Minutes of Meeting

# Date: 21/06/2022

## Things presented

* Important links:
  + <https://keras.io/examples/nlp/multimodal_entailment/>
  + <https://github.com/google-research-datasets/recognizing-multimodal-entailment>
  + <https://multimodal-entailment.github.io>

page break

# Date: 14/06/2022

## Discussion

* Word2VisualVec has a pre-trained Word2VisualVec

# page break

# Date: 07/06/2022

## Discussion

* Word2VisualVec has a pre-trained Word2VisualVec

# page break

# Date: 31/05/2022

## Discussion

* The dissertation Image recommendation for Wikipedia articles pointed to articles like Predicting Visual Features from text for image and video caption retrieval describing the Word2VisualVec architecture (source code freely available)
* Other works which were explored and can act as good sources for material and ideas: Survey on deep multi modal data analytics: collaboration, rivalry and fusion (this paper was given a light read and might be worth reading again); Deep multimodal representation learning: a survey
* From the paper Word2VisualVec is defined as a deep neural network architecture that learns to predict a visual feature representation from textual input
* My understanding is that it predicts a visual feature representation to try and match with an image and hence it is used to get the most similar caption to an image as in rather than making use of a joint space they rely on the visual space
* Plethora of work in image to text, text to image, caption retrieval field; challenge: requires training which for the news domain is difficult to acquire (same challenges as with the GPT)
* Cosine similarity concept which was encountered in the paper about the history of image manipulation and also in the Word2VisualVec
* Word2VisualVec has a pre-trained Word2VisualVec
* Multi scale sentence vectorization including BoW, word2vec and RNN based text encodings
* Problematic to use externally trained models however, it does depend on how robust the pre-trained models are. Thinking abstractly about it, if we are given a glass of water to hold, we may not be able to gauge exactly what temperature it is at however, we can guess at whether it was kept inside say a refrigerator or it was just taken out of the microwave or it was just sitting on the table. This correlates to the idea of granularity.
* Even using a pre-trained model we may be able to get some information
* Progressing slowly in the right direction
* The basic question is an observational one: we want to know what rhetorical relations exist and what the distribution of those is in representative text categories.
* Need to understand how the Word2VisuaVec is calibrated and what it does and how it can be useful for us.
* The nature of th training data; are the 5 sentences related with the image textual description of the image or is there a space of image text relations represented in the data; it would be advantageous if in all cases the text is a literal description of the image because that would mean that we have a single image text cross media relation in the training data; and then we can use this to assess the degree to which other texts have the same relation to their image
* Template for dissertation: best to use the past dissertations from the school of computer science and statistics to think about the structure of the dissertation
* Use the format submitted last year

## Agenda for next week

* Get an understanding of what the Word2VisualVec system is capable of delivering on the basis of pre-trained models; using some creative thinking and the already existing set of functionalities we might be able to make good use of the same (however some functionalities may not be entirely useful)

page break

**Minutes of Meeting**

Date: April 12, 2022

**Discussions**

* Discussed the paper by Katerina Pastra
* The paper is a good source of resources
* There is a formal format for the dissertations  (find this out from friends or maybe look at the projects.scss.tcd.ie website)
* Can use existing Github implementations
* Start writing about the dissertation little by little
* As we implement and assemble the things then write them down
* While reading the papers, start citing the sources and adding them into the Bibtex because towards the end it will become a cumbersome task to gather whatever was references
* Over the next month, ceasing the dissertation meetings (at least till the second week of May)
* The problem mentioned in the conclusion of a previous paper read (the Katerina Pastra one) is a smaller fragment and our problem is bigger

**THIS WEEK AGENDA for discussion on April 19, 2022**

No agenda as the meeting will not be taking place, send a reminder that the meeting will not be taking place.  Will still be updating the dissertation folder every week so that there is a record of when the meetings occurred and when they did not



Date: April 05, 2022

**Discussions**

* Meeting was not held

**THIS WEEK AGENDA for discussion on April 12, 2022**

Since the meeting was not held, same agenda as last week: Read the PDF at <https://ieeexplore.ieee.org/abstract/document/8424823> and try out the different search engines .



Date: March 29, 2022

**Discussions**

* The paper read this week also deals with a collection of images, same problem as the previous paper where the manipulation history was being mapped using a collection of papers
* We are going back to the previous google reverse image search thing where we search an image on the google search engine, and we take the captions of the visually similar images which turn up (taking only the text so that can deal with the memory constraints) Try other types of engines like Bing also.
* Twitter may be a viable option as well however, the problem remains that on twitter, people post generic images of say food and tuss. However, if we get the posts from the account of say guardian or something, it may be helpful (however, the collected data will not be containing the captions of the images as Twitter does not have a concept od captiontning the images with long captions. However, if we were to consider the tweet text as the caption perhaps?? How the tweet text related to teh imahe used.. Have to give more thought to this .
* possible to infer from the stuff google reverse image search returns the different locations the image is used
* To see if one can access the text that accompanies images
* Focusing on papers cited by and which cited the COSMOROE paper

**THIS WEEK AGENDA for discussion on April 05, 2022**

Read the PDF at <https://ieeexplore.ieee.org/abstract/document/8424823> and try out the different search engines .



Date: March 22, 2022

**Discussions**

* Focussing on extracting salient features from images rather than trying to find their original source because that will require a lot of memory and computation

**THIS WEEK AGENDA for discussion on March 29, 2022**

Read the paper Novel Visual and Statistical Image Features for Microblogs News Verification and present findings



Date: March 15, 2022

**Discussions**

* Meeting did not take place



Date: March 08, 2022

**Discussions**

* The current paper in question (Internet Image Archaeology: Automatically Tracing the Manipulation History of Photographs on the Web) is indeed an interesting paper: if we have an image at hand, which hashes onto some stock image then that would be helpful to know; however the mechanics for this may will be very complex; maybe we can make use of the top few ranked images
* 2 important takeaways from the current paper: the composite and residual images concept for extraction of the salient features of an image and the justification for the search terms used which can be helpful for choosing search terms in our study
* Not necessary for every paper deemed relevant from a quick reading to have methodology useful or relevant to us, just exploration of previous work space to inspire our methodology
* With regards to the methodology, for some fixed set of the data it will be possible to curate or annotate by hand but the data we are considering is in a dynamic environment
* Stock images will have many more hits (large and high connected VMMs) than an actual image of an event
* The methodology currently in question is very much different from what was being discussed on March 01, 2022. In the previous discussion we were more focused on the context vector representation of the image
* The current methodology in question will be infeasible memory wise (requires memory capacity owned by the likes of Google, hence infeasible for our study)
* Used Latex and Bibtex. Have a bib file for the references and a separate file for the summary of the papers read

**THIS WEEK AGENDA for discussion on March 15, 2022**

No detailed agenda, try to find relevant papers in the fields of zero shot classification and tracing the origin of a picture on the world wide web. 

Date: March 01, 2022

**Discussions**

* The paper’s technique is better than what was found till now in my exploration
* Using the web-scale data, they obtained captions that are less restrictive than the supervised methods
* Their main approach is to make use of an existing image-text alignment score for example, cosine similarity to direct an existing large scale language model for example, GPT-2 towards a given image without training
* The work done is this paper could possible hint at not needing to utilize or look for image-text datasets on the web as was being done in our research previously: have to check whether zero-shot or supervised methods for image classification, which one is more accurate and which one solves the problem of bias. (They have noted that because of the type of the data used, there is biasness as the data contains biased sources)
* We have to try and see whether there is anything we could do to get rid of the bias which could be a novel contribution if not done already.
* They have presented a visual arithmetic capability: how can that be utilized in our project in any way?
  + Distributional semantic representation of word meaning
  + Rather that the representation we get for queen is similar to the representation of girl, again using cosine similarity
  + If we take inspiration from equations, imagine instead of + and -, more complex operations such as mean (central tendency); say we take a particular image and other similar images, we should be able to understand what is salient about the given image as compared to the average.
  + For example, say we have images of cows, we obtain an average image; subtracting a particular image from the average image should give us the salient features of that particular image (this is just one idea off the top of the head)
  + In relation to the captions, say if we have a bunch of photos of a particular politician and a given photo; by virtue of telling us what are the salient features of the given image and maybe what the relation is to the caption
  + Other suggestions:
    1. Another thing to be considered is the cross media version of the equation
    2. Open question: whether in the rooster and egg and cow and milk example, the computation was made using the text for cow’s milk
    3. Question about the distributional relation of the caption and the image
    4. If given an image and we subtract away the representation of the caption then if we get a zero value or close to zero value (in image terms this could be either a blank white image or a blank back image or random noise) then we could say that the caption and the image are closely related
    5. Cosine similarity of caption to image then they fit
       1. If equal to 1, they fit (assuming cosine similarity values from 0-1)
       2. If equal to 0, then irony relation
    6. These discussions are just abstract and based on intuition keeping empirical issues beside, what values of the cosine similarity measure depict the different relations
    7. What alternative operations and outcomes would correspond to alternative cross media relations: Join and meet could be the different available options for distributional representations
    8. Extend this work to product of image and text representation: exemplar relation or not: whether 2 things providing independent information or related information
    9. One can define more complex operations
    10. Given some natural operations do those correspond to cross media relations (like addition in arithmetic can be disjunction in logic; subtraction: negation)
    11. What are the natural interpretations of available operations? (there can be many uninterpretable examples and operations; all this discussion is based on the analytical interpretation and understanding, the empirical or application side of things may or may not work out)
* All-in-all, this paper is a worthy contender and has given many directions to pursue and explore. **Must be cited!!**
* The paper also directs us towards exploring whether zero-shot learning has been utilized in analyzing the cross media relations (which should be highly likely) field or not. If not, then it could be a novel contribution but chances of this are minute
* The paper also provides a link to their code and a ready model for trying out a few examples. Some of the examples which were tried out have been included below:

This image is doing rounds on Twitter being associated with the current war. However, the model is able to caption it fairly well

Graphical user interface, application

Description automatically generated

 This image is of Elon Musk apparently from 2021 in Germany attending an event which was used in this and many other news articles. However, we can see that the model can make the association of the person in the image to Tesla but does not mention Elon Musk anywhereGraphical user interface, application

Description automatically generated

The above 2 images were downloaded from the internet and thus, this means that it is highly likely that the model was trained on these images. Hence trying these images means trying images which were in the model’s training set. Now we try with an image which has never been uploaded to the internet which means that the model has never seen this image. The conditional text was changed for the image on the left and was unchanged for the image on the right. We can see that the left one makes no sense whereas the right one mentions ‘dog’ but eating a sandwich whereas there is no sandwich (the shoe is very close to the dog’s mouth which was probably assigned as sandwich, so the model is not that inaccurate either)

A picture containing dog, black, mammal

Description automatically generatedA dog looking at the camera

Description automatically generated with low confidence

We need to come up with a mechanism to classify the image as well as to trace its source and origin

There will be a lot of work in related field of fake news detection which should be utilized

The book Beyond media borders will be used as a source for some non-technical background

**THIS WEEK AGENDA for discussion on March 8, 2022**

Read the paper ‘Internet Image Archaeology: Automatically Tracing the Manipulation History of Photographs on the Web’ and present findings



Date: February 22, 2022

**Discussions**

* The proposed strategy seems logical
* There are different reading strategies
  + Superficial: When trying to understand the scope of the work been done and is broad
  + Reading which is somewhat more concentrated and guided in order to answer certain questions motivated by work carried out
* From the cross-media relations angle: image classification is just one constituent
* Our research question is sort of figured out
* CLIP can be useful where the image is an example of the text
* Exemplars, irrelevant images, projected images, fabricated images
* CLIP will help textualising what an image is about
* Example. If we have a recipe and it contains a picture of a cat then CLIP can help us identify that the picture is of a cat and then what remains to be done is to figure out the relationship of the cat image with the text of the recipe we found that image in. (The recipe may include chocolate and it might have some warning on not spilling chocolate where the cat could be attached with a caption such as: You dont want her to get a stomach-ache do you? This does not mention the cat anywhere and it serves to support the warning)
* Sometimes the masters projects are defined in a way that one needs to find the most efficient solution
* In our problem, we are not guaranteed a robust or efficient or even a complete solution
* The distribution of cross media relations is a hard problem
* CLIP can be useful however it does not solve the cross-media relations; it helps convert the problem from an image text relation to a text text relations problem
* Apart from using pygooglenews for articles from random sites, we want a level of hand tuning as well. For example, to include sites such as TheOnion which contain some level of satire so that we can understand whether there are differences in the cross-media relations depending upon the category of the news (misinformation, propaganda, satirical…. (maybe to include sites which are known to tabloid: probably the same as propaganda))
* We have to do what Darwin did, but from our desk
* Sometimes when trying to harvest data it would be wise to put in a lag between requests (good general practice)

**THIS WEEK AGENDA for discussion on March 1, 2022**

* Try running CLIP on pictures of celebrities whether it can actually identify them as the particular person or just outputs: picture of a person
* Read one paper and write a summary on it and present your findings and observations from the paper: this can help with getting some background of the work done in this field (have a proper documentation for the paper read and keep its summary in the same folder for better reference later on)
* Start with the papers already found

