

**EXAMINING THE IMPACT OF STRATIFIED SAMPLING ON  
MODEL PERFORMANCE IN AUTOMATED IMAGE CAPTION:  
A TOPIC MODELLING APPROACH**

**by**

**Anh Nguyen**

**<ID: 20243144>**

**2021**

A Dissertation presented in part consideration for the degree of  
MSc. Business Analytics

## **ACKNOWLEDGEMENT**

First and foremost, I would like to express my sincere gratitude to my supervisor, Doctor Bertrand Perrat for his valuable and suggestive advice in both theoretical and practical aspects of this dissertation.

I would also want to highlight my deep appreciation towards all the teachers and professors at the N/LAB - Nottingham University Business School who have provided me with the necessary knowledge and skills.

Finally, I would like to convey my gratefulness to my family and friends for their encouragement and support throughout this arduous and difficult time.

## **ABSTRACT**

Deep learning's recent rapid development has prompted scientists to investigate a wide range of complex data problems. Among them, automated image captioning has increasingly drawn the attention of many researchers due to its challenging but intriguing architecture that involves a combination of image processing and text analytics. It has also attracted investment from businesses thanks to several practical applications, including image retrieval, impaired vision support, product tagging and automatic drive.

This dissertation investigates the application of a stratified sample split in evaluating the performance of the automated caption model. Despite its popularity in machine learning, stratification has not yet been directly applied in previous works of image captioning, as (1) researchers often utilise pre-defined sample split from data providers, (2) image and annotation are unstructured data that require more novel methodology in clustering versus structured data.

By applying topic modelling to images' annotations, this dissertation validated the positive impact of stratified sampling towards prediction results compared to the usage of a simple random split. The findings also specified the sample size territory where this strategy delivered the best performance and unveiled the reason behind this phenomenon. Finally, the study provided a more comprehensive understanding of the problem with insights on the behaviours of different support techniques in topic modelling and image encoding.

# TABLE OF CONTENT

<b>ACKNOWLEDGEMENT .....</b>	<b>i</b>
<b>ABSTRACT .....</b>	<b>ii</b>
<b>TABLE OF CONTENT .....</b>	<b>iii</b>
<b>LIST OF TABLES.....</b>	<b>v</b>
<b>LIST OF FIGURES.....</b>	<b>vi</b>
<b>1. Introduction .....</b>	<b>1</b>
1.1. Background .....	1
1.2. Relevance of research .....	2
1.3. Research objectives .....	2
1.4. Research outlines .....	3
<b>2. Literature review .....</b>	<b>4</b>
2.1. Automated image captioning .....	4
2.1.1. Methodologies and Model algorithms .....	4
2.1.2. Application direction .....	6
2.1.3. Evaluation metrics .....	7
2.2. Evaluation in Machine Learning.....	8
2.3. Topic modelling .....	9
<b>3. Methodology.....</b>	<b>11</b>
3.1. Resources .....	11
3.2. Baseline model.....	11
3.2.1. Overall structure .....	11
3.2.2. Model architecture .....	13
3.3. Research framework .....	16
3.3.1. Evaluation metric .....	17
3.3.2. Topic modelling .....	17
3.3.3. Stratification impact.....	18
3.3.4. Image encoders .....	19
<b>4. Result.....</b>	<b>21</b>
4.1. Evaluation metric .....	21
4.2. Topic modelling .....	22

4.3. Stratification impact.....	23
4.4. Image encoders .....	26
<b>5. Discussion .....</b>	<b>28</b>
5.1. Understanding the results.....	28
5.2. Results against hypothesis.....	35
5.3. Recommendation.....	36
5.4. Limitation and future work.....	37
<b>6. Conclusion .....</b>	<b>40</b>
<b>REFERENCE.....</b>	<b>41</b>

## LIST OF TABLES

Table 3.1. Baseline model parameters .....	12
Table 3.2. Input and output shape in different encoders.....	19
Table 4.1. BLEU average score and distribution in stratified and random approaches (5 run times) .....	24
Table 5.1. Summary of hypotheses .....	35

## LIST OF FIGURES

Figure 2.1. CRISP-DM machine learning cycle.....	8
Figure 3.1. Convolutional Neural Network architecture.....	13
Figure 3.2. Recurrent Neural Network architecture .....	14
Figure 3.3. Attention-based automated caption architecture .....	15
Figure 3.4. Prediction with attention-based automated caption concept.....	16
Figure 3.5. Research workflow by objectives .....	16
Figure 3.6. Classification accuracy of different CNN .....	20
Figure 4.1. BLEU distribution in 'one-to-one' and 'one-to-all' approaches.....	21
Figure 4.2. Coherence score by LDA and NMF: 'merged' vs. 'unmerged' .....	22
Figure 4.3. Number of captions with no assigned topic between 'merged' and 'unmerged' approaches .....	23
Figure 4.4. BLEU distribution in random and stratified sampling approaches ....	23
Figure 4.5. BLEU average score and distribution in stratified and random approaches with increase in sample size .....	25
Figure 4.6. Loss plot with VGG-16 (left), Inception-V3 (middle), EfficientNet-B7 (right) over 100 epochs.....	26
Figure 4.7. Average BLEU scores with 3 encoders with 20, 30 and 100 epochs.	27
Figure 5.1. Bad captions generated from the 'one-to-one' approach and their scores in 'one-to-all' approach.....	28
Figure 5.2. Coherence score in LDA and NMF methods with increase in sample	29
Figure 5.3. BLEU score and proportion of topics in test set .....	30
Figure 5.4. Captions with similar pattern in Stratified and Random approach ...	31
Figure 5.5. Topic distribution between random and stratified approaches with 1000 and 13000 samples .....	32
Figure 5.6. Changes in BLEU gap and topic distribution gap with increases in sample size .....	32
Figure 5.7. BLEU distribution with EfficientNet-B7 and Inception-V3 (5000 samples) .....	33
Figure 5.8. Bad captions with EfficientNet-B7 and their respective predictions with Inception-V3.....	34
Figure 5.9. An example of good caption with low BLEU score .....	38
Figure 5.10. An example of bad caption with high BLEU score.....	39

## **1. Introduction**

This chapter presents the business and technical background that fuels the study's objectives. A summary of the chapters' content is also listed to provide an overview of the dissertation's structure.

### **1.1. Background**

The explosion of digital technology comes with both opportunities and challenges. While it results in an enormous amount of information, the data generated from this process varies in shapes and formats, requiring different analysis techniques in processing and modelling (Bridgwater, 2018). Among them, images contribute a significant part to digital footprint and comprehend a rich source for data exploit. According to Mylio (2020), a business that provides photo storage and protection service, in 2020, nearly 7.4 trillion photos were created and stored, either on electrical devices, clouds or social network platforms. This development trend leads to the establishment and development of the computer vision field, which focuses on extracting and utilising information gained from graphical objects. Recent advancements in this branch, including the improvement of several deep learning models, have motivated scientists to investigate even more complicated realms of artificial intelligence (Hrga and Ivašić-Kos, 2019). One of them is automated image captioning, a problem that requires an integrated solution between visual recognition and natural language processing. The purpose is to not only unpack the image's information but also generating a sophisticated explanation from that.

As such, image description can be beneficial in numerous subjects. With the deep emergence of search tool, social network and e-commerce in our daily life, one noticeable prosperity of this application is to assist content retrieval and product tagging (Evergreen, 2020). As graphical data exponentially increases in our storage, the manual task of detection and categorization has become impossible. Another well-known and significant inspiration for the automated caption is to support visually impaired people. For instance, Facebook, a platform in which images account for 55.6% of daily posts (Corliss, 2017), has developed a new technology called Automatic Alternative Text that helps to decode photos on news feed into captions to bring a more fulfilling experience to its users (Facebook, 2021).



## **1.2. Relevance of research**

Due to its significant potential, several studies have been established to improve the model accuracy and application capability (Hossain et al., 2018). Despite varying in focus, improvements in this field can be classified into two aspects, algorithm architecture and evaluation methodology. In the second area, researchers have created several evaluation metrics to provide better understanding of model performance and leverage the optimisation process. However, no study has been done investigating the effect of a 'good split' between train and test set on prediction results. While stratification is often used in machine learning to ensure fairness in performance assessment and representativeness of dataset (Menon, 2020), application of this approach in the discussed field could be more complicated as the inputs (image and annotation) are unstructured data. Between these two data types, text appears to be the more approachable target in classification and clustering, since its processing task can be done at high accuracy using unsupervised dimensional reduction techniques.

This study, therefore, investigates the implementation of topic modelling in stratifying the samples to answer the question:

To what extent does stratified sampling affect to prediction result of the automated image captioning model?

Two sub-questions related to this topic were also included. First, is topic modelling an appropriate approach for stratification in this domain? Second, does a more advanced image encoder help to improve model performance?

## **1.3. Research objectives**

In delivering the research ideas, the study is designed to achieve the following objectives.

- Identifying appropriate metric for comparison.
- Finding an effective topic modelling approach for images' captions in sample preparation stage.
- Understanding the differences in model behaviours and accuracies with and without stratified sample split.
- Finding appropriate image encoder for the examined dataset.

As there can be several different approaches to achieve these objectives, it should be noted that due to time and computation constraints, the scope of this study is limited to observing model changes in a single-split procedure, not in a cross-validation and parameter tuning process. The focus is also to understand the model pattern, not to optimise prediction results.

#### **1.4. Research outlines**

This dissertation contains six key sections.

Chapter 1, Introduction, provides an overview of the background that led to the research questions and objectives.

Chapter 2, Literature Review, provides past research in automated image captioning and related fields that inspired the framework of this research.

Chapter 3, Methodology, details the research framework and analytics steps to achieve the study's goal. This section also highlights the hypotheses derived from previous works and the author's knowledge.

Chapter 4, Result, describes the output followed the framework structure.

Chapter 5, Discussion and Future work investigates the reason underlying collected results and presents conclusions on the hypotheses and research questions. Based on that, the impact and limitations of this study are drawn out to provide direction for future research.

Chapter 6, Conclusion, briefly summarise the insights and future steps resulted from this research.

## **2. Literature review**

This chapter describes the development of the discussed field and findings from previous research that supports the topic and methodology developed in this dissertation.

### **2.1. Automated image captioning**

Compared to other machine learning applications within the realm of computer vision, image captioning can be considered a newborn field. Starting with some of the first established research from Farhadi et al. in 2010 and Yang et al. in 2011, the industry has only boomed up since 2014 with the involvement and popularity of neural networks in image recognition and language translation. Since then, several studies have been conducted to improve model efficiency and applicability. While the classification of these techniques is still under investigation with only a few surveys recorded (Hossain et al., 2018, Bai and An, 2018, Wang et al., 2021), this review section will provide an introduction of previous works based on three directions of improvement: Methodologies and Model algorithms, Application directions, and Evaluation metrics.

#### **2.1.1. Methodologies and Model algorithms**

There are three main model architectures in image captioning, including retrieved-based, template-based and deep learning methodologies.

##### **Retrieved-based methodology**

The earliest works published in image description from Farhadi et al., 2010 and Ordonez et al., 2011 belong to the caption retrieval technique. Conceptually, this method concerns evaluating the association between images and sentences to predict a caption constructed from the pre-defined training list. Differences between studies in this sector involve an upgrade in similarity and ranking metrics and a shift from the whole-sentence retrieval to the phases-combined approach that allows more flexibility in generating statements. By using available literature sources, this methodology has the advantage of grammatical and semantical accurateness. However, its main challenge is predicting new object-action patterns with quality heavily depends on the size and diversity of inputs.

## **Template-based methodology**

An effort to deliver more flexible captions by breaking down images and sentences resulted in the template-based approach. Specifically, photos are projected to a range of visual concepts following a fixed sentence template. This research period also observed an improvement in integrated solutions for graphic and word elements, from using separate visual and language decoders to multimodal space approaches. Ushiku et al. (2015) were one of the first author groups to establish this direction when introducing a framework that maps image and phrase vectors into a common space to calculate their similarity and correlation.

While improving the relevancy of caption versus retrieve-based method, the template-based approach still faces the limitation in coverage due to the forcing connection of objects and actions to the sentence pre-defined elements. The captions though being correct in syntax can be semantically meaningless.

## **Deep learning methodology**

The involvement of deep learning opened a new horizon for the image captioning field. Since 2014, several studies in this knowledge zone have been published. Due to the complexity and robustness of neural networks, development direction varies depending on application areas and authors' interests. However, the common ground of these models is utilising a pre-trained Convolutional Neural Network (CNN) to encode the image, then using a language model to integrate CNN outputs with caption's elements (Hossain et al., 2018). However, the combination mechanism for these two models is divided into two categories: Encoder-Decoder and Compositional approach.

In the Encoder-Decoder architecture, two deep nets are involved. First, a sliced CNN that contains only certain hidden convolution layers is obtained to extract the image's vectors. Then, a translation model, Recurrent Neural Network (RNN) will use the graphical inputs taken from the first step, together with output in its recurrent cell to generate captions. By that, sentences are significantly improved in semantic and syntax.

Compositional architecture, on the other hand, is slightly more complicated with the appearance of another model in the process. Instead of extracting the whole image features, this approach uses CNN to detect key visual concepts from each image. These pieces of visual are then fed into a language model to acquire

multiple descriptions. Finally, the multimodal algorithm ranks these candidates to select the best option. As such, the compositional approach allows more flexibility and accuracy in detecting image components. However, as it is not an end-to-end process, the generated captions might be less semantically correct.

### **2.1.2. Application direction**

While a range of overall architectures has been created and studied, recent trend in model development is to focus more on incremental improvement based on application area.

#### **Attention-guided methodology**

First introduced by Xu et al. (2015), the attention-based approach focuses on detecting key image elements and hence provides more details of the graphic. In addition to the traditional encoder-decoder pipeline, attention-guided vectors taken from the last convolution layer of CNN are included as an input to the translation model. This helps to improve the weakness in older work when considering the image as a whole. Therefore, it is extremely valuable in several application zones such as image search and automatic driving, where object recall is more important than full context derivation. A series of research in this branch has been established, revolving around improvements on the attention vector. Some examples include review-based approach in which guided-vector contains multiple hidden states extracted from CNN (Wu et al., 2015), area-based approach that inputs attention vector on each layer of RNN to predict the next word (Pedersoli et al., 2017), or gaze simulation approach with a mechanism to identify human attention to important parts of an image (Sugano et al., 2016).

#### **Semantic-guided methodology**

Despite having a similar approach with the attention-guided procedure, the semantic-guided method pays more attention to the meaning of the generated text. The CNN's output, in this case, provides not only the image features but also their semantic concepts. These concepts later will also be added as an extra layer in the language decoder. This helps to generate more novel sentences that can be used directly to label vision in caption suggestion or non-audio video subtitles. Yao et al. (2017) came up with a model architecture that not only inserts semantic attributes to RNN but also controls the encoder's time step. This approach has limitations considering guided attributes separately and hence might result in

fragmented captions. Providing a broader view direction, Gan et al. (2017) established Semantic Compositional Network, in which semantic vector is generated with multiple context 'tags' that can generalise the overall meaning of pictures.

### **Style-guided methodology**

Aiming to achieve more humanised and creative sentences, this methodology uses separate corpora in training the RNN model, particularly LSTM to combine the factual and style elements from the text. From that, a more expressive and attractive caption can be generated, which is particularly helpful for suggestions in social networks. One of the most famous works in this field is StyleNet, a novel captioning system found by Gan et al. (2017). This method involves an LSTM that is capable of separating context and stylish elements from caption. With a more focus on the sentiment, Mathews et al. (2016) introduced a pipeline called SentiCap, which adds emotional factors to generated captions.

#### **2.1.3. Evaluation metrics**

Found in 2002, **BLEU** (Papineni et al., 2002) was the first and most common metric used to evaluate the accuracy of the machine-generated text. Using the n-grams (maximum 4) mechanism, it considers the similarity precision (percentage of replication) between the predicted caption and a set of alternative references. Despite the benefits of simplicity in concept and closeness to human judgement, the index has limitations including reference size dependency and syntactical ignorance, thus, being relevant only in the case of short text.

Using the same n-grams strategy but focusing more on measure the recall, **ROUGE** (Lin, 2004) calculates the percentage of elements from referenced summaries that prediction text can retrieve.

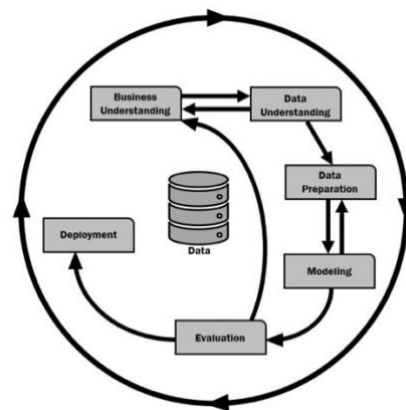
Embracing the strengths of both mentioned methods, **METEOR** (Agarwal & Lavie, 2008) takes into account both precision and recall ability, with extra consideration on the alignment of matched words.

Despite being commonly used in image description evaluation, the three above mentioned methods were created to evaluate the efficiency of the language-translation model. Not until the booming of research in the automated caption in 2015 that specialised measurements in this field were introduced. Building up from previous metrics but mitigating the duplication issue caused by n-grams, **SPICE**

(Anderson et al., 2016) creates a tree-like graph to store the semantic representation. On the other hand, **CIDER** (Vedantam et al., 2015) is a paradigm that pays attention to human judgement by applying Term Frequency Inverse Document Frequency in its calculation.

## 2.2. Evaluation in Machine Learning

Regardless of domain section and model type, a typical machine learning process can be described by a well-known six-phase life cycle called CRISP-DM as illustrated in Figure 2.1.



*Figure 2.1. CRISP-DM machine learning cycle*

*(Adapted from Shearer, 2018 CRISP-DM framework)*

One of the important steps that can be seen from the chart is Evaluation, in which a fundamental but reliable strategy is to split the data into train and test sets (Vabalas, 2019). Specifically, the training batch is used to 'fit' the model while the test one is reserved for performance validation. This helps to prevent overfitting and better understand how the model works with unseen data (Provost and Fawcett, 2013).

To facilitate a 'good split', two approaches are often involved namely Simple Random and Stratified sampling. The overall concept of Random Sampling is that all samples should have the same possibility to be selected (Meng, 2013). On the other hand, the stratification process first divides the whole dataset into several 'strata' – separated data subsets that share similar characteristics. A random selection is then made in each of the sub-groups with 'proportional allocation', meaning that the picked-up sample size in each stratum represents its contribution to the whole dataset (Thompson, 2012).

Between the two approaches, stratified sampling has been proved to provide a more accurate evaluation of model performance (Rao, 2019 and Menon, 2020).

As the strategy aims to create a train set that better represents the population, it often delivers higher accuracy in the test set (Esfahani and Dougherty, 2014, Cleartax, 2021). When there is an unbalance in sample characteristics between the train and test sets, overestimation or underestimation of prediction often occurs.

Ramezan et al. (2019), on the other hand, found out that when the sample size is very large (specifically, 10,000 samples), there is no significant difference between sampling methods. This indicates that the advantages of stratification over simple random could be restricted to a certain limit of sample base.

While stratified sampling is important as above mentioned, its implementation in an automated caption model can be tricky. As the problem deals with unstructured data, it requires a different method to unpack, categorise and stratify the samples versus the normal technique used in a structured data problem.

Past research papers, while focusing on accuracy optimisation, often use the pre-defined data split for train and test sets from the providers or other researchers (Karpathy, 2015). This dissertation, therefore, fills in this gap by investigating the impact of stratified sampling on model performance. The proposed approach used the topic modelling technique to cluster the captions into smaller sub-groups with the same context, with the assumption that this also helps segment the images' content.

### **2.3. Topic modelling**

Topic modelling stands as a commonly used technique in the natural language processing sector that allows the exploration of structure and theme classification within textual data (Wikipedia, 2021).

One of the most recognisable works in early-stage, Probabilistic Latent Semantic Indexing (pLSI) was created by Hofmann (1999). Conceptually, the model works under the assumption that each word represents a particular topic, and therefore each document is a combination of different themes. While successfully introducing the probabilistic concept in topic modelling, the model shows limitations in scalability and a high chance of overfitting with the exponential size of its obtained corpus.



Leveraging this base ground, Blei et al. (2002) established Latent Dirichlet Allocation, or LDA in short, a model that can capture '*the exchangeability of both words and documents*' (Alghamdi and Alfalqi, 2015). The topic, in this matter, consists of a mixture of words that follow a probability distribution, and therefore, helps to solve the constraints shown in previous works. To date, LDA is still considered as the most popular methodologies with several toolkits developed under its mechanism, including MALLET (Mccallum, 2002), Stanford TMT (Daniel et al., 2009), Gensim (Rehurek, 2010).

Besides LDA, a range of different algorithms has also been invented in this field. One of them is Non-negativity Matrix Factorization (NMF), introduced by Berry and Browne (2007). It uses factorizing correlation scheme to derive topics from co-occurrence words. Several studies have proved that NMF delivers a robust performance towards short texts in comparison with other techniques. (Chawla, 2017, Klos, 2020).

In the domain of caption modelling, a study that investigates the effectiveness of different clustering techniques on object-word sentences (Chen et al., 2017) showed that while NMF performs better with 'keywords only', LDA is more powerful with 'full sentence' data which involves more noise and hence is more topic-ambiguous.

In validating topic model efficiency, beside human judgement that is not scalable upon the number of topics, two quantitative metrics commonly involved are Perplexity (or Held out likelihood) and Coherence. While the first one uses prediction methodology to evaluate topic and word intrusion, it is unable to provide similar results to human interpretation (Chang et al., 2009). Coherence score, on the other hand, also consider the correlation between words in topics and topics in document, hence, become a widely used assessment method in topic modelling.

### **3. Methodology**

This chapter describes the research framework and analytics plan to answer the study's questions by first introducing the resources and the baseline model that was used as benchmark for testing. It also emphasises the hypotheses that were derived from previous works and the author's knowledge.

#### **3.1. Resources**

The dataset used for model training and evaluation is the 2014 version of **MS-COCO** (Microsoft Common Object in Context), a crowdsourced file that contains over 164,000 images, each has at least five reference captions. Due to time and computation constraints, only subsets of images and their associated annotations were involved in the analysis (starting with 5000 sample in baseline model). To validate the stratified sampling effect, these subsets of data were randomly selected (with 'random\_state'=42) without considering the predefined categories they belong to.

The coding environment used in this study was Google Colab Pro (Python language). This virtual machine provides T4 and P100 GPU with 12.68 GB RAM, 225.89 GB Disk Memory. It is important to mention that different devices and specifications would result in different numbers versus those that are shown in this dissertation.

In addition, as neural network is a stochastic algorithm, the result gained from re-runs of an exact same model, environment and machine might have a slight discrepancy with the original report. However, the author recorded no significant change in the findings from these re-runs throughout this research.

#### **3.2. Baseline model**

The code implementation in this study is based on the work pipeline for image captioning developed by the TensorFlow team (2018). The model architecture simulates the attention-based concept with an encoder-decoder mechanism introduced by Xu et al. (2015).

##### **3.2.1. Overall structure**

The model structure contains four main sections: sample preparation, image processing, caption processing and modelling.

In **preparing the sample**, the images are multiplied and matched with their respective descriptions to allow flexibility in prediction (for example, 5,000 images with five captions each will result in 25,000 samples), then randomly split into train and test sets with the default ratio of 80% and 20% respectively. To avoid information leakage, the splitting was carried on the image samples only. Associated captions were then accordingly mapped to their subset.

The **image processing** section consists of 3 key steps. First, photos are resized and transformed to the input format required by the encoding model. Next, a pre-trained CNN, in this case, Inception-V3 on ImageNet database is used to extract features from images. As the targeted results are vectors representing the 'attentions', or key objects in pictures, only outputs up to the last convolutional layer are taken (64 x 2048 in vector shape for Inception-V3).

In **annotation encoding**, the captions are first tokenized. Only the most frequent words are kept (5000 words in this case) while the rest is replaced with an unknown ('unk') mark. The purpose of this, strictly following attention-guided concept, is to only consider key components that describe the pictures, as well as reducing computational space. Second, to prepare for later decoding procedure that involves an RNN model, all vectors are padded with '0' indexes to have similar length with the longest sentence.

The **model architecture** involves three components, including CNN Encoder, Attention weighting function and RNN Decoder. The Encoder acts as a fully connected layer that squashes the extracted feature from the image processing stage. The aim of the second function, on the other hand, is to derive attention weight associated with each feature from CNN's output and merge them into a context vector. This vector will later be used as the direction guide in the Decoder process, which use a Gated Recurrent Unit (GRU) to predict caption for given image with a similar mechanism to the language translation model. Detailed parameters used in training are included in Table 3.1.

<b>Optimiser</b>	Adam	<b>Buffer Size</b>	1000
<b>Loss Function</b>	Sparse Categorical Loss	<b>Unit</b>	512
<b>Batch Size</b>	64	<b>Epoch</b>	20

*Table 3.1. Baseline model parameters*

### 3.2.2. Model architecture

The model architecture includes three main parts: a CNN that acts as the image encoder, an RNN as the decoder, and an Attention function to convert CNN's output to the direction vector in RNN.

#### Encoder - Convolutional Neural Network

CNN is an artificial neural network that unpacks graphical objects into structured data components to distinguish one image from another. More specifically, pictures are flattened to a matrix of pixels for further processing. In a normal classification task, the model can be divided into 2 key steps: feature engineering and result classifying as illustrated in Figure 3.1.

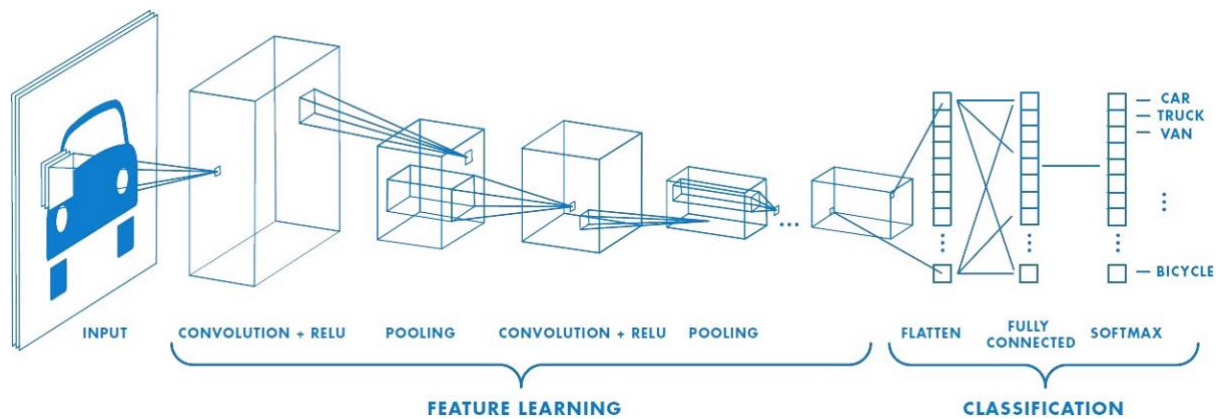


Figure 3.1. Convolutional Neural Network architecture

(Adapted from Saha, 2018 illustration)

To process the image, two network layers are involved including convolution and pooling. Generally, the first layer generates a feature map by performing a dot product between the image's matrix and kernel's matrix. The purpose is to transfer the complicated graphic into only two-dimensional information that helps reduce the learning parameters and computational power required for modelling. The pooling layer, on the other hand, compressed the feature map to a lower dimension. This allows flexibility in prediction, meaning the object can be recognised even with variance versus training input (Mishra, 2020).

In the classification stage, a fully connected layer is used to map the processed input with the output. Finally, a feed-forward network, commonly SoftMax

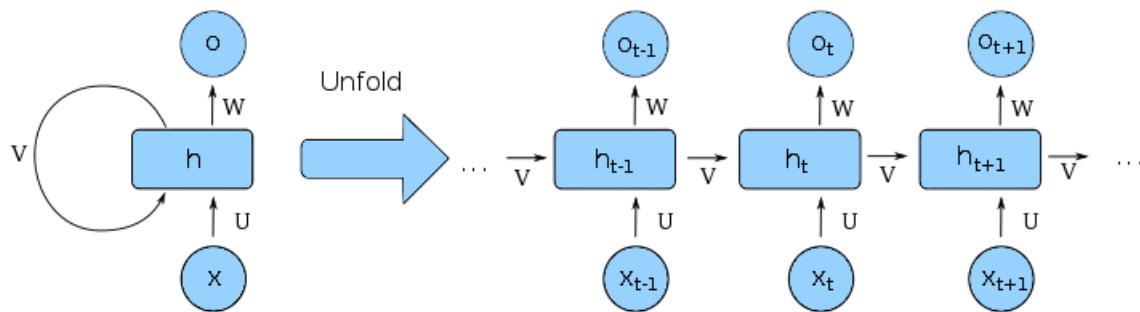
operation is included to provide the class label for the image based on its dominant probability (Saha, 2018).

For automated captioning, as image encoding is only one part of the feature engineering process, a common practice in most past research is to apply a highly accurate pre-trained CNN, up to the last convolutional layer to extract image vector for the decoding process.

As mentioned, the CNN model used in this study is InceptionV3, the third version of Google's Inception Convolutional Neural Network released in 2015. It is among the first models to apply batch normalisation for the fully connected layer which helps to provide regularisation and reducing generalisation error (Karim, 2019).

### Encoder - Recurrent Neural Network

RNN is a deep learning model that is extensively utilised in natural language processing. It detects the sequential properties of data and uses patterns to forecast the next most likely outcome. Figure 3.2 describes the overall structure of a basic RNN with its compressed form diagram on the left-handed side, and the expanded version on the right side (Phi, 2018).



*Figure 3.2. Recurrent Neural Network architecture  
(Adapted from Phi, 2018 illustration)*

Generally, inputs to RNNs are sequences with interconnections, which can be a time series, speech, or text data. They also represent or store some information about previous steps. For example, the output at time  $t+1$  as illustrated is not only based on input  $x_{t+1}$  but also the previous state  $h_t$ .

As discussed, the applied RNN in this dissertation is Gated Recurrent Units (GRU). The model outperforms traditional RNN as it stores both short-term and long-term

memories as input to predict the output of next time step. This not only results in better prediction due to closer simulation of human cognition but also helps to avoid vanishing and exploding gradient problem. Versus Long short-term memory (LSTM), another robust RNN, GRU is more lightweight in architecture, hence faster in training and more appropriate when dealing with a small sample size (Cho et al., 2014).

### Attention function

The involvement of Attention mechanism in caption generating process is illustrated in Figure 3.3.

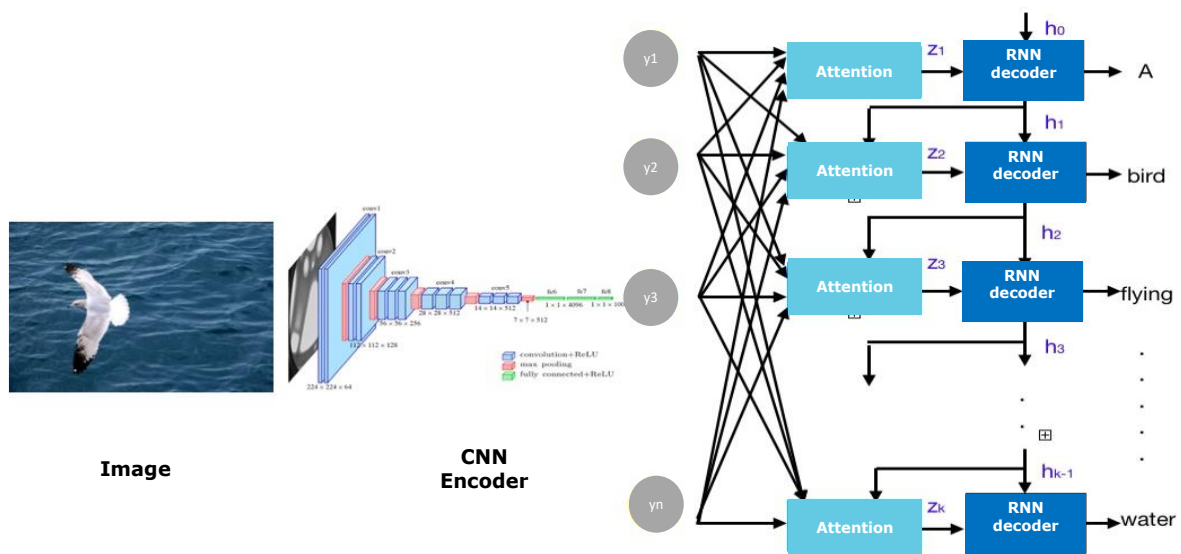


Figure 3.3. Attention-based automated caption architecture

(Adapted from Sarkar, 2020 illustration)

Different parts of the image generated from CNN under the last convolution layer are associated with their locations. Under the attention mechanism, they are converted into context vectors with respective weights. These weights later are fit to an attention function and then acts as direction guidance for the encoder (RNN) to concentrate on specific sections of the image.

Specifically, for this baseline model, the 'local' attention concept is applied. Instead of using all input words as source of attention just like the 'global' concept, it only concerns a subset of words that aligns with the predicted position in input sentence. By that, 'local' attention becomes less computationally expensive and more suitable for longer text generation (Sarkar, 2020).

An illustration of the prediction with attention-guided method can be found in Figure 3.4.

Prediction Caption: the person is riding a surfboard in the ocean <end>

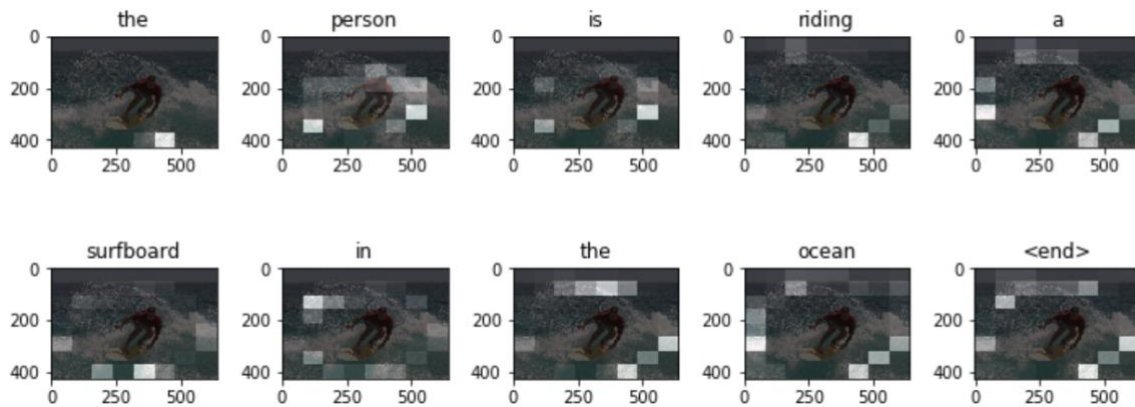


Figure 3.4. Prediction with attention-based automated caption concept

(Adapted from TensorFlow, 2018 illustration)

### 3.3. Research framework

To answer the research objectives, four respective testing phases were conducted as illustrated in Figure 3.5.

1	2	3	4
Evaluation Metric	Topic modelling	Stratification impact	Encoder comparison
Apply on Baseline model	Apply on Captions	Apply on Baseline model	Apply on Stratified model
<b>BLEU:</b> One-to-all   One-to-one  N = 5000	<b>Text:</b> Merged   Unmerged captions  <b>Model:</b> LDA   NMF  N = 5000	<b>Accuracy:</b> Random   Stratified <5 run times> N=5000  <b>Scalability:</b> Random   Stratified <N = 1000   5000   9000   13000>	<b>CNN:</b> Inception-V3   VGG-16   EfficientNet-B7  <epochs = 20   30   100>  N=5000

Figure 3.5. Research workflow by objectives

### 3.3.1. Evaluation metric

To deliver an accurate comparison between methods, the first important step is to have an appropriate evaluation metric. As such, a new section for performance evaluation was added to the pipeline in both the modified and control workflows. That includes a quantitative score for performance benchmark, and a qualitative analysis on the generated captions to provide more insights on the prediction.

BLEU-1 was chosen as the quantitative assessment parameter, due to its simple mechanism and hence reasonable computational expense. Plus, as the index only calculate the similarity between candidate and reference captions without considering grammar and syntax, it also well reflects the focus in object detection of the attention-based approach.

On the other hand, this advantage is also BLEU's limitation as it does not consider word's synonyms as acceptable replacements. As such, the author examined two different BLEU calculation methods on the baseline workflow to select the best one. The first option is called '**one-to-one**' approach in which the model predicted captions for all duplications of images. Each of these candidate annotations was then compared to the respective version of references to derive the average BLEU score for the whole caption set. In the '**one-to-all**' approach, the model only predicted one caption for each image. The evaluation was then made by comparing this single result to all reference versions.

**The first hypothesis** for this process was that the '**one-to-all**' approach would perform better as it allows more flexibility in caption matching.

### 3.3.2. Topic modelling

In delivering the equal sample split, topic modelling was applied on samples' captions to derive different context 'strata'. To define the best approach of topic modelling, two layers of testing were involved.

**The first layer** concerns choosing the representative caption for image. As each photo has different annotations describing different features, in some cases they could be divided into more than one topic group. With the stratified method in which train and test samples are randomly picked up from all context baskets, this multi-label approach might result in information leakage when same pictures appear in both the training and testing phase. To avoid this limitation, the study examined the model's effectiveness between a naïve method of using only the



first caption to represent the image' theme (hereinafter referred to as the '**unmerged**' approach) and the '**merged**' approach that combines all annotations of each image into a sequence for topic grouping.

**The second hypothesis** the author had was that longer sentence, in this case, the 'merged' route, would result in more stable performance of topic modelling.

In the **second testing layer**, LDA and NMF, two popular topic modelling methods (using Gensim package), were applied in both the '**unmerged**' and '**merged**' options. As a large dataset resulted in a significant number of topics that would be difficult to handle by manual judgement, performances were quantitatively compared using the **coherence score**. Before modelling, captions went through a cleaning process to remove stop words and punctuations. Bigram phrases, which are the combinations of two frequently accompanied words were also formed to identify dependency relations between words. Finally, a lemmatisation process was implemented, reserving only nouns, verbs, adjectives, and adverbs in each sentence to derive more concentrated topics.

Based on the study of Chen et al. (2017) on cross-situational object-word learning, **the third hypothesis** was that NMF would perform better in the 'unmerged' route with short captions, while LDA thrived in the merging option as combined sentences are longer and contain more noises.

### **3.3.3. Stratification impact**

To answer the most important question of this dissertation on the impact of stratified sampling impact to prediction, experiments between control and test prototypes of the model was carried out.

While the control version (baseline) used a random split to derive the train and test set, the test versions utilised topic modelling to divide caption samples and respective images into smaller baskets of content to support sample stratification.

Two assessed targets in this section were **accuracy** and **scalability**.

In **accuracy**, to ensure a fair and consistent evaluation, the two code samples went through five run times with different random states (from 0 to 4).

**The fourth hypothesis** was that prediction gained from the stratified process would perform better than its counterpart from the simple random split. This resulted from the assumption that unstructured machine learning problems have

a similar mechanism to structured ones. And therefore, as per previous findings from Esfahani and Dougherty (2014), sample stratification would increase the accuracy of prediction.

If the fourth hypothesis is confirmed, it would also validate the author's **fifth hypothesis** that topic modelling would be an appropriate method for stratification in automated image captioning.

For **scalability**, as per study from Ramezan et al. (2019), there could be a decrease in stratification impact when the samples become larger. This formed **the sixth hypothesis**, which was examined by an iteration of control and test models on a range of sample sizes (1,000; 5,000; 9,000; 13,000).

### 3.3.4. Image encoders

To enhance prediction results and gain more understanding of how different image encoders impact on overall performance, three CNN versions were applied to the baseline workflow.

**VGG-16**, the most used CNN architecture in past papers (Hossain et al., 2018), and **EfficientNet-B7**, one of the most updated pre-trained models were benchmarked versus the original **Inception-V3** from the baseline framework. Different choices of epoch number were also examined with each option.

Parameters of the three models are included in Table 3.2 based on the information provided by Fu, 2020 and Huilgol, 2020.

	VGG-16	Inception-V3	EfficientNet-B7
Input Image size	224 x 224 x 3	299 x 299 x 3	600 x 600 x 3
Output shape	49 x 512	64 x 2048	324 x 2560

*Table 3.2. Input and output shape in different encoders*

It should be noted that CNN models can perform normally with different image size inputs. However, following the finding from a study of Tan and Le (2019) that increasing image resolution resulted in better prediction for transfer learning, the author decided to resize the shape as original instruction from Keras's document as specified on above table.

Figure 3.6 illustrates a performance comparison between different CNN architectures (Tan and Le, 2019). Based on that, the author came up with **the seventh hypothesis** that EfficientNet-B7 would deliver the best performance among all, despite having a possible slowest run time.

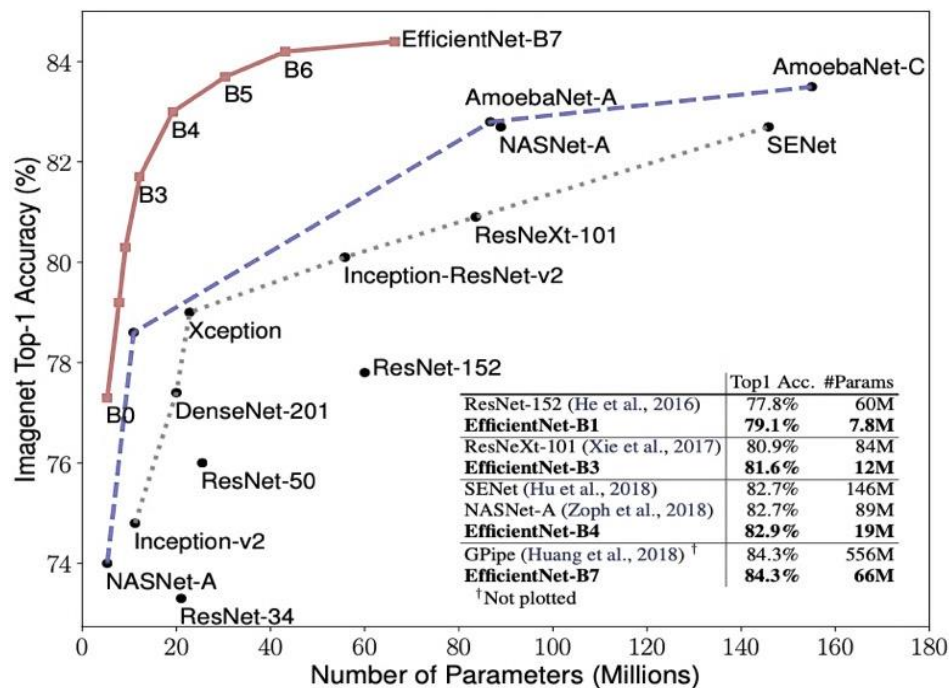


Figure 3.6. Classification accuracy of different CNN  
(Adapted from Tan and Le, 2019 illustration)

## 4. Result

This chapter provides the output resulted in four testing phases: Evaluation metric, Topic Modelling, Stratification impact and Image encoder comparison.

### 4.1. Evaluation metric

The prediction results on validation set from baseline model (total 5000 samples) went through two different approaches of evaluation as mentioned in Methodology section. Overall, BLEU score derived from 'one-to-all' method received over 5 points (5.2%) higher than it was in the 'one-to-one' path, 0.486 versus 0.434. Their detailed scores' distributions are included in Figure 4.1.

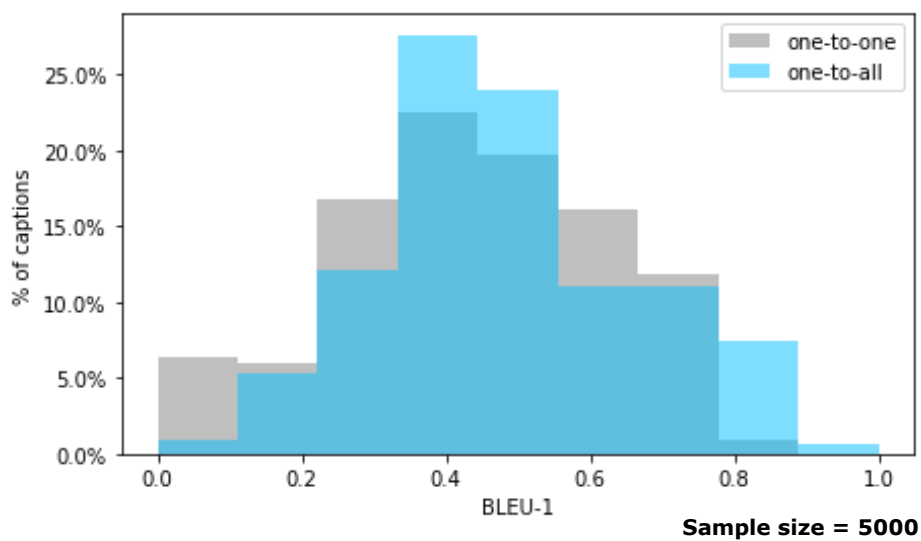


Figure 4.1. BLEU distribution in 'one-to-one' and 'one-to-all' approaches

There were two key differences between the two results observed from the chart that led to the gap in their average results.

First, the 'one-to-all' method had an extremely lower proportion of 'very bad' captions (BLEU < 0.1) compared to the 'one-to-one' option (0.8% versus 6.2%). Instead, it generated a significantly higher percentage of 'acceptable' predictions that have BLEU ranging from 0.4 to 0.6 (38.3% versus 33.9%).

Second, the contribution of 'very good' captions with BLEU > 0.8 for 'one-to-all' was also superior to the counterpart (4.8% versus 0.28%).

While more in-depth analysis regarding this matter is included in Chapter 5 – Discussion, generally 'one-to-all' route delivered not only higher but also fairer results in assessing model performance. Hence, it was selected as the quantitative metric for evaluation in following stages.

## 4.2. Topic modelling

In selecting the best methodology for topic modelling, two layers of testing between LDA and NMF, and between 'merged' and 'unmerged' approaches were intersected to derive respective coherence scores. Due to unknown point of optimisation, both options were first run in a range of 10 to 200 topics with jump step of 10 unit. The results can be found in Figure 4.2.

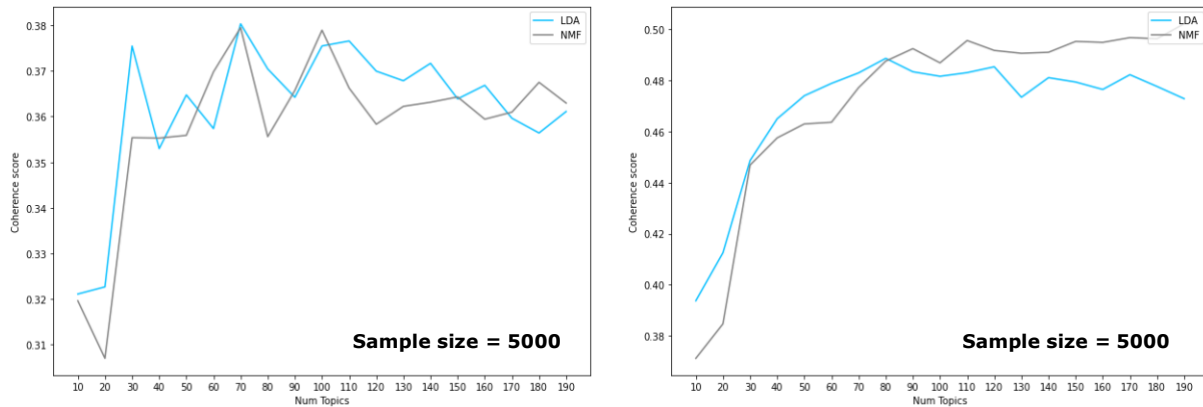


Figure 4.2. Coherence score by LDA and NMF in 'merged' (left) and 'unmerged' (right) approaches

Overall, in optimising coherence, 'merged' option produced considerably smaller groups of contexts versus 'unmerged' option.

It also can be seen that there was no significant difference between the performance of LDA and NMF across different options. However, LDA reached its peak at a smaller option of topic numbers (30 in merging and 70 in unmerging) compared to NMF (70 in merging and 90 in unmerging).

Iteration with larger numbers of topics in 'unmerged' method showed that coherence score continued to rise dramatically, leading to a sanity check on the transformed captions. It showed that considering caption individually resulted in a significant number of unlabelled captions (199 samples), while no such pattern found in the 'merging' concept (Figure 4.3).

With the proposed method, it is critical for all captions to be categorised. Therefore, 'merged' approach are the more sensible direction. A closer look at the performances of 'merged' caption in its optimisation range showed that the first break points after which no noticeable increase in coherence line witnessed was

LDA with 30 topics. Using elbow method, this option was chosen as the topic modelling technique for further steps.

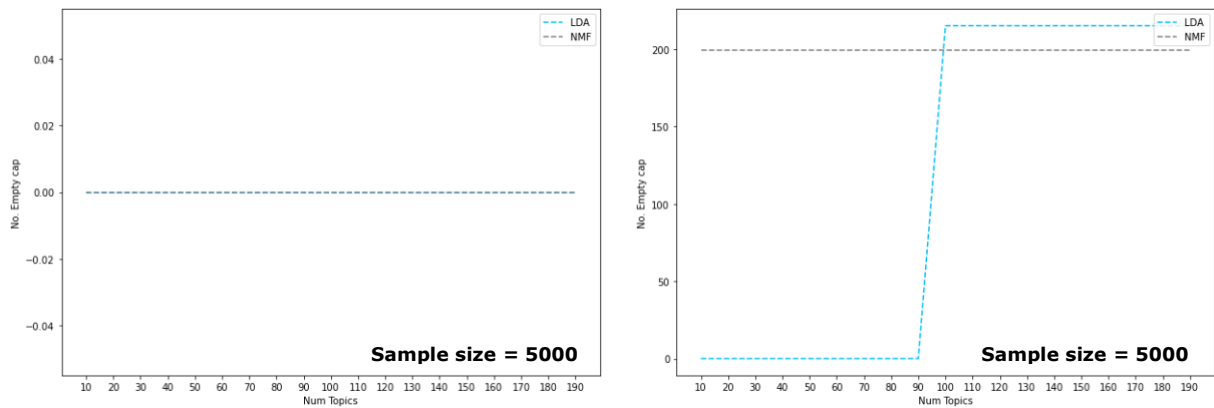


Figure 4.3. Number of captions with no assigned topic between 'merged' (left) and 'unmerged' (right) approaches

### 4.3. Stratification impact

As discussed, the key advantages of stratified sampling include **accuracy** improvement due to more representative data in train and test sets. However, when the sample size increases, the chance to derive a representative set from a random split is also higher, hence decrease the **scalability** of stratification. The investigation on impact, therefore, also focused on these two aspects.

In terms of **accuracy**, the involvement of stratified sampling with an appropriate topic modelling technique showed a slight improvement versus benchmark pipeline of an over 2-point increase, when average BLEU scores for baseline model, and the 'merged' versions were 0.486 and 0.510.

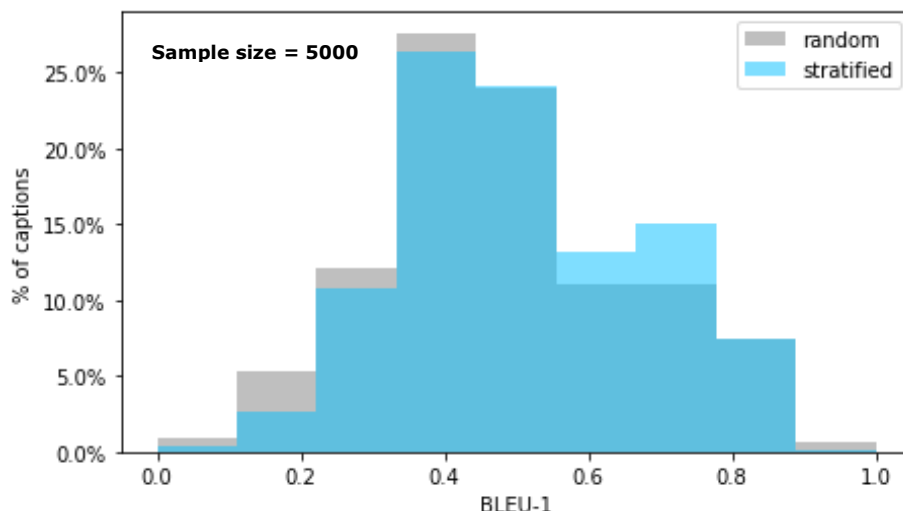


Figure 4.4. BLEU distribution in random and stratified sampling approaches

Further look at the distribution of the 2 approaches (Figure 4.4) showed that merging option with 30 topics was the more optimal selection.

Overall, the higher performance observed in the stratified model is driven by its ability to derive significantly lower number of 'bad' captions. In more details, the proportion of predictions having BLEU less than 0.3 in this version was 9.2%, while such numbers in benchmark approaches was 13.4%. Similarly, 'good' results (BLEU from 0.6 to 0.8) in stratified and random methods accounted for 25.8% and 22.7% of the total generated annotations respectively. However, it should be noted that there is no significant difference in the number of 'very good' captions (around 4.5%), as well as the achieved maximum scores (both are 0.9) between two approaches.

To validate this vulnerable gap and understand the stability of models' performances, 5 test runs with different random states (rd) from 0 to 4 were made.

Sample size = 5000

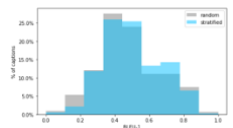
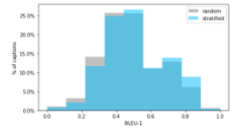
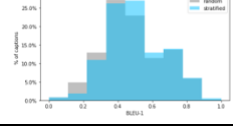
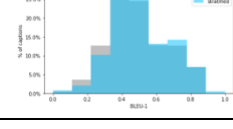
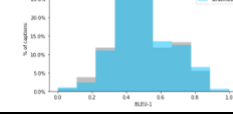
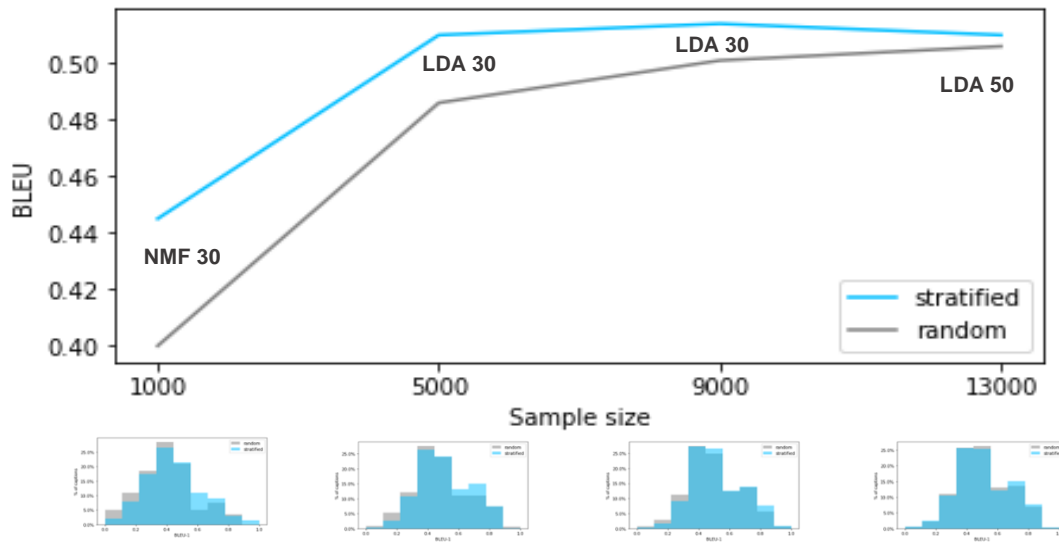
Trial	Random sampling	Stratified sampling	Histogram
1 (rd=0)	0.486	0.510	
2 (rd=1)	0.490	0.511	
3 (rd=2)	0.487	0.505	
4 (rd=3)	0.496	0.511	
5 (rd=4)	0.490	0.502	

Table 4.1. BLEU average score and distribution in stratified and random approaches (5 run times)

The collected BLEU scores for each sampling options shown in Table 4.1 proved stratified sampling consistently delivered higher BLEU versus the random approach. However, regarding stability, both models showed good performance when their prediction scores sliding within a small range of about 0.1.

To understand the solution's **scalability**, random and stratified models were iterated with different sample sizes ranging from 1000 to 9000. In each run time, the author reapplied the 'merged' topic modelling framework and compared the coherency between NMF and LDA to select the best number of topics. In situation when the global optimisation was not clear, elbow method was applied.

The results of topic choices, BLEU scores in average and full distribution version are presented in Figure 4.5.



*Figure 4.5. BLEU average score and distribution in stratified and random approaches with increase in sample size*

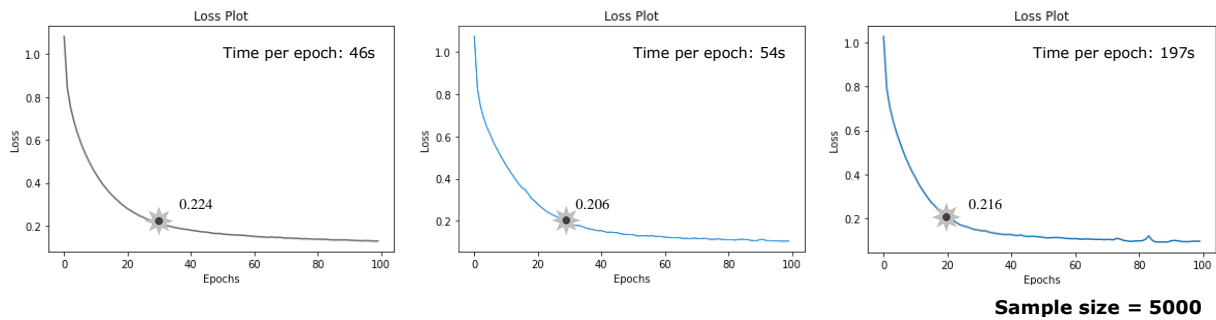
Overall, the gap in accuracy between random and stratified models gradually reduced with the increase in sample size. Up to 13000 samples, no difference between the two versions was observed.

It also can be seen that both models had a boost in performance when the samples increase from 1000 to 5000 images. After this point, the baseline model witnessed only slight improvement while the stratified route stabilised with expansion in sample base.



#### 4.4. Image encoders

To understand how different encoders contribute to the overall performance, the change in model's loss value with 3 options of VGG-16, Incpetion-V3 and EfficientNet-B7 over 100 epochs was examined with the standard 5000 samples (Figure 4.6). Since in automated caption model, loss value computation relies on the imperfect word-to-word metric, only the learning curve from training set, not the validation set, was considered as an indicator of how well the model fit.



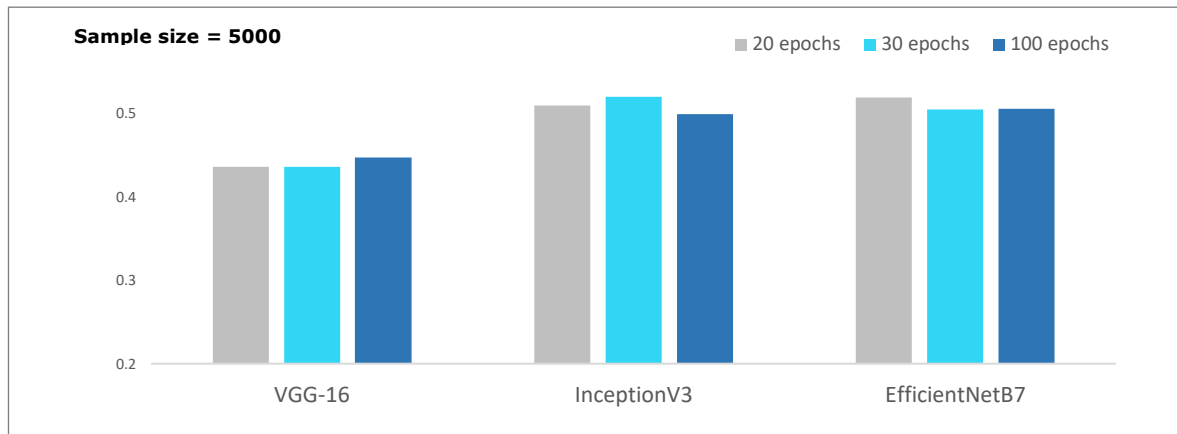
*Figure 4.6. Loss plot with VGG-16 (left), Inception-V3 (middle), EfficientNet-B7 (right) over 100 epochs*

In terms of computation, VGG-16 delivered the fastest run per epoch (46 seconds), while the Inception-V3 and was not far from behind with 54 seconds. As expected, EfficientNet-B7 with the highest image resolution requirement showed the longest run time per epoch of 197 seconds.

Overall, all 3 models showed good learning rates. EfficientNet-B7 had the best performance when its plateau points in loss was achieved at 20 epochs. VGG-16 and Inception-V3 observed similar pattern with optimal point showed at around 30 epochs. The three models were re-run with these options to see which one delivers the best performance in BLEU. Figure 4.7 illustrates the results from this process.

It can be seen that VGG-16 delivered the lowest score in prediction versus the other two techniques. On the other hand, Inception-V3 and EfficientNet-B7 performed equally well at their optimal number of epochs of 30 and 20 respectively. However, regarding computation cost, Inception-V3 achieved this result after around 27 minutes, while for EfficientNet-B7, it took approximately 66 minutes.

In addition, while being the more advanced image classification technique as discussed in the Methodology section, EfficientNet-B7 did not provide a better result as expected. More insights on this phenomenon will be unpacked in the Discussion section.



*Figure 4.7. Average BLEU scores with 3 encoders with 20, 30 and 100 epochs*

## 5. Discussion

This chapter discusses the rationale for the collected results and draws conclusions about the hypotheses and research questions. Based on that, the study's impact and limitations are identified in order to provide suggestions for future research.

### 5.1. Understanding the results






Further investigation into the results was made to better understand the model's behaviours. This section follows four steps of the framework including Evaluation metric, Topic modelling, Stratification impact and Image encoders.

#### Evaluation metric

As previously mentioned, the 'one-to-all' evaluation approach showed a higher prediction score and was referred to as the key metric in further steps. This matched the first hypothesis in assuming when being compared to different versions of benchmark, generated captions will have a higher chance of matching.

A qualitative look into the results seconded this argument. Examples of what were considered 'bad' captions from the 'one-to-one' approach ('true' vs. 'pred') are included in Figure 5.1. On the right-hand side, BLEU scores for these exact same captions when comparing in 'one-to-all' concept are shown (each 'pred' vs. all 'true')

Overall, each of the candidates received a higher score when being compared to all 5 references. Plus, while the first prediction could have received 90% of similarity, especially to the fifth reference, it only matched 49% with the chosen benchmark and therefore reducing the overall evaluation score.

	true: two people stand near bicycles behind a motorcyclist driving on the street pred: a person is on top of a motor cycle BLEU: 0.4919625503668659	0.904
	true: in the road a man is riding a motorcycle and two people are standing with bicycles pred: a person on a motorcycle rope next to a silly wheeled gear on a street with lots of <unk> BLEU: 0.22793076437070336	0.367
	true: a man on a red motorcycle driving on the road pred: a surfer is riding on a street with a helmet attached to a bunch of wave behind him BLEU: 0.3146660996956415	0.492
	true: person on motorcycle rides pass couple of people on bikes pred: a person on a motorcycle with a helmet attached to a group on a motorcycle BLEU: 0.3715011599826719	0.526
	true: a couple of people on a city street pred: a person on a cycle BLEU: 0.34154357946095637	0.436

*Figure 5.1. Bad captions generated from the 'one-to-one' approach and their scores in 'one-to-all' approach*

## Topic modelling

By observing the results when applying 2 testing layers to a range of different sample sizes, the author comes up with 3 key findings.

First, merging captions helped to reduce the ambiguity of sentence context and hence results in no unlabelled caption as well as a smaller number of topics. On possible explanation was that short captions provides insufficient information for algorithms to perform properly. A qualitative look into the empty-topic cases derived from the NMF model in the 'unmerged' option provides more insights regarding this matter.

As mentioned in the **Methodology** section, in reducing the dimension and focusing on the key content of the text, one important step was to keep only the nouns, verbs, adjectives and adverbs. However, it is observed that in all the unclassified sentences, words were identified as proper nouns and therefore removed in the lemmatised dataset. This included captions that have objects stand at the beginning (e.g., 'bathroom', 'kitchen'), or have incorrect grammar (e.g., 'A bathroom with a TV near the mirror') (details in supported notebook for 'unmerged' topic modelling).

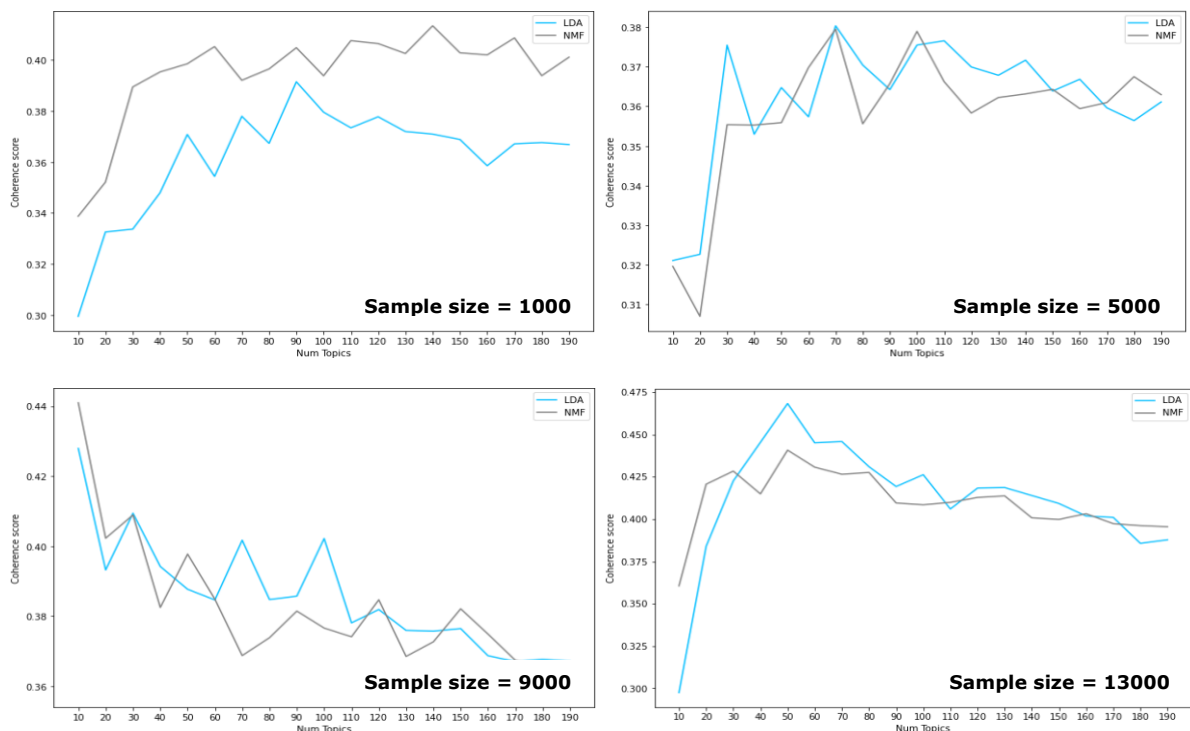


Figure 5.2. Coherence score in LDA and NMF methods with increase in samples ('merged' approach)

On the other hand, the merging option, by combining caption and lengthening the considered text, eliminated the possibility of empty captions. While some important keywords could still be incorrectly encoded, the rest remained and defined the topic of the sentence.

Second, increasing samples did not necessarily lead to more topics but even relatively reduce the number of topics in coherence optimisation (Figure 5.2). The peak score achieved in large sample sizes also surpassed the smaller options.

Third, comparative analysis between techniques' coherency showed that NMF performed better with a smaller sample size. When data rose to 5000 samples, there was no difference between the two methods. Up to 13,000 samples, LDA delivered a higher and clearer peak of coherency.

### Stratification impact

With an appropriate selection of topic modelling methodology, stratified sampling showed an improvement in prediction as mentioned in the Result section. To validate the hypothesis that equal distribution of topics in train and test set was the key driver for this increment, a topic mapping step was made between the topics' contribution and prediction accuracy in random and stratified approaches. The result is illustrated in Figure 5.3.

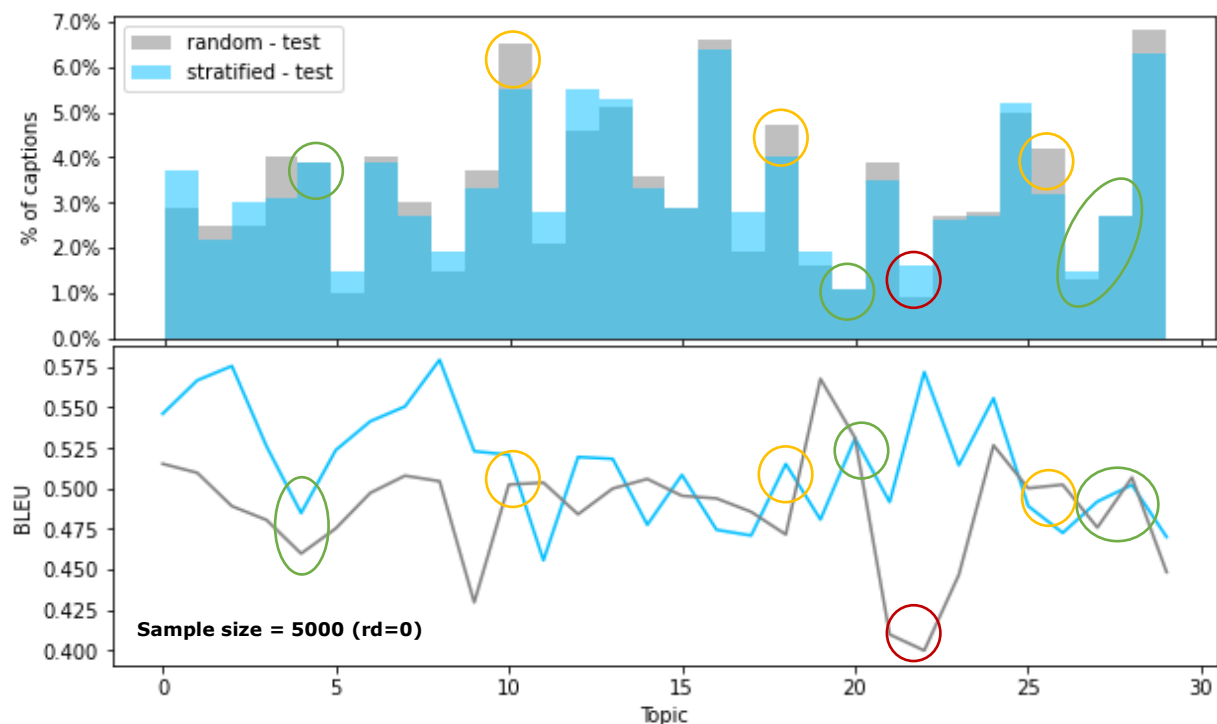


Figure 5.3. BLEU score and proportion of topics in test set: Stratified vs. Random

From the chart, there are 3 insights than can be drawn out regarding the effectiveness of stratified sampling.

**First**, the stratified approach delivered more consistent performance with less variance across topics compared to the random approach. This proves that having an equal split provides sufficient information for prediction in each 'strata', even the ones with less sample size.

**Second**, the random approach performed worst in minor topics that have fewer samples in the test set than they should (highlighted red), however, showed no difference with the stratified approach in larger topics that have the same issue (highlighted orange). This indicates that when the input expands to a certain size, prediction results become more consistent.

**Third**, topics with equivalent sample allocation between the two approaches also received similar BLEU scores (highlighted green). This shows that regardless of the detailed context, same number of 'mutually inclusive' samples will result in same prediction performance. Hence, topic modelling on captions does help in clustering graphical objects as per assumption.

To illustrate this argument, examples of bad and acceptable captions in Topic 4 generated from two methods are included in Figure 5.4 below. In both cases, the two models showed very similar word and syntax patterns.



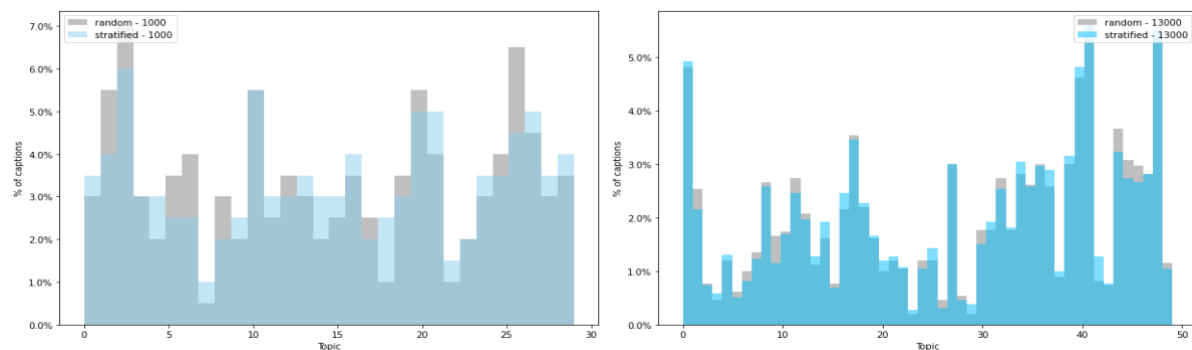
Bad caption	BLEU
Random: The train on a train	0.42
Stratified: A bus is strain at a bus on a city	0.30



Acceptable caption	BLEU
Random: A train traveling on tracks near a pass	0.64
Stratified: A train on the tracks on a fence	0.42

*Figure 5.4. Captions with similar pattern in Stratified and Random approach*

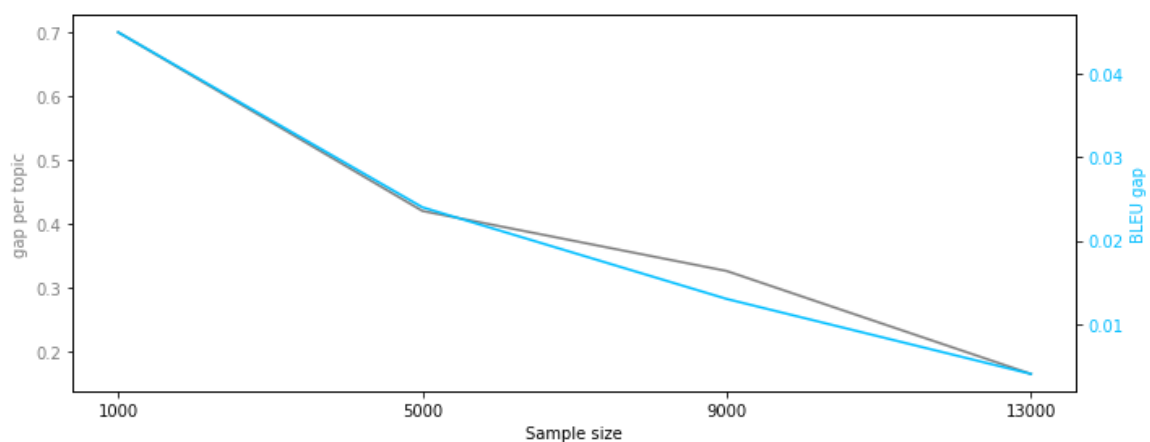
Similarly, when examining the model's scalability with changes in sample size, the result showed that a more balanced split led to better prediction. More specifically, in a random trial that provided a more equal 'split' as showed with 13,000 samples, the accuracy increased to a similar rate with stratified model result. In contrast, the significant improvement observed from 1,000 samples was due to the considerable gap in topic distributions between its random and stratified approaches (Figure 5.5).



*Figure 5.5. Topic distribution between random and stratified approaches with 1000 and 13000 samples*

To provide a clearer view, Figure 5.6 demonstrates the difference between topic samples in random and stratified approaches with an increase in sample size (grey line). The index, in each approach, was computed by first adding up the absolute gaps of samples' proportions in each topic. The total number was then divided to the number of topics to derive the average rate.

The associated BLEU gaps between the two methods are also added (blue line).

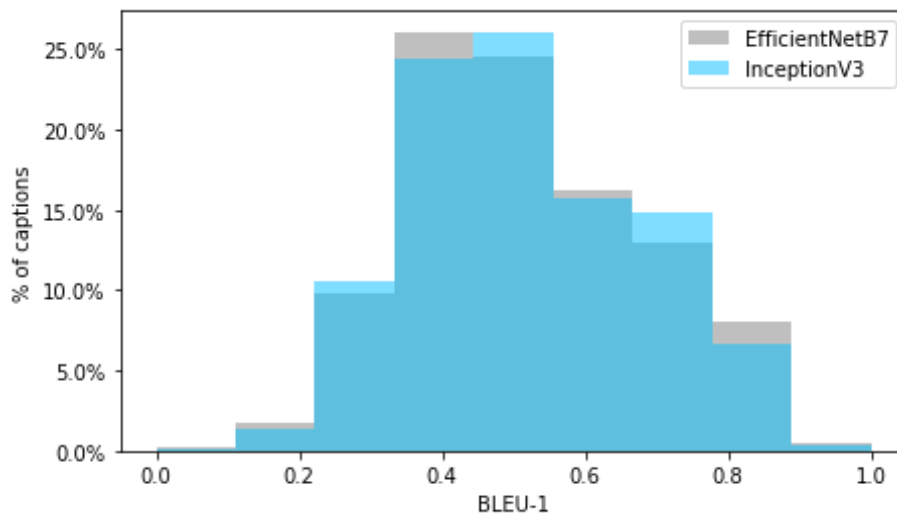


*Figure 5.6. Changes in BLEU gap and topic distribution gap with increases in sample size*

This results in two nearly identical slopes, proving that the more equal distributions between the train and test set, the smaller gap in the observed prediction accuracy between simple random and stratified methods.

### Image encoders

As mentioned in the Result section, despite being the more robust CNN, EfficientNet-B7 did not provide a superior result versus Inception-V3. In fact, it took a longer time run to achieve similar prediction score.



*Figure 5.7. BLEU distribution with EfficientNet-B7 and Inception-V3 (5000 samples)*

Figure 5.7 illustrates the distribution of BLEU score derived from the two approaches. It can be seen that while EfficientNet-B7 delivered a higher percentage of 'very good' captions (BLEU  $\geq 0.8$ ) compared to Inception-V3, it also generated more 'bad' caption (BLEU  $\leq 0.3$ ).

A qualitative look into the captions were carried out to better understand this phenomenon. Based on the gap between 2 encoders as mentioned above, similar pictures receiving under 0.3 in BLEU from EfficientNet-B7 were taken into consideration. The examples in Figure 5.8 demonstrated that bad captions from



EfficientNet-B7 method having much longer sequences versus their counterparts using Inception-V3.



Caption	BLEU
<b>InceptionV3:</b> A couple with a purple coloured cake and a large room	0.82
<b>EfficientNetB7:</b> Several people in dress clothes cutting a pair of people in the room with a wedding cake	0.24



Caption	BLEU
<b>InceptionV3:</b> A yellow and red train on some tracks	0.57
<b>EfficientNetB7:</b> A train is on the tracks near a blue starting to its starting green house	0.27

*Figure 5.8. Bad captions with EfficientNet-B7 and their respective predictions with Inception-V3*

To validate that this was not a coincidence, the average length of captions was calculated. The results showed that while original captions have average length of 10.5 words, EfficientNet-B7 generated on average 11.3 words per caption, and 17.5 words per its bad caption (BLEU < 0.3). The related number from Inception-V3 were 10.5 and 10.8 respectively, proving the argument from qualitative analysis.

Looking back at the two model structures and parameters, one explanation would be that EfficientNet-B7 separated image into more pieces (generated vector shape = 324 x 2560) compared to Inception-V3 (vector shape = 64 x 2048). This mechanism could be helpful in classification task as it provided more image's information and hence more precise result. However, in generating caption, it led to a fragmented attention and therefore more lengthy and less concise

sentences. The examples of bad captions as above illustrated also support this explanation when more details in pictures were detected with EfficientNet-B7.

## 5.2. Results against hypothesis

Summary of the hypotheses and respective references are included in Table 5.1.

	<b>Hypotheses</b>	<b>Reference</b>
No. 1	'One-to-all' comparison delivers better result in BLEU versus 'one-to-one'	The author
No. 2	'Merged' captions delivers more stable topic modelling result versus 'unmerged'	The author
No. 3	NMF performs better with 'unmerged', LDA performs better with 'merged'	Chen et al., 2017
No. 4	Stratified sampling provides more accurate prediction versus random sampling	Esfahani & Dougherty, 2014
No. 5	Topic modelling is an appropriate method for stratified sampling in automated caption domain	The author
No. 6	Stratification impact decrease when sample size become larger	Ramezan et al., 2019
No. 7	EfficientNet-B7 delivers better performance versus Inception-V3 and VGG-16	Tan & Le, 2019

*Table 5.1. Summary of hypotheses*

Overall, the findings have confirmed the most important hypothesis (No. 4) which is also the key objective of this study: stratified sampling had a positive impact on the prediction result. In line with the works of Rao (2019) and Esfahani and Dougherty (2014), this study proved that stratification in splitting samples consistently increased model accuracy with different run tests. More specifically, the increment was achieved due to equal allocation of different graphical contexts in both train and test sets. The performance on the stratified model was also stable across topics even minor ones, which was not attained by the random approach.

The results also supported Ramezan et al. (2019) in arguing that when the sample base becomes very large, the difference in accuracy reduces (No. 6). Moreover, further investigation into this phenomenon showed that in large sample sizes, a

simple random split behaved similarly with a stratified split and therefore the gap in accuracies decreased.

Topic modelling, as per the author's speculation, was proved to be a quick, simple but effective method in stratification (No. 5). To improve the model performance, an approach of merging caption was examined and evidenced as the better choice versus the unmerging path (No. 2).

In addition, for the more ambiguous type of text such as image's caption in 'merged' approach, LDA performed slightly better versus NMF as stated by Chen et al. (2017) (No. 3). Another finding on this section was that when the sample size increased, LDA showed a clearer win versus NMF.

For evaluation metrics, the study seconded the first hypothesis (No. 1) of the author that using only one evaluation score for image by comparing its generated caption with all references helped to boost the average result of the whole dataset.

Finally, despite delivering an opposite result versus hypothesis (No. 7), the study provided an interesting insight regarding how to choose an appropriate image encoder. It showed that a wider and deeper network did not necessarily deliver better captions in attention-based concept. Instead, CNN's output from the last convolutional layer should have a 'just right' shape following the complex of captions. In this case specifically, a vector shape of 64x2048 was the best fit to predict 10-word sentences.

### **5.3. Recommendation**

First and foremost, the study suggested that a stratified sampling approach in validation was necessary to achieve a more accurate evaluation of model prediction in automated image captioning. In a common machine learning procedure where a tuning stage is involved to derive effective hyperparameters, inappropriate split provides incorrect direction for indexes optimisation and hence results in lowering the overall accuracy (Esfahani and Dougherty, 2014).

In new research or business fields where the image data is not robust, or there is no pre-defined caption on the images that requires manual labelling beforehand, this could become particularly important to derive highly accurate predictions with less requirement on sample size.

Second, as the findings showed that stratification delivered consistent performance across 'strata', it would be prosperous in dealing with unbalanced classes situation. Not to mention, by identifying minor topics, re-sampling strategies could be used to obtain the targeted result.

Third, researchers could narrow down their options of encoders in testing phase by examining the complexity of images and sentences. As per finding from this study, CNN's configures should match with the training input. For a simple caption set like MS-COCO, InceptionV3 or EfficientNet up to B3 version (required size = 300 pixels x 300 pixels) would be sufficient for prediction.

Fourth, a potential application recommended from this research is to use topic modelling as a means for image clustering and classification. In traditional prediction, pre-defined classes are required as the input for machine learning. In the image-caption domain, specifically social media content, this could become a challenge due to the enormous dataset and diverse graphical contexts. As this study showed that the images could be segmented using their descriptions, businesses can utilise textual data to define the key topic of photo.

This approach could have 3 major benefits. The most important one would be reducing the cost of classification, due to either manual effort in categorisation or computational expense in processing images. In addition, as topic modelling is an unsupervised task, it requires less or even none of the researchers' domain knowledge in defining the labels. Finally, the method would also be robust in identifying new topics of images, which cannot be easily achieved by using solely graphical data.

#### **5.4. Limitation and future work**

Despite satisfying the objectives and providing insights about the impact of stratified sampling on the image captioning model, the research has 3 major constraints that can be further improved and deep dived in future studies.

Regarding the scope of work, as mentioned in the Introduction, this study only focuses on validating and explaining the impact of stratification on test set accuracy without concerning prediction after the model tuning process. This results in an incomplete understanding of the matter, as hyperparameters adjustment is known for driving the incorrect result even further with inappropriate sampling (Esfahani and Dougherty, 2014). Future studies,

therefore, can investigate this direction based on the initial findings from this research.

Similarly, in examining different model options, the author used a pre-defined set of parameters including the number of most frequent words to keep in caption tokenisation, the batch, buffer size and number of units in neural network layer. While this helped to keep a consistent framework for validation and reduce the time cost, it might not reflect the model capacity with changes in sample size and processing technique.

In model selection, this study is only based on the attention-guided approach and might not generalise the results in other model architectures. As mentioned in the literature review section, image captioning is currently a highly invested field with a large number of new methodologies published yearly. As these techniques vary in structures and algorithms, a potential direction of research is to examine this framework with different benchmarks. This could fuel new insights on models' robustness and their behaviours with seen and unseen data.

In the evaluation stage, despite being a simple and straightforward metric, BLEU highly depends on the referenced texts and therefore did not provide a completely fair evaluation.



**Reference:**

- Two zebras grazing in a nearby grassy plains area
- Two zebra's standing side by side in the grass dirt
- Two zebras standing next together in their den
- Two zebra standing next to each other on a dry grass covered field
- Two zebra standing near each other on a field of grass

**Prediction:** A number of zebras standing in the grass (BLEU = 0.5)

*Figure 5.9. An example of good caption with low BLEU score*

The generated caption in Figure 5.9 showed a good summary of the content with relatively correct grammar. However, since it does not match completely with the words provided in references, its achieved BLEU score is quite low. In contrast, Figure 5.10 shows a caption with a very high BLEU score. However, under manual judgment, it was clearly a bad prediction.



**Reference:**

- The elderly man is looking at the bushels of bananas on a table
- A man bending over the table topped with lots of bananas
- An older man is examining a table of bananas
- An older man is inspecting some bananas at an outdoor market
- An older man standing in front of a table with bunches of bananas

**Prediction:** A bunch of fruit are looking up containers of food on concrete (BLEU = 0.8)

*Figure 5.10. An example of bad caption with high BLEU score*

To solve this problem, researchers can apply other metrics to provide a more comprehensive and accurate assessment of the result. Furthermore, scholars can also study new metrics that highly resemble human judgement towards text's coherency.

## **6. Conclusion**

The development of artificial intelligence opens a new era of automation even in one of the most challenging data realms like image captioning. Several studies have been made to improve the capability of the model and the accuracy of evaluation in this domain. While varying in research areas and applied techniques, most of the previous works focus on producing a new, novel but also compiled methodology and algorithm.

This study examines the impact of a traditional validation technique, stratified sampling, on the effectiveness in prediction by comparing the results of the model under simple random and stratified splits. The findings have proved that despite being a simple and lightweight approach, stratification resulted in an increment of prediction accuracy without the requirement of heavy computation when the sample size is limited. A recommendation on applying this technique in new domains with constraints in database was made to help researchers and businesses achieve a better result with minor effort.

On a side note, this dissertation confirmed the close connection between image and associated caption, which suggested that segmentation tasks on graphical objects can be done through clustering their descriptions. Plus, the relationship between encoder's configures and the complexity of dataset was also unpacked, providing a more straightforward direction for other works.

Leveraging the insights from this study, future works can be done to comprehensively understand the impact of stratification, including but not limited to parameters' adjustment process, different model architectures and evaluation metrics. This would help to further drive the development of the automated image caption field and its contribution to society and business.



## REFERENCE

- Agarwal, A. and Lavie, A. (2008), "METEOR, M-BLEU and M-TER: Evaluation metrics for high-correlation with human rankings of machine translation output", *Proceedings of the Third Workshop on Statistical Machine Translation*, pp. 115–118.
- Alghamdi, R., Alfalqi, K. (2015), "A Survey of Topic Modelling in Text Mining", *International Journal of Advanced Computer Science and Applications* 6, Issue 1.
- Anderson, P., Fernando, B., Johnson, M. and Gould, S. (2016), "SPICE: Semantic propositional image caption evaluation", *European Conference on Computer Vision*, pp. 382–398.
- Bai, S. and An, S. (2018), "A Survey on Automatic Image Caption Generation", *Neurocomputing*, Vol. 311.
- Berry, M. W., Browne, M., Langville, A. N., Pauca, V. P., Plemmons, R. J. (2007), "Algorithms and applications for approximate nonnegative matrix factorization", *Computational Statistics & Data Analysis*, Vol. 52, pp. 155-173.
- Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003), "Latent Dirichlet Allocation", *JMLR*, pp. 993–1022.
- Bridgwater, A. (2018), *The 13 Types Of Data?* [online]. Available at: <https://www.forbes.com/sites/adrianbridgwater/2018/07/05/the-13-types-of-data/?sh=c74355533624> [Accessed 28 Jul 2021].
- Carrington D. (2020), *How many photo will be taken in 2020?* [online]. Available at: <https://blog.myllo.com/how-many-photos-will-be-taken-in-2020/> [Accessed 28 Jul 2021].
- Chang, J., Boyd-Graber, J., Wang, C., Gerrish S., Blei, D. M. (2009), "Reading Tea Leaves: How Humans Interpret Topic Models", *Neural Information Processing Systems* 32, pp. 288-296.



Chawla, R. (2017), *Topic Modeling with LDA and NMF on the ABC News Headlines dataset* [online]. Available at: <https://medium.com/ml2vec/topic-modeling-is-an-unsupervised-learning-approach-to-clustering-documents-to-discover-topics-fdfbf30e27df> [Accessed 6 Sep 2021].

Chen, T.H., Liao, Y.H., Chuang, C.Y., Hsu, W.T., Fu J. and Sun, M. (2017), "Show, Adapt and Tell: Adversarial Training of Cross-domain Image Captioner", *The IEEE International Conference on Computer Vision (ICCV)*, Vol. 2.

Chen, Y., Bordes, J. and Filliat, D. (2017), "An experimental comparison between NMF and LDA for active cross-situational object-word learning", *2016 Joint IEEE International Conference on Development and Learning and Epigenetic Robotics*, pp. 217-222.

Cho, K., Merriënboer, B. V., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H. and Bengio, Y. (2014), "Learning phrase representations using RNN encoder-decoder for statistical machine translation", *2014 Conference on Empirical Methods in Natural Language Processing*, pp. 1724-1734.

Cleartax (2021), *Stratified Random Sampling* [online]. Available at: <https://cleartax.in/g/terms/stratified-random-sampling> [Accessed 6 Sep 2021].

Corliss, R. (2017), *Photos on Facebook Generate 53% More Likes Than the Average Post* [online]. Available at: <https://blog.hubspot.com/blog/tabid/6307/bid/33800/photos-on-facebook-generate-53-more-likes-than-the-average-post-new-data.aspx> [Accessed 28 Jul 2021].

Cornia, M., Baraldi, L., and Cucchiara, R. (2019), "Show, control and tell: A framework for generating controllable and grounded captions", *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 8307-8316.

Daniel, R., Rosen, E., Chuang, J., Christopher Manning, D., and Daniel McFarland, A. (2009), "Topic modeling for the social sciences," *Workshop on Applications for Topic Models: Text and Beyond*.

Esfahani, M. S. and Dougherty, E. (2014), "Effect of separate sampling on classification accuracy", *Bioinformatics*, Vol. 30, pp. 242-250.

Evergreen (2020), *Automatic Image Captioning Using Neural Networks* [online]. Available at: <https://evergreen.team/articles/automatic-image-captioning.html> [Accessed 28 Jul 2021].

Facebook (2021), *How Facebook is using AI to improve photo descriptions for people who are blind or visually impaired* [online]. Available at: <https://tech.fb.com/how-facebook-is-using-ai-to-improve-photo-descriptions-for-people-who-are-blind-or-visually-impaired/> [Accessed 28 Jul 2021].

Farhadi, A., Hejrati, M., Sadeghi, M.A., Young, P., Rashtchian, C., Hockenmaier, J. and Forsyth, D. (2010), "Every picture tells a story: Generating sentences from images", *European conference on computer vision*, pp. 15–29.

Fu, Y. (2020), *Image classification via fine-tuning with EfficientNet* [online]. Available at: [https://keras.io/examples/vision/image\\_classification\\_efficientnet\\_fine\\_tuning/](https://keras.io/examples/vision/image_classification_efficientnet_fine_tuning/) [Accessed 6 Sep 2021].

Hofmann, T. (1999), "Probabilistic latent semantic indexing", *Proceedings of the Twenty-Second Annual International SIGIR Conference*, pp. 50-57.

Hossain, M. Z., Soheli, F., Shiratuddin, M.F. and Laga, H. (2018), "A Comprehensive Survey of Deep Learning for Image Captioning", *ACM Computing Survey*, Art. 0.

Hrga, I. and Ivašić-Kos, M. (2019), "Deep Image Captioning: An overview", *42nd International Convention on Information and Communication Technology, Electronics and Microelectronics, MIPRO 2019 – Proceedings*, pp. 995-1000.

Huilgol, P. (2018), *Top 4 Pre-Trained Models for Image Classification with Python Code* [online]. Available at: <https://www.analyticsvidhya.com/blog/2020/08/top-4-pre-trained-models-for-image-classification-with-python-code/> [Accessed 6 Sep 2021].

Jolliffe, I. T. (1986), "Principal Component Analysis and Factor Analysis", *Principal Component Analysis*, pp. 115-128.

Karim, R. (2019), *Illustrated: 10 CNN Architectures* [online]. Available at: <https://towardsdatascience.com/illustrated-10-cnn-architectures-95d78ace614d#c5a6> [Accessed 6 Sep 2021].

Karpathy, A. and Fei-Fei, L. (2015), "Deep Visual-Semantic Alignments for Generating Image Descriptions", *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3128-3137.

Khare, A. and Huber, M. (2019), "Show, Infer and Tell: Contextual Inference for Creative Captioning", *30th British Machine Vision Conference 2019*.

Klos, A. (2020), *Topic modeling: LDA vs. NMF for newbies* [online]. Available at: <https://alexklos.ca/blog/natural-language-processing-lda-vs-nmf-for-newbies/> [Accessed 6 Sep 2021].

Lin., C.Y. (2004), "ROUGE: A package for automatic evaluation of summaries", *Text summarization branches out: Proceedings of the ACL-04 workshop*, Vol. 8.

Mccallum, A. K. (2002), *MALLET: A Machine Learning for Language Toolkit* [online]. Available at: <http://mallet.cs.umass.edu/> [Accessed 28 Jul 2021].

Meng, X. (2013), "Scalable simple random sampling and stratified sampling", *30th International Conference on Machine Learning*, Vol. 28, pp. 1568-1567.

Menon, S. (2020), *Stratified sampling in Machine Learning?* [online]. Available at: <https://medium.com/analytics-vidhya/stratified-sampling-in-machine-learning-f5112b5b9cfe> [Accessed 28 Jul 2021].

Mishra, M. (2020), *Convolutional Neural Networks, Explained* [online]. Available at: <https://towardsdatascience.com/convolutional-neural-networks-explained-9cc5188c4939> [Accessed 6 Sep 2021].

Ordonez, V., Kulkarni, G. and Berg, T.L. (2011), "Im2text: Describing images using 1 million captioned photographs", *Advances in Neural Information Processing Systems*, pp. 1143-1151.

Papineni, K., Roukos, S., Ward, T. and Zhu, W. (2002), "BLEU: A method for automatic evaluation of machine translation", *Meeting on Association for Computational Linguistics*, Vol. 4.

Phi, M. (2018), *Illustrated Guide to LSTM's and GRU's: A step by step explanation* [online]. Available at: <https://towardsdatascience.com/illustrated-guide-to-lstms-and-gru-s-a-step-by-step-explanation-44e9eb85bf21> [Accessed 6 Sep 2021].

Prabhakaran, S. (2018), *Topic Modeling with Gensim (Python)* [online]. Available at: <https://www.machinelearningplus.com/nlp/topic-modeling-gensim-python/> [Accessed 6 Sep 2021].

Provost, F. and Fawcett, T. (2013), *Data Science for Business*, O'Reilly Media, Sebastopol.

Ramezan, C. A., Warner, T. A. and Maxwell, A. E. (2019), "Evaluation of sampling and cross-validation tuning strategies for regional-scale machine learning classification", *Remote Sensing*, Vol. 11, Issue 2.

Rao, D. (2019), *Stratified Sampling – Machine Learning* [online]. Available at: <https://medium.com/@dhivyarao94/stratified-sampling-machine-learning-b622189ae77> [Accessed 6 Sep 2021].

Rehurek, R. (2010), "Software framework for topic modelling with large corpora", *Proceedings of LREC 2010 workshop New Challenges for NLP Frameworks*, pp. 46–50.

Saha, S. (2018), *A Comprehensive Guide to Convolutional Neural Networks – the ELI5 way* [online]. Available at: <https://towardsdatascience.com/a-comprehensive-guide-to-convolutional-neural-networks-the-eli5-way-3bd2b1164a53> [Accessed 6 Sep 2021].

Sarkar, S. (2020), *Image Captioning using Attention Mechanism* [online]. Available at: <https://medium.com/swlh/image-captioning-using-attention-mechanism-f3d7fc96eb0e> [Accessed 6 Sep 2021].

Shearer, C. (2000), "The CRISP-DM Model: The New Blueprint for Data Mining", *Journal of Data Warehousing* 5, pp. 13-22.

Tan, M., and Le, Q.V. (2019), "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks", *ArXiv*, *abs/1905.11946*.

TensorFlow (2018), *Image captioning with visual attention* [online]. Available at: [https://colab.research.google.com/github/tensorflow/docs/blob/master/site/en/tutorials/text/image\\_captioning.ipynb](https://colab.research.google.com/github/tensorflow/docs/blob/master/site/en/tutorials/text/image_captioning.ipynb) [Accessed 6 Sep 2021].

Thompson, S. K. (2012), *Sampling* (3rd ed.), John Wiley & Sons, New Jersey.

Ushiku, Y., Yamaguchi, M., Mukuta, Y. and Harada, T. (2015), "Common subspace for model and similarity: Phrase learning for caption generation from images", *IEEE International Conference on Computer Vision*, pp. 2668–2676.

Vabalas, A., Gowen, E., Poliakoff, E. and Casson, A. J. (2019), "Machine learning algorithm validation with a limited sample size", *PLoS ONE*, Vol. 14, pp. 1-20.

Vedantam, R., Zitnick, C.L. and Parikh, D. (2015). "CIDER: Consensus-based image description evaluation", *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4566–4575.

Wang, C., Zhou, Z. and Xu, L. (2021), "An Integrative Review of Image Captioning Research", *Journal of Physics: Conference Series*, Vol. 1748.

Wikipedia (2021), *Topic model* [online]. Available at: [https://en.wikipedia.org/wiki/Topic\\_model](https://en.wikipedia.org/wiki/Topic_model) [Accessed 6 Sep 2021].

Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., Zemel, R. and Bengio, Y. (2015), "Show, Attend and Tell: Neural image caption generation with visual attention", *International Conference on Machine Learning*, pp. 2048–2057.

Yang, Y., Teo, C.L., Daume, H. and Aloimono, Y. (2011), "Corpus-guided sentence generation of natural images", *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 444–454.