Synapse Task 3

3.1
Firstly one would need some good quality camera to capture the different wildlife.We can then collect a diverse dataset containing the different animals and birds!This data can be used used to train the model.
The data needs to be preprocessed for more accuracy.The various animals can be zoomed into to get a much clearer picture of them.With the help of ML,you can even increase resolution of pictures .This will be helpful in training in the model as there won't be too much noise.
We will be using convolutional neural networks for image classification.
CNN is most effective for image classification ,object detection, etc.This would be extremely helpful to even identify animals or bird who are there in the frame for even few seconds.
Pre-trained CNN models like VGG, ResNet, or Inception can be used as feature extractors.
We can fine-tune the pre-trained CNN by replacing the final classification layer and training on our dataset.

Evaluate the trained model on the validation set to assess its performance. Metrics like accuracy, precision, recall, and F1-score can help quantify how well the model is identifying species.
F1 Score:The F1 score is the harmonic mean of precision and recall

We should consider ethical and privacy aspects, especially when dealing with wildlife. Make sure that data collection and model deployment do not disturb or harm the animals in any way.
In summary, this approach combines the power of deep learning (CNNs) for feature extraction and traditional machine learning techniques for classification.

3.2

Collect translations of the same text in multiple languages.This will help in developing more data for the model to learn on.For translations RNNs are the best.(Recurrent Neural Networks)
RNNs can help with word relationships to maintain structure and meaning.
Many Languages don't use the subject verb object form and hence when translating, the literal translation may not make sense.
After processing sentences can be reordered for better structure and to be more meaningful.
Sequence to sequence:The core technique for word conversion is the use of sequence-to-sequence models. These models take a sequence of words from the source language and generate a sequence of words in the target language.
Subward Tokenization:Languages often have varying word structures, so

subword tokenization is used. Words are broken down into smaller meaningful units like subword pieces or characters. This is especially useful for handling morphologically rich languages and dealing with out-of-vocabulary words.

We can better these methods by having actual meaning of various idioms in different languages than their literal meanings.We must also take into account the local jargon used(if any) which can help fine tune our language model?
In summary Deep learning, subword tokenization, and sequence-to-sequence models improve translation quality.

3.3
Identifying spam texts from unlabelled data is challenging and often tackled through unsupervised methods.
We can clean the text by removing all the ascii characters.We can remove all emojis for better simplicity in analysing data.We can apply clustering methods (e.g., K-Means, DBSCAN) on feature vectors to group similar texts.Hybrid approaches or incorporating labeled data could enhance robustness and accuracy.
*Tokenization:** Split the text into individual words (tokens). Tokenization is a crucial step as it breaks down the text into manageable units for analysis.

Inspect the clusters and identify the ones that seem to have spam-like characteristics. Spam messages often exhibit patterns such as excessive use of certain words, promotional language, or links.

While unsupervised methods can provide valuable insights, they might lack the precision of supervised approaches due to the absence of labeled data for validation.