECT MOTIVATION

Traditional image search relies on alt-text contextualization of image data. We developed a learned image search system that matches queries directly to image

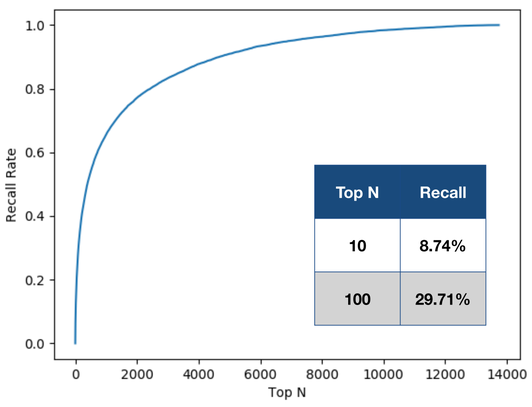


Training architecture for image embedding generation used for image search.



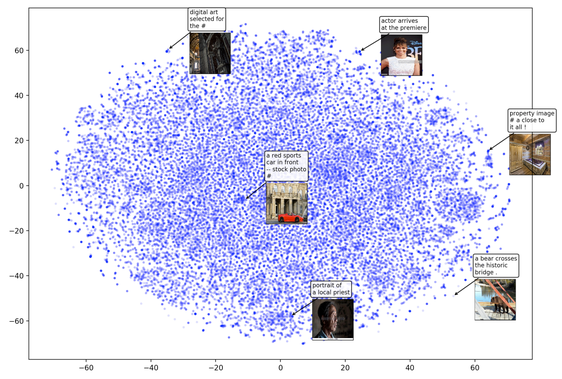
Search string embedding to image embedding search architecture.

TOP-K ACCURACY CURVE



Top N	Recall
10	8.74%
100	29.71%

T-SNE

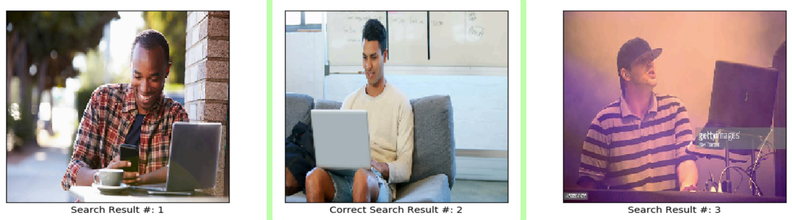


T-SNE chart displaying low-dimensional embeddings of image caption BERT vectors.

VALIDATION SET RESULTS

"man having a video chat sitting on the couch"

Model recognizes that best results have both a man and laptop.



"the moon oribits the no atmosphere some water . as ice"

The model embeds powerpoint slides similarly.

"close up of a pizza"

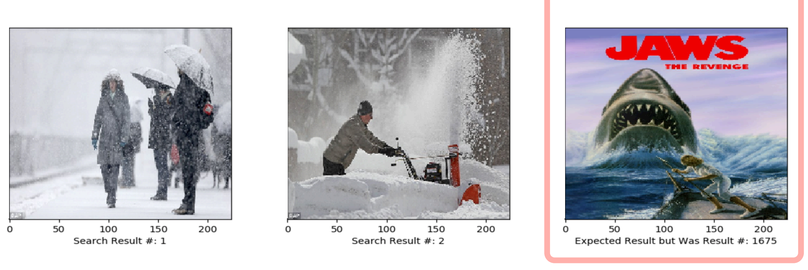
The model completely fails.

"stuffed roasted red peppers with cherry tomatoes , white beans and olives ."

The model correctly identifies cooking photos, but not specific ingredients.

"natural horror film - review : it is the christmas season ."

Poor captions yield poor results.

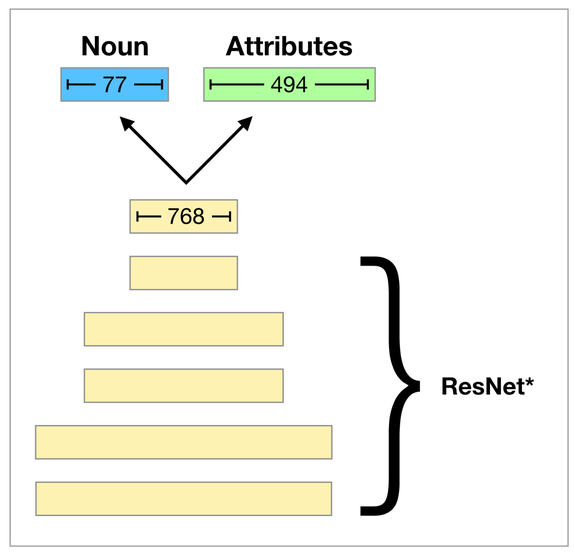


HYBRID IMAGE-TEXT MODEL

MOTIVATION

Past work composes fixed format caption and image data for search. We propose an adapted transformer architecture for learning on mixed sequences of image and text tokens to allow composite input.

IMAGE PRE-TRAINING ARCHITECTURE



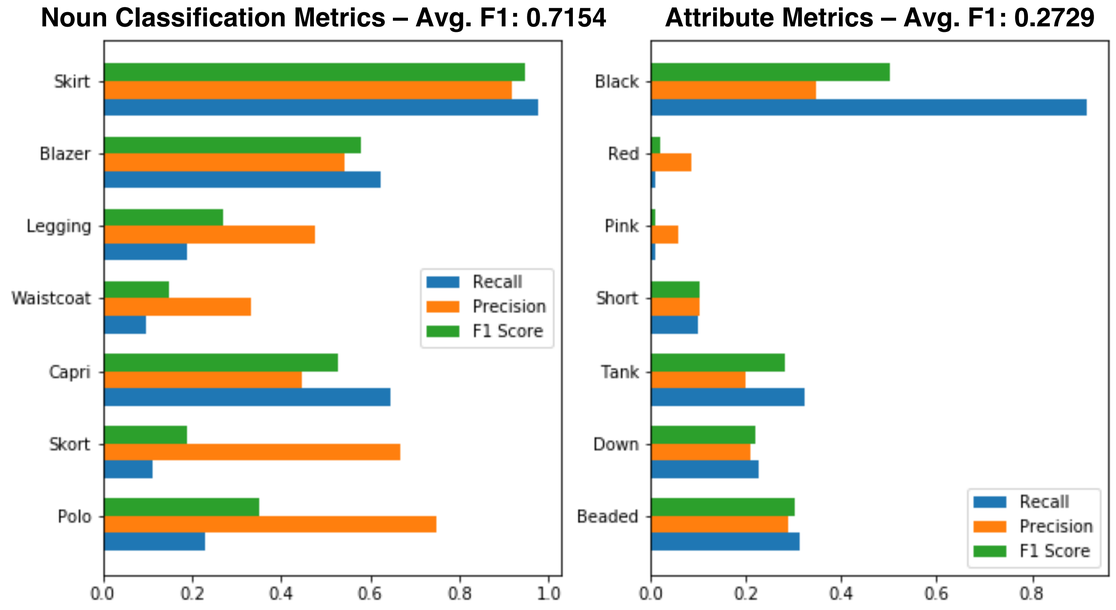
We pretrain the image encoder by training to classify fashion nouns and attributes.

After pretraining, the ResNet model is used to initialize the image encoder in the model below.

PRE-TRAINING RESULTS

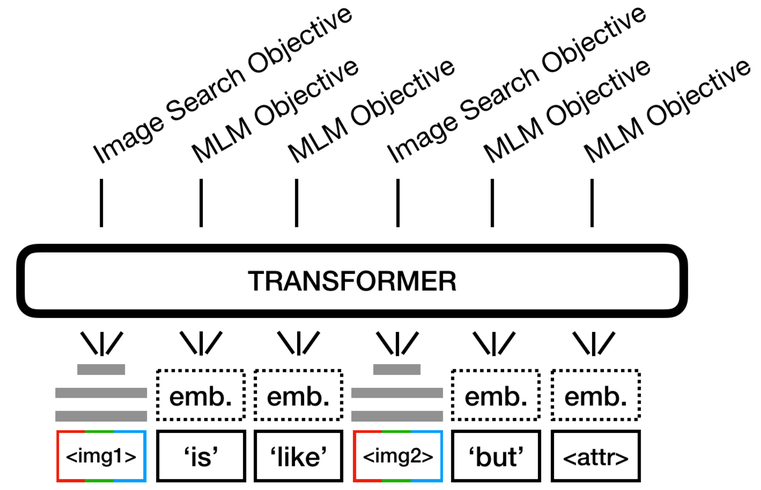
Noun Classification Metrics – Avg. F1: 0.7154

Attribute Metrics – Avg. F1: 0.2729



Nouns (left), attributes (right) precision recall metrics showing Y-Net performance on 'interesting' examples. Note that performance is much better on nouns than attributes on average.

HYBRID MODEL ARCHITECTURE



BERT and Image Embedder Model used for image search from multimodal text and image input.