CCT College Dublin

Assessment Cover Page

|  |  |
| --- | --- |
| Module Title: | Advanced Data Analytics  Big Data Storage and Processing |
| Assessment Title: | Integrated CA1 Sem 2 MSc in Data Analytics |
| Lecturer Name: | David McQuaid  Muhammad Iqbal |
| Student Full Name: | Mohamed Noor |
| Student Number: | SBS22070 |
| Assessment Due Date: | 03/10/2022 |
| Date of Submission: | 02/10/2022 |

Declaration

By submitting this assessment, I confirm that I have read the CCT policy on Academic Misconduct and understand the implications of submitting work that is not my own or does not appropriately reference material taken from a third party or other source. I declare it to be my own work and that all material from third parties has been appropriately referenced. I further confirm that this work has not previously been submitted for assessment by myself or someone else in CCT College Dublin or any other higher education institution.

Image Recognition for Learning Arabic Words

Mohamed Noor   
*Information and Communications Technology (ICT) Faculty   
CCT College Dublin*Dubln, Ireland  
sbs22070@student.cct.ie

*Abstract*—IEEE format requires you to include an abstract at the start of your paper, followed by a list of keywords. In the “Styles” section in Microsoft Word, you can find the appropriate styles for all the different sections and headings in the paper, which are already applied here. For example, the “abstract” style is applied to this text, the “keywords” style to the next section. Note that the titles “Abstract” and “Keyword” should remain as they are written here: italicized and followed by an em dash.

Keywords—image classification, Arabic, big data

# INTRODUCTION

Level 1 and 2 headings (as well as the paper title) should be written with title case capitalization, while level 3 and 4 headings are written in sentence case.

With more than 7,000 languages currently spoken [1], their study can be complex and biased by researchers’ preferences and the availability of funding. In general, languages can be grouped into families, of which there are more than 100 [2]. Arabic and English are spoken by about 360-400 million native speakers [3], [4] and are official United Nations languages [5]. However, English is a *lingua franca* through the prevalent inter-connected global economic structure [6]. Arabic is a part of the Afro-Asiatic family [7] while English is an Indo-European language [8], yet historical contact through a series of commerce and wars [9] has allowed for an intermingling (borrowing) of words. Arabic words, such as ‘sine’ and ‘alkali’ have penetrated English while words such as بلاستيك (‘bi-laastik’, plastic) have an English origin [10].

Unlike English, Arabic is an unusual language. For example, there are 200 words that refer to lion in Arabic [11]. The definite article ‘the’ in Arabic (‘al’, ال) never stands alone. Yet, when joined with another word will change its morphology (for example, lion (أسد) will become الأسد). There are also regional variations, which is why learners of Arabic usually tend to follow the Modern Standard Arabic [12].

Given the complexity of Arabic compared to English, it would be helpful for learners of Arabic, especially those who are not native, to have the ability to instantaneously convert images to Arabic words. The ubiquitous smartphones can play a significant role in helping these learners by snapping images of new objects that can be classified rapidly in Arabic.

Image recognition through deep learning is a particularly useful method for classifying images. Two types of solutions are relevant in the context of helping Arabic learners – (i) multi-class algorithms where each image is assumed to only take one label and the entire dataset consists of multiple labels, or (ii) a multi-label problem where each image can take multiple labels. Taking an image of a woman walking with a dog as an example, the image can be labelled either as a woman or a dog, but not both (multi-class), or as a woman and a dog (multi-label).

## Convolutional neural network (CNN)

CNN is xxx. Unlike recurrent neural networks, CNN is highly parallelisable, making training much faster with graphical processing units (GPUs).

Image recognition is a well-studied domain of artificial intelligence with wide-ranging applications in real life. As such, complex pre-trained models have been published and are easily accessible within the Tensorflow ecosystem. The use of pre-trained models hinges on the ability to use deep learning models across different use cases. This ability, termed transfer learning, is XXX.

Unlike traditional machine learning, deep learning does not require extensive feature engineering. Neural networks also lends itself to modular and complex architectures, including the ability for branching (multi-headed) training that can take advantage of distributed training (model parallelism).

<< table of models + number of parameters + references >>

<< discuss CNN, pre-trained models, transfer learning, CNN architectures: Inception, Xception etc (essentially a lit review of models) >>

## Big image data

Success of deep learning depends on the availability of large datasets. As shown in Table XX above, these architectures consist of millions of trainable parameters, necessitating the use of large training datasets. Fortuitously, public image datasets are available, allowing for like-for-like comparison of various image classification algorithms.

ImageNet is XXX.

<< some description of large image datasets, with a focus on ImageNet and MMID >>

<< image sources: Rationale for selecting the MMID – why this dataset and not the others???, ImageNet … >>

MMID reference paper: [14]

## Study objectives

Learners of new languages are often advised to use flashcards that has an image of an object on one side and the word on the other [13]. To avoid confusion, multi-class algorithms may be of more benefit than multi-label for learners.

In this study, XXX

As a proof-of-concept, this paper discusses the use of pre-trained image classification algorithms for identifying images in Arabic. Pre-trained algorithms take advantage of the transfer learning concept in deep learning, where a model trained on a problem can be used in another problem if these are relatively similar. For instance, a model training on recognizing XXX can be used for YYY (give examples peer-reviewed).

## This Is a Level 2 Heading

### And this is a level 3 heading: Equations should be typed in either Times New Roman or Symbol font, or, if the equation is multileveled, inserted into your text as a graphic instead. On the far right of the line containing the equation, number it in parentheses, and use this number to refer to it in the text (1).

*a**b* 

### This is another level 3 heading: Lorem ipsum…

| Table Head | Table Column Head | | |
| --- | --- | --- | --- |
| Table column subhead | Subhead | Subhead |
| text | Texta |  |  |

#### And this is a level 4 heading: Make sure that the

1. This is a figure caption. It appears directly underneath the figure.

appropriate style is still applied to each section, reapplying styles if necessary.

#### This is another level 4 heading: It’s also possible to add bullet points when appropriate, using the “bullet list” style:

* Treat the word “data” as plural, not singular.
* For example, “the data indicate that …”

## This Is Another Level 2 Heading

Place any figures or tables you use at the top or bottom of a column. Don’t place them in the middle of a column. If particularly wide, a table or figure can span across both columns. Insert a table or figure after the point where it is first cited in the text.

When inserting a figure, such as a photograph or infographic, use 8 pt. Times New Roman for any labeling text within the image and for the figure caption. You can see an example of a figure caption in Fig. 1, above. Refer to figures like that, using the abbreviation “Fig.” and the figure’s number.

A table heading (using the “table head” style) appears above a table. This will automatically number the table for you. Any footnotes appear below the table, using the “table footnote” style. Footnotes are indicated by superscript lowercase letters within the table. An example of a table can be seen in Table I, below.

# RESULTS

All the headings in the main body of your paper are numbered (automatically).

## Image selection

For this project, Python XX was used together with YYY.

## Programming environment

Python 3.9.12 was used for data exploration and deep learning. Given the size of the data, PySpark 3.3.0 was used. Other relevant packages are TensorFlow 2.10.0 (cite??), Pandas 1.5.0, Numpy 1.23.3 and Matplotlib 3.6.0.

An interactive Python application was created using Visual Studio Code as an IDE through the Dash 2.6.2 and Plotly 5.10.0 packages.

<< ethics + licensing: use of people’s images 🡪 is this allowed?, did MMID only use copyright-free images?? >>

## Exploratory data analysis

Images from public sources can be expected to vary in their dimensions. As CNN models can only deal with one input size, it is important to understand the distribution of image heights and widths. Some images can also be represented by symbols. For example, ‘+’ images 🡪 give some examples.

<< plots of image dimensions >>

Table . XXX IS this the right format for an IEEE table?

|  |  |
| --- | --- |
| **Symbol** | **Example images** |
| $ | A picture containing text, clipart  Description automatically generatedA golden trophy with a white background  Description automatically generated with low confidenceA picture containing diagram  Description automatically generated |
| + | Icon  Description automatically generatedIcon  Description automatically generated |
| = | Graphical user interface  Description automatically generatedTable  Description automatically generatedText  Description automatically generated with medium confidenceGraphical user interface, application, table  Description automatically generated |
| ᵒ | Qr code  Description automatically generatedA picture containing tool, brush  Description automatically generated |

<<example of images: grid of 3x3>>

<< example of images of symbols >>

1. This Is the Heading for a Table
2. This is a table footnote.

Lorem ipsum….

##### Acknowledgments

This study used data from MMID ….

As you can see, the formatting ensures that the text ends in two equal-sized columns rather than only displaying one column on the last page.

This template was adapted from those provided by the IEEE on their own website.

##### References

[1] Eberhard, David M, Gary F. Simons, and C. D. Fennig, Eds., *Ethnologue: Languages of the World. Online version*, Twenty-Fifth edition. Dallas, Texas: SIL International, 2022. [Online]. Available: http://www.ethnologue.com.

[2] S. Malik-Moraleda *et al.*, “An investigation across 45 languages and 12 language families reveals a universal language network,” *Nat. Neurosci.*, vol. 25, no. 8, Art. no. 8, Aug. 2022, doi: 10.1038/s41593-022-01114-5.

[3] A. S. Kaye, “Arabic,” in *The World’s Major Languages*, 3rd ed., Routledge, 2018.

[4] “English Speaking Countries List | Lingoda Online English Language School,” *Lingoda*. https://www.lingoda.com/en/content/english-speaking-countries/ (accessed Oct. 01, 2022).

[5] U. Nations, “Official Languages,” *United Nations*. https://www.un.org/en/our-work/official-languages (accessed Oct. 01, 2022).

[6] B. Seidlhofer, “Language Variation and Change: The Case of English as a Lingua Franca,” in *English Pronunciation Models: A Changing Scene*, 2nd edition., K. Dziubalska-Kolaczyk and J. Przedlacka, Eds. Verlag Peter Lang, 2008.

[7] A. Howell *et al.*, “(No) variation in the grammar of alternatives,” *Linguist. Var.*, vol. 22, no. 1, pp. 1–77, Jan. 2022, doi: 10.1075/lv.19010.how.

[8] R. Guarasci, S. Silvestri, G. De Pietro, H. Fujita, and M. Esposito, “BERT syntactic transfer: A computational experiment on Italian, French and English languages,” *Comput. Speech Lang.*, vol. 71, p. 101261, Jan. 2022, doi: 10.1016/j.csl.2021.101261.

[9] A. Hashim, G. Leitner, and M. A. Aqad, “Arabic in contact with English in Asia: Linguistic, social and political influences of Arabic in the region,” *Engl. Today*, vol. 33, no. 1, pp. 25–32, Mar. 2017, doi: 10.1017/S0266078416000377.

[10] Y. Bader, “Semantic change in Arabic loanwords from English and French,” *Abhath Al-Yarmouk*, vol. 8, no. 2, pp. 33–48, 1998.

[11] P. Zemánek, “Two Hundred Ways to Call a Lion in Arabic: Names or Epithets?,” *Z. Dtsch. Morgenländischen Ges.*, vol. 171, no. 2, pp. 343–374, 2021, doi: 10.13173/zeitdeutmorggese.171.2.0343.

[12] A. Mahmoud, “Modern Standard Arabic vs. Non-Standard Arabic: Where Do Arab Students of EFL Transfer From?,” *Lang. Cult. Curric.*, vol. 13, no. 2, pp. 126–136, Jul. 2000, doi: 10.1080/07908310008666594.

[13] S. Rosenbloom, “Want to Learn French? Italian? Russian? There’s No Time Like the Present,” *The New York Times*, Apr. 28, 2020. Accessed: Oct. 01, 2022. [Online]. Available: https://www.nytimes.com/2020/04/28/travel/language-instruction-apps-television-youtube.html

[14] J. Hewitt, D. Ippolito, B. Callahan, R. Kriz, D. T. Wijaya, and C. Callison-Burch, “Learning Translations via Images with a Massively Multilingual Image Dataset,” in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Melbourne, Australia, Jul. 2018, pp. 2566–2576. doi: 10.18653/v1/P18-1239.