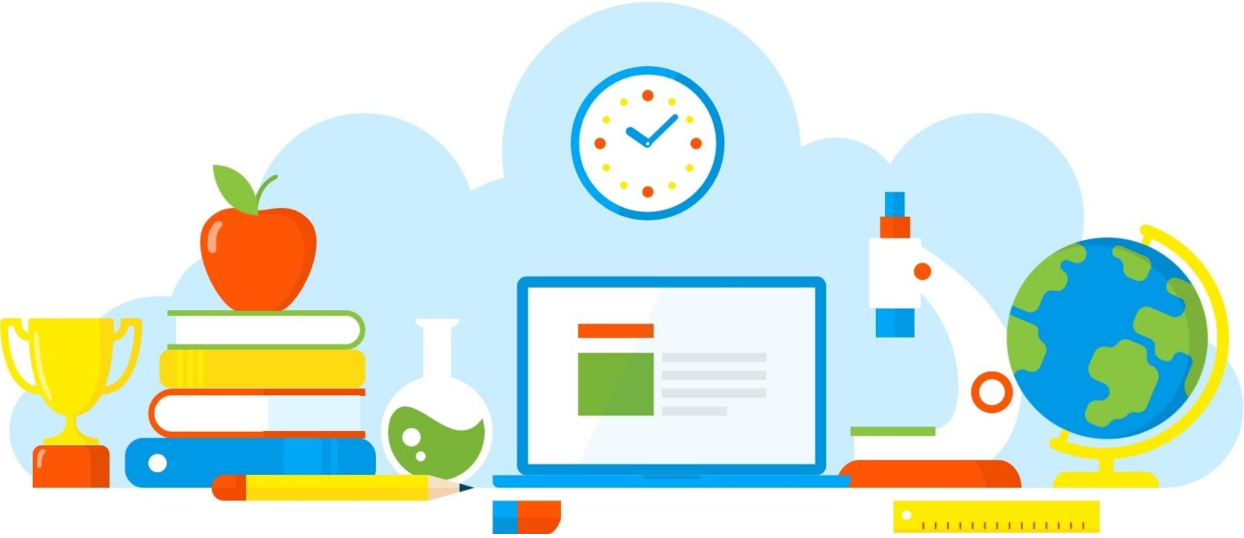


Independent Study Report

Text to Image Generation

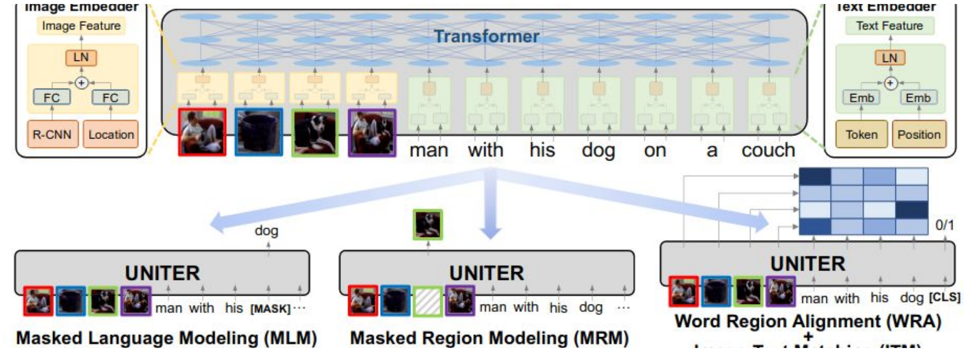
Weekly Report:Neural Text Transfer

Pulkit Gera
20171035

Introduction

The objective of this study is to examine the different methods of text to image tasks ranging from GANs to Transformer based and produce a novel captioned dataset to accompany the same. We do a study of different methods as well as datasets and their feasibility to be applied in this setting. We also implement MirrorGAN, a text to image GAN based model.

Transformer Architecture



Transformers architecture was first introduced in ViLBERT. It extends BERT to a multi-modal architecture tackling problems like Caption Based Image Retrieval, Visual Q&A. Both visual and textual information are processed in separate streams and then combined through co attentional transformer layers.The pretrained model can be then used for a variety of tasks like Visual Question Answering, Visual Commonsense Reasoning, Caption Based Image Retrieval,etc.

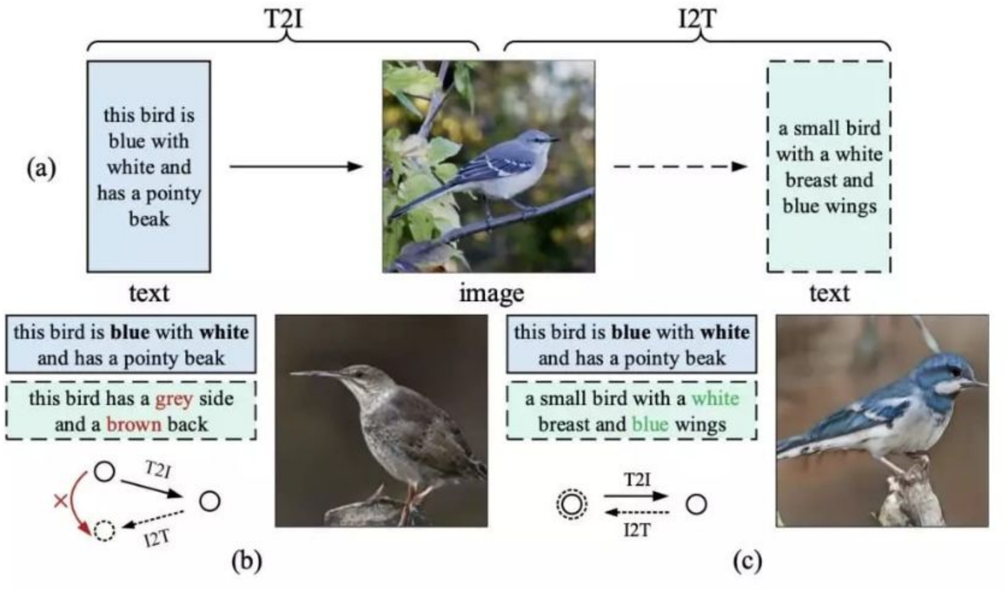
LXMERT consists of 3 Transformer encoders: Object Relationship Encoder, Language Encoder, Cross Modality Encoder. The idea is to not only improve cross modality but also intra modality i.e infer masked modality either from same elements in same modality, or from aligned components in other modality.

UNITER first encodes image and text regions into a common embedding space. Then a transformer module is applied to learn a cross modality contextualized embedding across the 2 spaces.The idea is to minimize the cost of transporting one distribution to another. The idea is to optimize towards better cross modal alignment.

LXMERT is similar to ViLBERT in the fact that they both model modalities separately. However, the feature representation is very different in LXMERT with it using RoI as well Bounding Box coordinates from Fast RCNN. Also it makes an effort to optimize for Visual QA.

UNIter on the other hand tries to optimize them in the same embedding space.

GAN based Architecture


T2I
I2T
(a)
this bird is blue with white and has a pointy beak
a small bird with a white breast and blue wings
text
image
text
this bird is blue with white and has a pointy beak
this bird has a grey side and a brown back
T2I
I2T
(b)
this bird is blue with white and has a pointy beak
a small bird with a white breast and blue wings
T2I
I2T
(c)

Stack- GAN comprises 2 GANs to generate photo-realistic images conditioned on text descriptions. The Stage-I GAN sketches the primitive shape and basic colors of the object based on the given text description, yielding Stage-I low resolution images. The Stage-II GAN takes Stage-I results and text descriptions as inputs, and generates high resolution images with photo-realistic details.

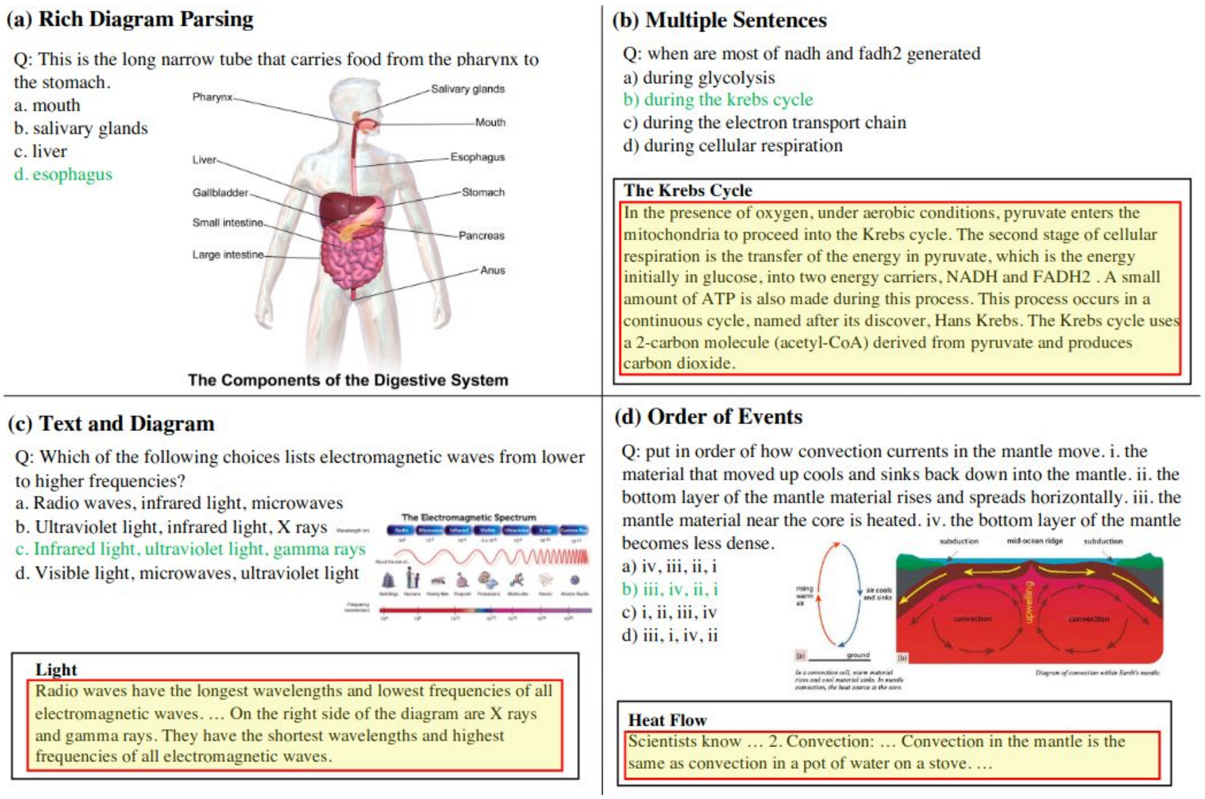
AttnGAN use a combination of Attention and GAN while iteratively fixing the images. Instead of just using word embeddings, we use Attention to better capture the context.It captures both both the global sentence level information and the fine-grained word level information

MirrorGAN works similar to CycleGAN in the sense that the text caption for the image and the caption generated by the image must be semantically consistent. Therefore reducing the T2I and I2T distance. More details given below.

These networks encode the text through RNN and LSTM based architecture and while they are good at understanding Image features, they dont fully generalize to the word space as well as Transformers are able to do

Datasets

We first attempted to create a leaves datasets by leveraging captions and images of wikipedia. However that failed due to the captions not describing any visual part of the image. This fails the purpose. We do an extensive study on how Oxford-102 and CUB-200 dataset were curated. We use the same methodology for curating the dogs dataset.



(a) Rich Diagram Parsing

Q: This is the long narrow tube that carries food from the pharynx to the stomach.
a. mouth
b. salivary glands
c. liver
d. esophagus

(b) Multiple Sentences

Q: when are most of nadh and fadh2 generated
a) during glycolysis
b) during the krebs cycle
c) during the electron transport chain
d) during cellular respiration

The Krebs Cycle

In the presence of oxygen, under aerobic conditions, pyruvate enters the mitochondria to proceed into the Krebs cycle. The second stage of cellular respiration is the transfer of the energy in pyruvate, which is the energy initially in glucose, into two energy carriers, NADH and FADH2 . A small amount of ATP is also made during this process. This process occurs in a continuous cycle, named after its discover, Hans Krebs. The Krebs cycle uses a 2-carbon molecule (acetyl-CoA) derived from pyruvate and produces carbon dioxide.

(c) Text and Diagram

Q: Which of the following choices lists electromagnetic waves from lower to higher frequencies?
a. Radio waves, infrared light, microwaves
b. Ultraviolet light, infrared light, X rays
c. Infrared light, ultraviolet light, gamma rays
d. Visible light, microwaves, ultraviolet light

Light

Radio waves have the longest wavelengths and lowest frequencies of all electromagnetic waves. … On the right side of the diagram are X rays and gamma rays. They have the shortest wavelengths and highest frequencies of all electromagnetic waves.

(d) Order of Events

Q: put in order of how convection currents in the mantle move. i. the material that moved up cools and sinks back down into the mantle. ii. the bottom layer of the mantle material rises and spreads horizontally. iii. the mantle material near the core is heated. iv. the bottom layer of the mantle becomes less dense.
a) iv, iii, ii, i
b) iii, iv, ii, i
c) i, ii, iii, iv
d) iii, i, iv, ii

subduction
mid-ocean ridge
subduction
rising warm air
air cools and sinks
convection
upwelling
convection
(a)
ground
(b)
In a convection cell, warm material rises and cool material sinks. In mantle convection, the heat source is the core.
Diagram of convection within Earth's mantle.

Heat Flow

Scientists know … 2. Convection: … Convection in the mantle is the same as convection in a pot of water on a stove. …

Meanwhile we also try to leverage NCERT textbooks as a potential source for datasets. They work well because they explain a concept in detail and have images which are closely related to the same concept. We find a similar dataset curated by Allen AI called TQA dataset which consists of text and images explaining a concept and performing Question Answering for the problem. We contrast this with Recipe QA, another multimodal dataset which involves filling the gaps with words and images. They help us understand that these datasets do not necessarily work in Text to Image tasks because they are complementary and help to explain a step or concept and when taken in isolation don't quite give us enough cues to create the image. For eg, we have a diagram explaining water cycle and explanation for water cycle but in isolation isnt enough to generate the image.

Captioned Dogs Dataset

The curated dataset has 3177 images out of which 3000 were kept for training and 177 for testing.

Instructions:
https://iiitaphyd-my.sharepoint.com/:b:/g/personal/pulkit_gera_research_iiit_ac_in/EdsuFtvSLV1NhKs3WY7ZpqQBMOA0tMwFLkVbo1UbbeI5Qw?e=SxsAgZ

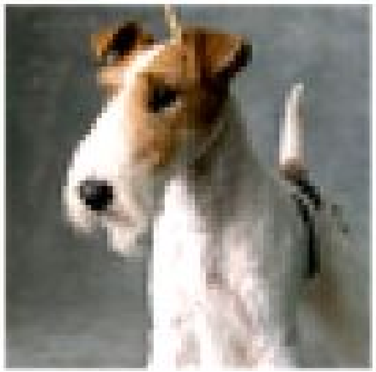
Dataset:

https://iiitaphyd-my.sharepoint.com/:u:/g/personal/pulkit_gera_research_iiit_ac_in/EfWpQsl6KedFod1C56u0pXsBRaWSC7KlCgbKbHmO1HZYkg?e=JVOB2s

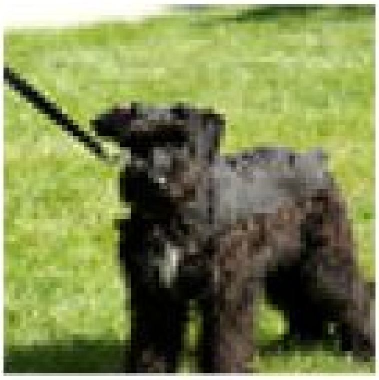
Mirror GAN

Results:

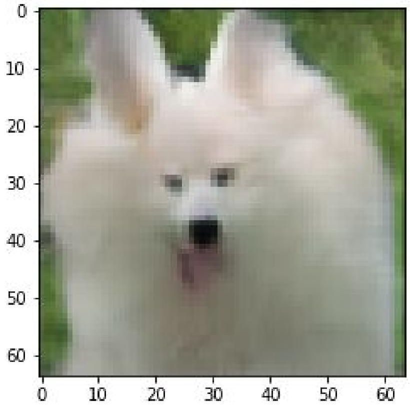
Text: Medium sized.white and brown coloured dog with a long snout and droopy ears



Text: Small black coated dog with drooping ears.



Text: White fur dog with upright ears and small eyes



These were the best images that we got. These are 64x64 and more training and data would help in generalizing the images.

Code : https://github.com/darthgera123/Mirror-GAN

Future Directions

As discussed before we observe that Transformers have been able to generalize well with the text embeddings. However we havent seen these embeddings being used as input to a C-GAN based architecture which is much better at generating images. We also observe how UNITER is capable of making mappings between images and words by having a shared embedding space. This embedding could be used for creating images as well which would generalize as well.