Identifying Complex Metadata Relationships to Objects using NLP and Computer Vision

Benjamin Wilke

Tej Tenmattam

Anand Rajan

Data Science @SMU

Team & Advisors

Tej Tenmattam, Anand Rajan and Benjamin Wilke – Graduate Students SMU

Joel Lindsey – Adjunct Professor SMU

Andrew Pollock - Getty Images

Agenda

- Problem Statement
- Use Case Review and Refinement
- Encoder/Decoder Architectures
- Training Dataset
- Key Next Steps
- Key Novel Contributions

Problem Statement

With simple text searching methods often the order and context of words are lost or conflated.

Through integration of natural language processing (NLP), image classification, and transfer learning our goal is to effectively match specific relationships of words in image metadata with the corresponding relationships of objects that are present in an image

Use Cases

Assisting the Disabled - the integration of these technologies can provide for better computer vision generated descriptions about events, themes, moods, and object interactions found in an image. Conversely, descriptions and word pairs can match images to those who cannot speak.

Improving Search Results – by maintaining the context and ordering of adjectives, prepositions, and prepositional phrases the results of a search can more closely align with user intentions.

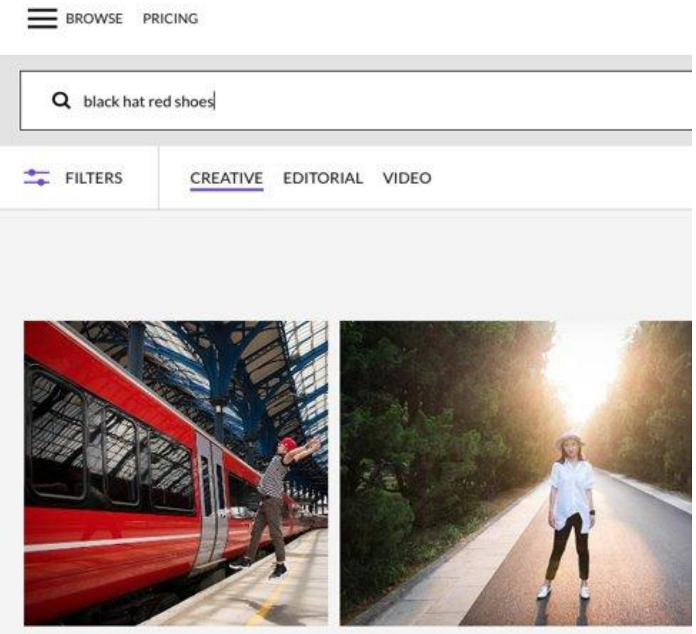
Improving Search Results

Simple search:

"black + hat + red + shoes"

Search preserving context:

"(black + hat) + (red + shoes)"



black hat red shoes

Improving Search Results - Value Verification

Getty Images has identified this as a key challenge in their business.

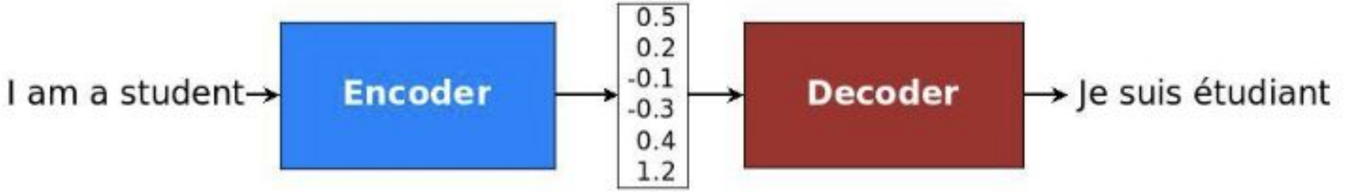
Image search on something like Google Images may “cast a wider net” by including related images - however, search features for image brokers must be able to be tuned to precise results.

Market differentiation for image brokers is tightly coupled with search capabilities as this is the primary function on their sites.

Encoder/Decoder Architecture

There is a lot of work that has been done on generating captions for images using encoder/decoder or sequence-to-sequence deep neural architectures

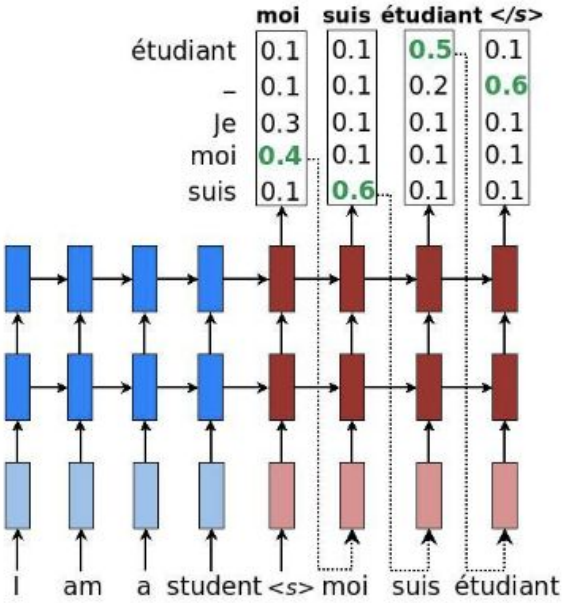
Historically, these architectures innovated neural machine translation (NMT)



Greedy Decoding

This is a greedy implementation of the decoder

At each step the most likely word (maximum logit value) is chosen



	moi	suis	étudiant	</s>
étudiant	0.1	0.1	0.5	0.1
_	0.1	0.1	0.2	0.6
Je	0.3	0.1	0.1	0.1
moi	0.4	0.1	0.1	0.1
suis	0.1	0.6	0.1	0.1

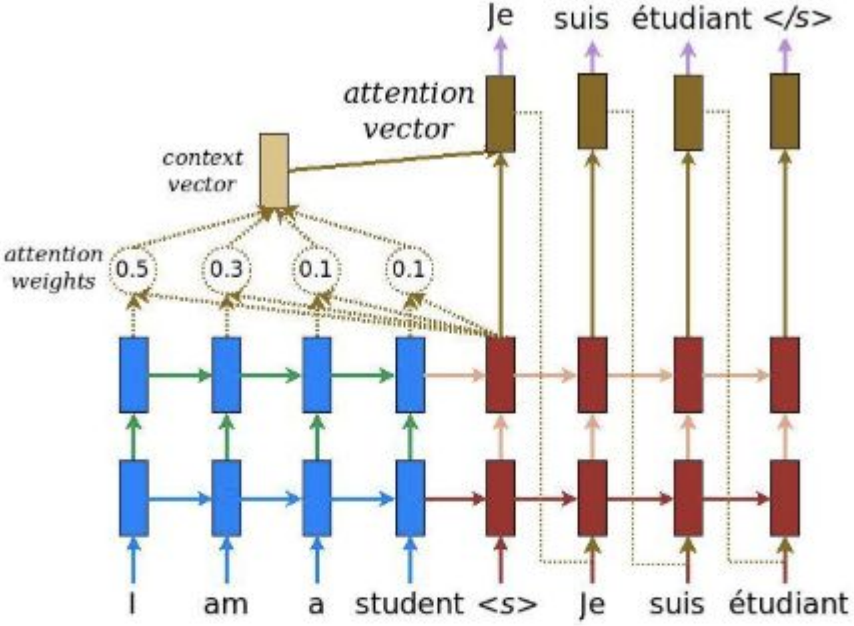
Decoding with “Attention”

Popularized in a paper by Bahdanau et al.

At each decoding step current target hidden state is compared with source states to derive attention weights.

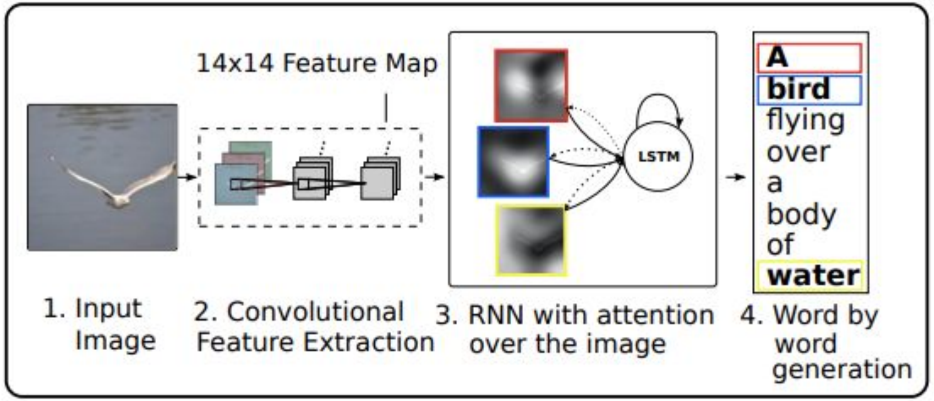
This produces a context vector, which is the weighted average of the source states.

The context vector is added to the current hidden state to produce final attention vector



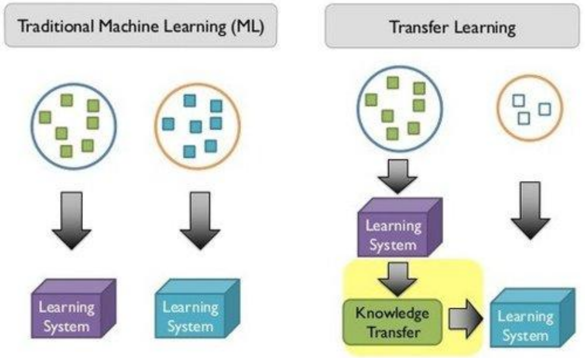
Encoder/Decoder Architecture with Images

Intuitively, the Encoder could be replaced with a CNN to process images as inputs and produce the embeddings to be fed to the decoder.


14x14 Feature Map
LSTM
A
bird
flying
over
a
body
of
water
1. Input Image
2. Convolutional Feature Extraction
3. RNN with attention over the image
4. Word by word generation

CNN Transfer Learning

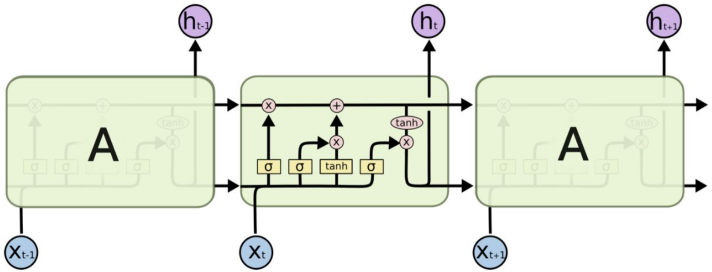
- We will still be employing Transfer Learning for our CNN encoder as proposed in our last presentation
- We have explored implementation schemes utilizing feature extraction and “fine tuning” approaches
- Our training iterations will explore many popular pretrained architectures, including:
  - VGG
  - ResNet
  - Inception
  - Inception-ResNet
  - Xception



RNN Architecture (LSTM)

Our decoder will be an RNN consisting of Long Short Term Memory (LSTM) modules.

This is required to fulfill requirements of long-term dependencies of word generation, while also avoiding the vanishing gradient problems with standard RNN modules



Microsoft COCO: Common Objects in Context

The COCO image dataset seeks to elevate simple object detection by placing objects in their natural context and encouraging complex scene understanding.

To achieve these goals the COCO dataset provides non-iconic views of common objects in scenes comprised of many objects versus objects in isolation.

Ultimately their goal is to provide a dataset to advance research on contextual reasoning via computer vision related tasks.

Key Next Steps

- Identify subset of object (noun) and simple bigram or trigram descriptors (adjectives) to focus training and testing effort
- Align with Getty Images on availability of testing images
- Mine MSCOCO for existing language relationships in provided descriptions
- Manually annotate training images as required
- Build encoder (CNN) / decoder (RNN) architecture implementing attention
- Evaluate, iterate, and improve results over validation sets

Key Novel Contributions

- Searching for specific object attributes in complex images with objects in context
- Most of the work we have encountered in our research involves generating entire scene descriptions, which speak more to the interactions between objects than accurately elaborating on descriptors of objects within the scene
- This can hopefully be achieved by mining validated training datasets for specific short adjective to noun relationships (bigrams, trigrams) or manually annotating images for specific target results
- Other ideas specific to Getty Images include building a feedback loop into the results - where a human can validate that the search terms they provided are contained in the image.
    - This information can be stored against the image for additional model training or image weighting when it sees those search terms again

Questions?