Quick Summary of the Deep Document Understanding Workshop held in conjunction with ICFHR2022

MINUTES OF THE KEYNOTE PRESENTATIONS

1. **Dr. Filip Gralinski**

Dimensions of Document Understanding - Multidimensional Measurement of Progress in DIAR:

a. How do we define document understanding - Information extraction from legal documents (amounts, seller/buyer, etc). Usually interest lies only in the information and not where the information exist in the document.

b. How to tackle the problem of evaluation in a more principled way? c. DUE : A proposal to benchmark document understanding from an industry perspective.

d. Assumptions in Document Understanding:

i. Documents are by humans for humans. Documents act as API. ii. We should not limit humans' way of solving problems. Document Understanding challenges should take both into account.

iii. Pyramid of needs: Understanding <- Abstractive extraction <- Structure Recognition <- Text Recognition (OCR). Eg, in an invoice, an OCR mistake can be figured out by overall semantic understanding of the invoice. Hence, there are some loops in the pyramid. Abstractive extraction is therefore most important.

iv. Transformer limitation: Number of tokens/patches. Larger documents become difficult to represent. Example, newspapers with narrow columns and dense information.

v. How to get rid of metadata in documents that may make challenges trivial?

vi. How to handle multi-pages documents?

vii. Create a challenge where there are no biases in the collection of documents. Why do we follow fixed steps in documents collection?

viii. Temporal dimension imperative for processing historical documents. E.g., changes in laws/regulations, etc.

ix. Multilinguality - Brings in practical issues such as datasets, resources, etc, but not necessarily from an algorithmic point of view?

x. Other dimensions - domain, language, format, quality, printed/handwritten, etc.

e. Gold Approach for Document Understanding - Question Answering Format

i. Abstractive along with extractive.

ii. Add references both in inputs/outputs - in a more human way, not just boxes.

iii. How to measure model calibration?

iv. Avoid traps in QA.

f. Principled evaluation: Instead of simple matching between predicted and actual, we need to take into account factors such as probability of occurrence, confidence measures, etc.

g. How to get insights from document datasets and models instead of simply treating them as datasets.

2. Dr Seiichi Uchida

Challenges Beyond Recognition

a. Character recognition, Scene Text Detection and Recognition have become very accurate.

b. What to do with the perfectly recognized text from images?

c. Label (to disambiguate object), Message (verbalized information), Design (non-verbal impressions), Code (readability in presence of noise) d. Proposal of Total OCR - recognize text everywhere. On top of it, create applications such as education, enable visually disabled and so on. e. Discriminate between significant and trivial information. - A possible direction of research.

f. Understanding interaction between text and humans.

g. Understanding interactions between text and objects within an image. TextVQA is an example of an integrated task.

h. Understanding text and object co-occurrences.

i. Analysis of font styles. This may find applications in automatic poster creation.

How are the fonts used and why do we have font variations. Eg. book genres correlate with font styles and colors of titles.

j. Font impression analysis. How do shapes and impressions of font impressions correlate? Important parts for legibility? Why do font shapes and impressions correlate?

k. Generative models - Handwriting generation?

l. Affection related annotations in documents?

m. Document as a new area where a new learning paradigm may emerge? (other than supervised, unsupervised, etc.) Language + Vision for new learning paradigm specific for document understanding?

PANEL DISCUSSION

The second session was about the panel discussion. The goal is to figure out grand challenges that can drive the document understanding community at large. This has been happening in other sub-areas of AI such as object detection, instance segmentation in Computer Vision; Text Classification, summarization, etc. in Natural Language processing.

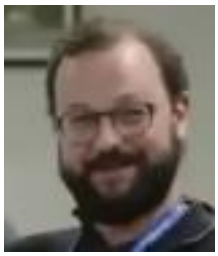
- **Prof Jawahar - Introduction:**

In ICDAR we see many challenges organized every year, however we have not seen a “grand challenge” for document understanding at large with the aim of setting up new goals for the community and moving forward. On the other hand, we observe much fewer challenges in other vision and NLP areas.

One reason may be that the size of datasets in document understanding space is much smaller than those in the parallel areas of research. As a part of this discussion, it was aimed to understand how to move forward - whether to have many subproblems or we have a larger generic problem to solve. Also, for the smaller sub-problems, how can we have synergies between the datasets and other resources. Do we need any changes in the very fundamental way in which the document research community has been moving forward as of today.

The panel more-or-less agreed on the fact that despite having many new smaller challenges every year, there hasn't been a single "grand challenge" as yet that could propel the research in document research space at large just as what ImageNet did to the Computer Vision community. We need to develop a challenge that could result in a leap in the methodology. While it has been observed 4-6 participants on the smaller datasets, there is not a single dataset as of now that has been of interest to the document community at large.

- **Speaker 1 (Andreas Fischer):**



Long-term competitions, those which have been hosted across years may not be "grand" as yet but it may be a step towards that direction. One perspective of looking at a grand challenge is to prepare a dataset that includes many different scripts, languages, domains etc, put together even in the absence of the ground truth. While most linguists might not agree given the intricacies of each script and language, that itself would be the idea of a grand challenge. In order to create that "grand dataset", it is imperative to go beyond the metadata of date, location, language, domain, etc. Back in 2010, Google Books project hit a roadblock which tried to digitize all the historical books. Maybe something from historical documents may be linked to the "grand challenge" given the enormity of the data.

- **Speaker 2 (Milan Šulc):**

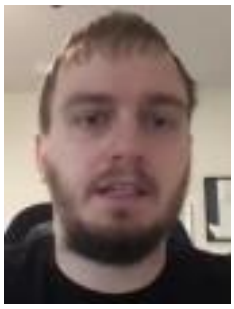
ImageNet is taken as a reference of grand challenge. Some aspects such as scale of the data, variety of possible less correlated tasks may make the notion of "grand challenge" difficult for the document community. The tasks generally in document information extraction are rather complex than "detection" or "classification". For a grand challenge, a task needs to be simple, accessible and generalized well enough such that results of some sub-tasks may correlate to results of some other

subtasks. In the sense that the “document embedding” could be shared between several document understanding subtasks. Many documents that are of interest to the industry are also of a sensitive nature that could not be publicly released to the academic community. Some techniques around maintaining privacy around this also need to be thought about. This makes it a challenge to get realistic data at a scale. Annotations in document understanding space are also a lot more expensive than simple classification annotations. Like ImageNet which has helped answer the question - “which architectures have better captured the image representations”, we need something like that in our space.

- **Speaker 3 (Srirangaraj Setlur):**

Defining the end task is not as simple. Annotations for layouts, tables, charts, etc are quite complex despite having an abundance of data. Amount of online lecture data can also be utilized for “education for all”. Annotating such data for scale is a big challenge even in a semi-automated manner. The fundamental requirement of a grand challenge is to choose an area where large scale data is easily accessible and then other issues on top of it may be worked towards to be addressed. Two such proposed areas are lecture videos and scientific articles.

- **Speaker 4 (Chris Tensmeyer):**



One of the most successful characteristics of ImageNet is that the task is universal, which is not the case in documents. Which image is this can be asked for any image irrespective of the domain. For documents, there are many specialized domains (historical, invoices, poems, etc) which bring about domain specific

datasets and tasks. We could ask “What type of document is this?” for a grand challenge, but that would require building up of taxonomy which again may be a point of arguments. Perhaps a broad category of attributes may be built up to take specific definitions into account. Even for document lay outing, taxonomy would be required for building solutions for domain-specific documents. There may be datasets where records may need to be extracted in a certain way, which could be hard to scale for a grand challenge. It may therefore be worthwhile to restrict to a single large domain such as scientific articles where attributes are relatively well defined. However, for a grand challenge, that may limit the reach. For a grand challenge, we need to figure out an annotation scheme that can be applied to all types of documents which may be hard to start with but that’s the direction which needs to be taken. Documents combine vision and linguistic modalities in a non-trivial manner which needs to reflect the grand challenge of document understanding.

- **Speaker 5 (Dimosthenis Karatzas):**

A grand challenge needs to be thought of very differently from ICDAR challenges which are solved for a very specific purpose. All competitions proposed once should be available for several years ahead. Multi-script and multi-linguality is also a challenge practically but may not be theoretically as pointed out by Dr. Philip. Tackling privacy concerns needs to be introduced into document communities which are already utilized in other areas. Further, we do not have an easy way to crawl for documents such that they are unbiased.

- **Speaker 6 (Manish Srivastava):**

“What is this document about?” need not have a taxonomy or a layout requirement. It may be answered with a simple description. E.g., it is a legal contract between

parties A and B and so on. The description can itself be at various levels. This may become a universal question regarding documents, though it is coming from an NLP background. Also, the dataset creation side of it would be very complex to say the least.

Document communities also need to collaborate more among themselves and not have competitions that divulge. There are many layers in documents as opposed to the primitive task of image classification which requires thorough collaboration.

Conclusion: What could be the problems for prospective PhD or other research students? Non-trivial to answer. A problem may be interesting either because it is hard to solve right now and further from application and other may be that the research community thinks is already solved but may be far away from applicability.