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: | 03/10/2022 |

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*Dublin, Ireland  
sbs22070@student.cct.ie

*Abstract*—Unlike English, Arabic is a complex language with the morphology of letters changes upon connecting with other letters. The diacritics are also present in modern typography, although they may be absent in certain contexts or in old literature. Image recognition through deep neural networks presents an opportunity for learners of the Arabic language to identify new objects by simply uploading an image to a Python application. In this study, a simple and several complex pre-trained networks were used to classify images from the Massively Multilingual Image Dataset (MMID). Exploratory data analysis on such a big dataset was undertaken by using distributed computing. Preliminary screening was performed to identify the most promising models, followed by a longer training on MobileNetV3Large and VGG19. Although the accuracy was limited, the resulting model could classify the broad category of images as judged by a human annotator. The MMID dataset has a disadvantage of being biased by the Google Image Search algorithm, the working of which is not fully understood. In addition, the complexity of labels should be reduced by only focusing on nouns and ‘bigger picture’ context, instead of specific locations or groups of individuals.

Keywords—image classification, Arabic, big data

# INTRODUCTION

With more than 7,000 languages currently spoken (Eberhard, David M et al., 2022), 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 (Malik-Moraleda et al., 2022). Arabic and English are spoken by about 360-400 million native speakers (“English Speaking Countries List | Lingoda Online English Language School,” n.d.; Kaye, 2018) and are official United Nations languages (Nations, n.d.). However, English is a *lingua franca* through the prevalent inter-connected global economic structure (Seidlhofer, 2008). Arabic is a part of the Afro-Asiatic family (Howell et al., 2022) while English is an Indo-European language (Guarasci et al., 2022), yet historical contact through a series of commerce and wars (Hashim et al., 2017) 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 (Bader, 1998).

Unlike English, Arabic is an unusual language. For example, there are 200 words that refer to lion in Arabic (Zemánek, 2021). 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 (Mahmoud, 2000). 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 a neural network that consists of one or more convolutional layer, non-linearity layer, pooling layer and finally fully-connected layer (Albawi et al., 2017). As the relative position of an object in an image is less relevant, the convolution layer allows for the extraction of features irrespective of object orientation. In fact, small datasets may be augmented by artificial shearing and rotation. Unlike recurrent neural networks, CNN is highly parallelisable especially when graphical processing units (GPUs) are used.

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 (Sanchez et al., 2020). The use of pre-trained models hinges on the ability to adapt deep learning models across different use cases (You et al., 2021). This ability, termed transfer learning, is based on how humans learn to identify new objects. A model trained on pictures of animals for a binary classification (‘is it an animal or not?’) can be used to classify the type of animal (‘what animal is it?’). Transfer learning will not be appropriate if the training dataset is completely different than the original dataset. When the datasets are similar, however, the use of pre-trained models can result in highly accurate models at a fraction of time needed for full training.

Another advantage of deep learning is that it requires no feature engineering. In the context of image classification, no annotation will be required to identify the legs or ears of an animal. This becomes useful when human cannot pre-identify features of high relevance to learning. Neural networks also lend itself to modular and complex architectures, including the ability for branching (multi-headed) training that can take advantage of distributed training (model parallelism). In this mode, one GPU may be used for one branch while another can be simultaneously used for the other branch, further reducing the total training time.

Several architectures of deep neural networks have been proposed for the use of image classification as summarised in TABLE 1. These include various MobileNet architectures (Howard et al., 2017) that originate from Google researchers, including MobileNetV2 (Sandler et al., 2019) and MobileNetV3 Large (Howard et al., 2019). These models are targeted for mobile and embedded devices where the computational power is more limited compared to personal computers. Although deeper and complex networks may achieve higher accuracy, they also require longer time and power for training. Through MobileNet architectures, the researchers aimed to create more efficient networks by using depthwise separable convolutions, which themselves are a combination of 3x3 depthwise convolution and 1x1 (pointwise) convolution (Howard et al., 2017). MobileNetV2 builds on the original MobileNet work by including an inverted residual structure with shortcut connections between thin bottleneck layers (Sandler et al., 2019). Overall, the use of depthwise separable convolution significantly reduces computation – for instance, MobileNetV2 requires about 8 to 9 times less power than standard convolution (Howard et al., 2017). The bottleneck layers accept a low-dimensional compressed representation, which is then expanded to a high dimension and filtered with depthwise convolution (Fig. 1).

In deep learning, it can be difficult to judge whether one architecture will be better than the other. The use of network architecture search (NAS) is essentially ‘a neural network to design a neural network’ as it automated the manual search for the optimal architecture (Elsken et al., 2019). MobileNetV3-Large is a product of NAS and network design, with the resulting increase of 3.2 % in accuracy on ImageNet classification with 20 % reduction of latency compared to MobileNetV2 (Howard et al., 2019).

Diagram

Description automatically generated

Fig. 1. Evolution of MobileNet networks through the use of separarable and bottleneck layers. Diagonal hatches indicate linear layers. The final (lightly coloured) layer is also the beginning of the next block. Upon stacking, the blocks in (c) and (d) are equivalent. Figure taken from (Sandler et al., 2019).

Another family of relation architectures is the Inception network, including InceptionV3 (Szegedy et al., 2015) and InceptionResNetV2 (Szegedy et al., 2016). This family is also aimed at simplifying network structure for its use in mobile and other low performant devices. The Inception networks arose from architectural considerations around convolutional layers as well as balancing of the network width and depth. In general, convolutions with spatial filters larger than 3x3 tend to be more computationally expensive. Nevertheless, they can capture long-range dependencies (Szegedy et al., 2015). As a workaround, a multi-layer network with 3x3 convolutions was proposed with a resulting network depth of 42 layers for InceptionV2. InceptionResNetV2 extends the thinking further by incorporating residual connections, which was first introduced in (He et al., 2015) to reduce training time.

The NasNetMobile architecture (Zoph et al., 2018) also uses NAS to design the best model where accuracy was increased by 1.2 % with a concomitant reduction in computation of 28 %. The NAS search space was limited to only the best cell structure instead of re-designing the entire architecture. Further limits were also placed on activation (ReLU only) and no batch normalisation/ReLU for depthwise separable convolution. The VGG19 (Visual Geometry Group 19 weight layers) (Simonyan and Zisserman, 2015) is one of the two models that take advantage of depth as the main driver for image recognition algorithms through the use of very small (3x3) spatial filters. In contrast, the Xception network (Chollet, 2017) was propositioned as an interpretation of the Inception module, forming an intermediate between regular convolution and depthwise separable convolution. Both InceptionV3 and Xception have the same number of parameters but performance gains in the latter were achieved through a better use of parameters.

|  |  |  |
| --- | --- | --- |
| **Model** | **Number of parameters** | **Commentary** |
| InceptionV3 | 23.9M | Captures long-range relationships through deep small spatial filter layers, instead of large spatial filters. |
| InceptionResNetV2 | 55.9M | Incorporates residual connections to reduce training time. |
| MobileNetV2 | 3.5M | Adds bottleneck layers with separable layers. |
| MobileNetV3-Large | 1.4M | Used NAS to arrive at the optimal architecture. |
| NasNetMobile | 5.3M | Used optimised design space for NAS. |
| VGG19 | 143.7M | Deep CNN with 19 layers. |
| Xception | 22.9M | Improvement over InceptionV3. |

TABLE I.

SUMMARY OF VARIOUS NEURAL NETWORK ARCHITECTURES FOR IMAGE CLASSIFICATION.

## Big image data

Success of deep learning research depends on the availability of large datasets. As shown in TABLE 1 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. A popular dataset often used as the standard for benchmarking is ImageNet, a database of more than 14 million images organised according to the WordNet hierarchy (Deng et al., 2009; Shanmugamani, 2018). The hierarchy organises over 150,000 words into 117,000 synsets (synonym sets/categories), thereby establishing ontological and lexical relationships for natural language processing tasks (Miller, 1995). For instance, German shepherd is organised under ‘dog’, which in turn is under ‘mammal’, potentially allowing for the same images to be used for different model training depending on the requirements (‘cats *vs*. dogs’, or ‘what breed of dog is this?’) More than 20,000 synsets have been indexed in ImageNet. To help the research community, a subset of the dataset has been human annotated using Amazon Mechanical Turk (Deng et al., 2009), which also forms the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) dataset (Russakovsky et al., 2015).

In contrast to ImageNet, the Massively Multilingual Image Dataset (MMID) consists of 98 languages and 10,000 words per language (Hewitt et al., 2018). The images were collected using Google Image Search, which has a major disadvantage of being prone towards highly optimised webpages. Search engine optimisation is a digital marketing strategy that could lead to bias in the type of images being included in the dataset, although the authors performed a filtering step. Unlike ImageNet, the words are not limited to nouns although nouns and adjectives were found to be translated with significantly higher accuracy than verbs and adverbs (Hewitt et al., 2018). For example, searching for coffee could result in images of cups and mugs and not necessarily coffee powder. MMID is also not hand-annotated, which could represent a more realistic opportunity for automated learning by algorithms. Searching for images in a target language may result in images that are biased towards the demographics of native speakers. By default, Google Image Search filters inappropriate search results, which can remove the ability of researchers to use the dataset for an automated image flagging algorithm.

As will be demonstrated in later sections, the Arabic language is highly complex and requires an understanding of grammar. The MMID dataset is significantly less researched than ImageNet with a resulting 33 and 43,168 citations on Google Scholar, respectively, and the former is also not hand annotated. Due to these factors, this study will be focused on using pre-trained models for transfer learning from ImageNet English words to MMID Arabic words.

## Study objectives

Learners of new languages are often advised to use flashcards that have an image of an object on one side and the word on the other (Rosenbloom, 2020). Two types of image recognition from a given image are the multi-label and multi-class classification problems. For the specific purpose of learning new languages, a multi-class algorithm will be of more benefit than a multi-label algorithm as each image will be associated with only a single text. With the Arabic dataset being relatively large, big data tools and techniques will be used and the resulting model will be deployed as an interactive user-friendly application.

# RESULTS

## 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, 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, and the Dash 2.6.2 and Plotly 5.10.0 packages. GitHub was used as a source control service, and the repository is available at <https://github.com/mnoor-ds/Sem2_CA1_Data>. For project management, the CRISP-DM framework was used beginning from exploratory data analysis up to final model deployment.

## Image selection

The MMID datasets consist of 100 images per word per language. Downloading the entire dataset will be time and storage consuming. Therefore, the Amazon Web Services (AWS) CLI v2 was used to browse through the dataset as stored in a Simple Storage Service (S3) bucket. Specifically, only the ‘scale-arabic-package.tgz’ file with a size of about 16 GiB was downloaded. Upon extraction, the entire dataset size was about 20 GiB, which precluded loading it into memory at once unless a high-specification machine is available.

## Exploratory data analysis

Images from public sources can be expected to vary in their dimensions, quality, colour (black-and-white, grayscale or full colour). In contrast, CNN models can only accept one input size. This restriction becomes even more critical when using pre-trained models, which often work with square images. Exploratory data analysis (EDA) was performed to understand the distribution of image dimensions, as well as to ensure consistency in data folder structure and readability of images for downstream purposes.

A picture containing logo

Description automatically generated

Fig. 2. A sample of images present in the MMID Arabic dataset. Each panel has been labelled with the Arabic word, which can be literally translated to ‘circles’, ‘the cloud’, ‘annual’, ‘style’ or ‘type’, ‘and the North’, ‘in Latin’, ‘ar-Rusafi’ (an Iraqi poet), ‘food’ and ‘books’ (left to right, top to bottom). Note that the ‘in Latin’ panel shows an image of Roman and Hindu-Arabic numerals.

A cursory EDA (Fig. 2) shows the presence of a mixture of image types, including illustrations, text-like images, and photographs. Moreover, several folders referring to symbols were found as shown in Table II. As stated earlier, the Arabic language, especially the written form, can be complex. Table III shows how the addition of vowels and diacritics can alter the meaning of the root word ‘*ab’* (father). Coupled with borrowed non-Arabic words, such as ABBA (the Swedish pop-rock group), the complexity increases even further. In the first panel, the words ‘*ab’* and ‘*ibb’* share the exact same letter but each has a diacritic ‘*fathah’* and ‘*kasrah’* representing the English-equivalent vowels ‘*a*’ and ‘*i*'. In addition, the word ‘*ibb’* also has a ‘*shadda’*, which places stress on the second letter. Relatively minor modifications to the root word can also alter the meaning from a generic ‘father’ to a specific ‘to father’ and ‘his father’ as shown in the second and bottom panels. Incidentally, there is also a subtle bias towards Arabic-looking male images.

Together, these image subsets provide a glimpse into the complexity of using deep learning for natural language processing and image processing tasks, particularly when no context is given to a data scientist who will be training a model.

|  |  |
| --- | --- |
| **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 |

TABLE II

EXAMPLE IMAGES THAT ARE LABELLED AS SYMBOLS INSTEAD OF WORDS.

The sample images discussed so far clearly have different dimension, albeit mostly square-like. To quantitatively understand dimension distribution, PySpark was used together with simple Python wrapper functions to create a list of image heights and widths (refer to Section 2.3 in the accompanying Jupyter notebook). This list was then used to obtain a graphical representation of image dimensions. As shown in Fig. 3 and Fig. 4, there are heavy tail distributions for both image heights and widths. About 12 % of images have a height of 192 pixels, 9 % with 144 pixels and 8 % with 256 pixels. In contrast, about 95 % of images have a width of 256 pixels and other width values represents less than 1 % each (Section 2.3.2 in the Jupyter notebook). In terms of image proportions, about 12 % has a 4:3 ratio, 9 % has a 16:9 ratio and another 9 % are square. Therefore, about 70 % of images do not have a standard square, 4:3 or 16:9 ratios, although the latter two are popular standards for television (Fischer, 2020).

|  |  |
| --- | --- |
| **Word** | **Sample pictures** |
| Labelled: أب  Noun: father (‘ab’)  Could also refer to a proper noun: إِبّ (‘ibb’, a Yemeni city) | A picture containing outdoor, grass, nature, mountain  Description automatically generated A close-up of some water  Description automatically generated with medium confidenceA picture containing mountain, sky, outdoor, nature  Description automatically generatedA black and white drawing of two people  Description automatically generated with low confidenceA picture containing mountain, nature, outdoor, ravine  Description automatically generated |
| Labelled: أبا  Verb: to father (‘a-baa’)  Accusative: ‘father of’  Could also refer to a proper noun: أبّأ (‘ab-ba’, the Swedish pop-rock group ABBA) | A group of people posing for a photo  Description automatically generated A picture containing person, sky, person, outdoor  Description automatically generated A person and person posing for a picture  Description automatically generated with medium confidenceA group of people posing for a photo  Description automatically generated A person wearing a head scarf  Description automatically generated with medium confidence |
| Labelled: أباه ()  Genitive noun: his father (‘a-baa-hu’) | A picture containing text, indoor  Description automatically generatedA person in a suit and tie  Description automatically generated with medium confidenceA person sleeping with a baby  Description automatically generated with medium confidenceA person wearing glasses  Description automatically generated with low confidenceA picture containing text  Description automatically generated |

*TABLE III*

A sample of similar words in Arabic with different labels present in the MMID Arabic dataset. Potentially offensive images have not been included.

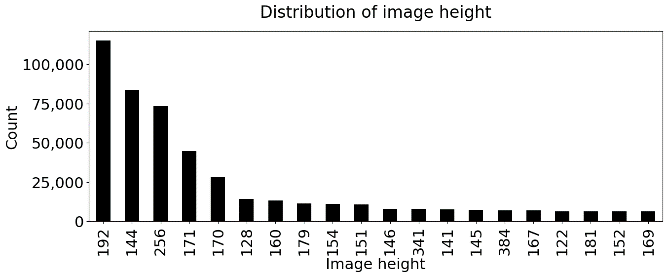


Fig. 3. Distribution of image heights in the MMID Arabic dataset. Note that the distribution appears to be Pareto-like with most images having 192 pixels in height. Only the first 20 height categories are shown.

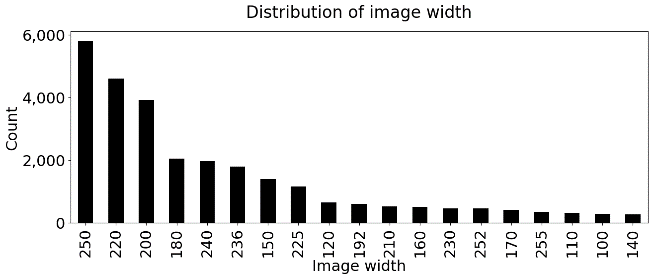


Fig. 4. Distribution of image width, excluding the single largest value of 256 pixels. This value was removed as it disproportionately represents more than 95 % of the images.

## Training of CNN models

As discussed in the Introduction section, various architectures for image recognition have been invented. The benchmark ImageNet dataset is often used to rank the quality of these models. However, the ranking may not reflect complex data structures, such as the one in the MMID Arabic dataset where there are more than 9000 classes. Accordingly, in this study, a set of models were selected based on their depth and number of trainable parameters.

Overfitting, a case where a model memorises an image class, instead of learning the features relevant to the class, is a common problem. To overcome this, validation must be incorporated. As was seen at the data pre-processing stage, it is impossible to fit the entire dataset into memory. Therefore, TensorFlow image generator was used to load a small number of images into the model, one batch at a time while simultaneously reducing the time taken to read data from disk. Through appropriate structure of directory where training, validation and test dataset are created, the image generator can be used during training to ensure overfitting does not occur. In this study, 70 % of images for each label was randomly assigned to the training set and the rest were split into validation and test sets. Random assignment, instead of using the first 70 % of images, was judged to the important as there is a significant effort in search engine optimisation done by website developers that can affect Google Image Search results. Moreover, the Google algorithm might also prioritise images that are relevant to the Arabic culture, such as Arabic facial features, over those from other cultures. Yet, a cross-cultural dataset will be required if an application is to be developed to allow new Arabic learners to understand the language.

A preliminary selection of a subset of promising pre-trained models was performed by training models with a short epoch of 3. As a baseline, a simple and relatively shallow CNN model was also trained. Nonetheless, the simple model did not result in effective learning (Fig. 5) and other well-characterised pre-trained models (MobileNetV2, NASNetMobile, MobileNetV3Large, InceptionV3, InceptionResNetV2, VGG19 and Xception) were included. The Mobile networks were targeted for devices with low computing power with the resulting small number of parameters of up to 5.3 million. In contrast, the Inception and Xception networks have about 25-56 million parameters. The very deep VGG19 is even more complex with 143 million parameters. This wide-ranging variation in depth and complexity was hypothesis to provide an insight into the most effective architecture for the MMID Arabic dataset.

MobileNetV3Large and VGG19 had the largest percentage increase in their validation accuracy (Fig. 5). In terms of validation loss improvement, only VGG19 showed a decrease. Given these metrics, these two models were then selected for further training with modifications to the model structure and epoch numbers. An early stopping callback was also introduced. Further experiments were also performed by removing pre-trained weights. Fig. 6 shows that pre-trained weights are of significant hindrance to the learning process. For instance, the VGG19 training loss increased by 20 % when using the pre-trained weights whereas the loss increase was only 10 % when training from scratch. Although loss increase is undesirable, it conflicts with the increase in the validation accuracy of 15 % and a decrease of about 2 % for pre-trained vs. from-scratch training, respectively.

Taken into consideration the actual validation accuracies of only about 3 % for MobileNetV3Large and 2 % for VGG19, there is a significant scope for improvement in model development. For MobileNetV3Large, there is a large gap of about 17 % between training and validation accuracy, whereby the training accuracy is about 20 % compared to only 3 % for validation. This large gap indicates overfitting, which could be reduced by adding dropout layers. For VGG19, the training-validation gap is much smaller at about 3-4 % despite it being about 100 times larger in terms of the number of trainable parameters.

Chart, waterfall chart

Description automatically generated

Fig. 5. Change in the validation accuracy and loss of simple and pre-trained CNN models during preliminary screening.

## Model deployment

As a proof of concept of model deployment after training, the fully trained MobileNetV3Large was saved onto disk. This model was then loaded into a Dash-based application, which provides a Pythonic environment to create a user-friendly interface. The goal of this deployment is to enable non-technical users to upload one or more images and use the trained model to classify those images. Dash is also amenable for production use, such as a webpage that can be accessed from anywhere in the world. The entire application hinges on the use of callbacks, which are initiated upon image upload followed by inference using the saved model. Images are resized and rescaled during inference, as the appropriate layers were added to the model during model building. This strategy reduces the amount of code needed for model deployment, especially if the application were to be built by non-data scientist or a web developer. A screenshot of the web application is shown in Fig. 6 whereby two sample images have largely accurate labels in Arabic.

Graphical user interface, application, website

Description automatically generated

Fig. 6. Screenshot of the deployment using Dash for a web-based application. Note the resulting Arabic label that represent the words ‘castles’ and ‘civilisation’, respectively, for the top and bottom images.

# DISCUSSION AND FUTURE WORK

New architectures are often based on a standard benchmark dataset. In the case of image recognition, the ImageNet dataset that has been hand annotated becomes useful for developing architectures that are geared for accuracy. In some cases, minor decreases in accuracy may be tolerated as long as the latency is improved significantly as is the case with mobile and embedded devices.

In this study, a framework has been established for the use of relatively simple and more complex deep neural networks for classify images in Arabic. Although the validation accuracy is not of production quality, it opens further opportunities through a more systematic evaluation of optimisers, learning rates and batch sizes. More importantly, the quality of the dataset will need to be improved. Instead of relying on the labels given, manual annotation will be required while taking into consideration the complexity of the Arabic language.

Several key steps that can be taken to improve the model performance are outlined as follows. Firstly, the complexity of the Arabic language must be reduced for an effective deep learning. This can be achieved by first removing symbol-like labels (‘+’, ‘=’, and so on). The reasoning behind this is that new Arabic learners are unlikely to use an application to learn new symbols as they are almost universally understood. Secondly, images must be labelled in the ‘bigger picture'. For example, a picture labelled as a particular rock group should be re-labelled as ‘singers’, ‘female/male artists’ and ‘musicians’. Similarly, photographs of a landscape or a tourist attraction should be labelled as ‘beach’, ‘hill’ or ‘forest’ instead of specific locations. A hierarchy must be established whereby only root words are considered, instead of relying on grammatical rules. Once these basic actions have been undertaken, human annotations may be used to complement the labelling efforts.

Such a clean dataset will reduce the complexity of the dataset and can be semi-automated, especially given that the proper nouns in Arabic have the ‘al-‘ prefix. In deep learning, it can be difficult to categorically determine if one algorithm is better than another without experimentation. Hence, both relatively shallow and deep, complex networks must be evaluated in tandem. A more automated approach could also be taken by using NAS, as implemented in the Auto-Keras library (Jin et al., 2019), where the number of layers can be tuned along with the use of non-traditional separable convolution. Finally, model performance might also be improved by image augmentation, where images are artificially flipped and/or sheared to increase the number of images per label. Together, the use of a robust dataset with a variety of model architectures can improve the final model accuracy.

Chart, waterfall chart

Description automatically generated

Fig. 7. Change in the validation accuracy and loss of refined pre-trained models. The ‘unfrozen’ tags refer to the unfreezing of all layers, making the entire model to be trained from scratch.

##### Acknowledgments

The author is grateful for the provision of MMID dataset by Amazon Web Services for no charge.

##### References

Albawi, S., Mohammed, T.A., Al-Zawi, S., 2017. Understanding of a convolutional neural network, in: 2017 International Conference on Engineering and Technology (ICET). Presented at the 2017 International Conference on Engineering and Technology (ICET), pp. 1–6. https://doi.org/10.1109/ICEngTechnol.2017.8308186

Bader, Y., 1998. Semantic change in Arabic loanwords from English and French. Abhath Al-Yarmouk 8, 33–48.

Chollet, F., 2017. Xception: Deep Learning with Depthwise Separable Convolutions. https://doi.org/10.48550/arXiv.1610.02357

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., Fei-Fei, L., 2009. ImageNet: A large-scale hierarchical image database, in: 2009 IEEE Conference on Computer Vision and Pattern Recognition. Presented at the 2009 IEEE Conference on Computer Vision and Pattern Recognition, pp. 248–255. https://doi.org/10.1109/CVPR.2009.5206848

Eberhard, David M, Gary F. Simons, Fennig, C.D. (Eds.), 2022. Ethnologue: Languages of the World. Online version, Twenty-fifth edition. ed. SIL International, Dallas, Texas.

Elsken, T., Metzen, J.H., Hutter, F., 2019. Neural Architecture Search: A Survey. J. Mach. Learn. Res. 20, 1–21.

English Speaking Countries List | Lingoda Online English Language School [WWW Document], n.d. . Lingoda. URL https://www.lingoda.com/en/content/english-speaking-countries/ (accessed 10.1.22).

Fischer, W., 2020. Video Signal Formats for HDTV and UHDTV, in: Fischer, W. (Ed.), Digital Video and Audio Broadcasting Technology: A Practical Engineering Guide, Signals and Communication Technology. Springer International Publishing, Cham, pp. 97–105. https://doi.org/10.1007/978-3-030-32185-7\_5

Guarasci, R., Silvestri, S., De Pietro, G., Fujita, H., Esposito, M., 2022. BERT syntactic transfer: A computational experiment on Italian, French and English languages. Comput. Speech Lang. 71, 101261. https://doi.org/10.1016/j.csl.2021.101261

Hashim, A., Leitner, G., Aqad, M.A., 2017. Arabic in contact with English in Asia: Linguistic, social and political influences of Arabic in the region. Engl. Today 33, 25–32. https://doi.org/10.1017/S0266078416000377

He, K., Zhang, X., Ren, S., Sun, J., 2015. Deep Residual Learning for Image Recognition. https://doi.org/10.48550/arXiv.1512.03385

Hewitt, J., Ippolito, D., Callahan, B., Kriz, R., Wijaya, D.T., Callison-Burch, C., 2018. 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). Presented at the ACL 2018, Association for Computational Linguistics, Melbourne, Australia, pp. 2566–2576. https://doi.org/10.18653/v1/P18-1239

Howard, A., Sandler, M., Chu, G., Chen, L.-C., Chen, B., Tan, M., Wang, W., Zhu, Y., Pang, R., Vasudevan, V., Le, Q.V., Adam, H., 2019. Searching for MobileNetV3.

Howard, A.G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., Adam, H., 2017. MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications. https://doi.org/10.48550/arXiv.1704.04861

Howell, A., Hohaus, V., Berezovskaya, P., Sachs, K., Braun, J., Durmaz, Ş., Beck, S., 2022. (No) variation in the grammar of alternatives. Linguist. Var. 22, 1–77. https://doi.org/10.1075/lv.19010.how

Jin, H., Song, Q., Hu, X., 2019. Auto-Keras: An Efficient Neural Architecture Search System, in: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD ’19. Association for Computing Machinery, New York, NY, USA, pp. 1946–1956. https://doi.org/10.1145/3292500.3330648

Kaye, A.S., 2018. Arabic, in: The World’s Major Languages. Routledge.

Mahmoud, A., 2000. Modern Standard Arabic vs. Non-Standard Arabic: Where Do Arab Students of EFL Transfer From? Lang. Cult. Curric. 13, 126–136. https://doi.org/10.1080/07908310008666594

Malik-Moraleda, S., Ayyash, D., Gallée, J., Affourtit, J., Hoffmann, M., Mineroff, Z., Jouravlev, O., Fedorenko, E., 2022. An investigation across 45 languages and 12 language families reveals a universal language network. Nat. Neurosci. 25, 1014–1019. https://doi.org/10.1038/s41593-022-01114-5

Miller, G.A., 1995. WordNet: A Lexical Database for English. Commun. ACM 38, 39–41.

Nations, U., n.d. Official Languages [WWW Document]. U. N. URL https://www.un.org/en/our-work/official-languages (accessed 10.1.22).

Rosenbloom, S., 2020. Want to Learn French? Italian? Russian? There’s No Time Like the Present. N. Y. Times.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A.C., Fei-Fei, L., 2015. ImageNet Large Scale Visual Recognition Challenge. Int. J. Comput. Vis. 115, 211–252. https://doi.org/10.1007/s11263-015-0816-y

Sanchez, S.A., Romero, H.J., Morales, A.D., 2020. A review: Comparison of performance metrics of pretrained models for object detection using the TensorFlow framework. IOP Conf. Ser. Mater. Sci. Eng. 844, 012024. https://doi.org/10.1088/1757-899X/844/1/012024

Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., Chen, L.-C., 2019. MobileNetV2: Inverted Residuals and Linear Bottlenecks. https://doi.org/10.48550/arXiv.1801.04381

Seidlhofer, B., 2008. Language Variation and Change: The Case of English as a Lingua Franca, in: Dziubalska-Kolaczyk, K., Przedlacka, J. (Eds.), English Pronunciation Models: A Changing Scene. Verlag Peter Lang.

Shanmugamani, R., 2018. Deep Learning for Computer Vision. Packt Publishing.

Simonyan, K., Zisserman, A., 2015. Very Deep Convolutional Networks for Large-Scale Image Recognition. https://doi.org/10.48550/arXiv.1409.1556

Szegedy, C., Ioffe, S., Vanhoucke, V., Alemi, A., 2016. Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning. https://doi.org/10.48550/arXiv.1602.07261

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z., 2015. Rethinking the Inception Architecture for Computer Vision. https://doi.org/10.48550/arXiv.1512.00567

You, K., Liu, Y., Wang, J., Long, M., 2021. LogME: Practical Assessment of Pre-trained Models for Transfer Learning, in: Proceedings of the 38th International Conference on Machine Learning. Presented at the International Conference on Machine Learning, PMLR, pp. 12133–12143.

Zemánek, P., 2021. Two Hundred Ways to Call a Lion in Arabic: Names or Epithets? Z. Dtsch. Morgenländischen Ges. 171, 343–374. https://doi.org/10.13173/zeitdeutmorggese.171.2.0343

Zoph, B., Vasudevan, V., Shlens, J., Le, Q.V., 2018. Learning Transferable Architectures for Scalable Image Recognition. https://doi.org/10.48550/arXiv.1707.07012