Tamil Language Image Annotation Using Semantic Order Learning And Attention Based Neural Machine Translation

Gokkul Anbazhagan[1], Saairam Venkatesan[2], Abishiek Srinivasan[3]
[1, 2, 3]College Of Engineering Guindy

Abstract- *Nowadays, picture and sentence matching has made tremendous strides, but because of the current broad visual-semantic gap, it remains difficult. The words in the phrases are structured in a grammatical way sequentially, while the semantic meanings in the images are typically unorganised. Generally, the image and phrase matching task relates to calculating the visual-semantic resemblance between an image and a phrase. Here, the words in the phrases are structured in a grammatical way sequentially, while the semantic meanings in the images are typically unorganised. Thus, semantic concepts are suggested and an order learning system for the matching of images and phrases in this work, which can enhance the representation of images by first predicting semantic concepts and then arranging them in the correct semantic order. Further the sentence generated is translated to Tamil language since there is no image annotation system for the same.*

I. INTRODUCTION

Automatically generating captions to an image shows the understanding of the image by computers and is a fundamental task of intelligence. This caption model not only finds which objects are contained in the image but also expresses their relationships in a natural language, English and translates them to Tamil.

Image captioning methods can also be expanded by first generating the sentence given to an image and then comparing the sentence produced with ground truth one to deal with image-sentence matching. However these models are not very good. Unlike them, the project's work focuses on the calculation of resemblance, which is particularly appropriate for picture and sentence matching tasks . Additionally, the sen- tence generated for the given image is language dependent i.e English and there are no diverse datasets for image - sentence matching in Tamil language. To remove this barrier, a neural machine translation system is proposed with an attention based approach to translate sentences to Tamil language. So the objective is to generate sentences from the given image, using Semantic feature extraction and to translate sentences in Tamil language from English using the Neural Machine Translation System with an attention based approach.

II. SYSTEM DESIGN

The image is passed through two separate ResNet-152 mod- els. One model is pre-trained on ImageNet dataset and is used to retrieve the global context features and the feature maps from the image. The other model trained on the MSCOCO dataset retrieves the semantic concepts from the Image. These values from the models are vectors which are passed on to the gated-fusion module. This module calculates the weightage needed to be given to the global context and the semantic concepts. The output of this module is used to train the LSTM which sees the different parts of the image and where it needs to attend based on the gated-fusion output. Once the model is trained, the weight file is used to generate sentences for further test images. This trained model is used to generate the image annotation based on the conventional LSTM as in Figure 3.1.

The generated sentence is given to the Neural Machine Translation Model. This model is built using an Encoder- Decoder architecture with attention. In order to improve the correctness of translation, the model is trained on a parallel corpora preprocessed using Byte Pair Encoding and Word Em- bedding Techniques. The model accepts the English sentence, performs Beam-Search and returns a sentence in Tamil.

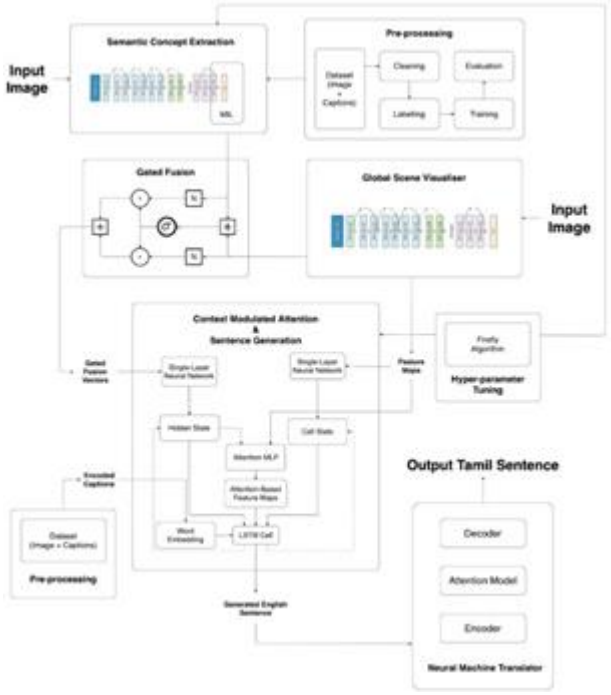

Fig. 1. Overall architecture diagram of the mode

III. MODULE DESIGN

The overall framework of the project is shown in the block diagram, which consists of a preprocessor, an extractor, a generator and a translator. The model is first trained and then the entire network is trained end-to-end as shown in Figure 1.

The modules used are as follows:
1) Preprocessing
2) Semantic Feature Extraction
3) Global Scene Visualizer
4) Gated Fusion
5) Context Modulated Attention
6) Neural Machine Translation with Attention

A. PREPROCESSING

Tokenizing the captions and Tagging their corresponding part of speech. Map the image dataset with the corresponding semantic concepts. Finetune the pretrained ResNet using py- torch framework and MIL layer. Store the detections and the benchmarks(precision and recall) in a pickle file.

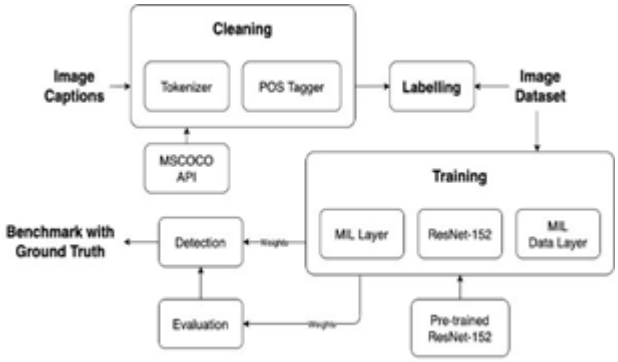

Fig. 2. Preprocessing architecture diagram of the model

B . SEMANTICFEATUREEXTRACTION

For images, their semantic conceptsreferto various objects, properties and actions, which are represented in the visual content. To learn such a model, a training dataset is built man- ually keeping only the nouns, adjectives, verbs and numbers as
semantic concepts, and eliminating all the semantic-irrelevant words from the sentences. Considering that the size of concept vocabulary could be very large, words are ignored having very low use frequencies. In addition, the different tenses of verbs, and the singular and plural forms of nouns to further reduce the vocabulary size. Finally, a vocabulary containing K semantic concepts is obtained. Based on this vocabulary, a training dataset is generated by selecting multiple words from sentences as the ground truth semantic concepts. To extract the concepts from the image, ResNet is used which is pretrained on the ImageNet dataset as multi-label CNN. This ResNet uses the trained vocabulary in extracting the semantic concepts.

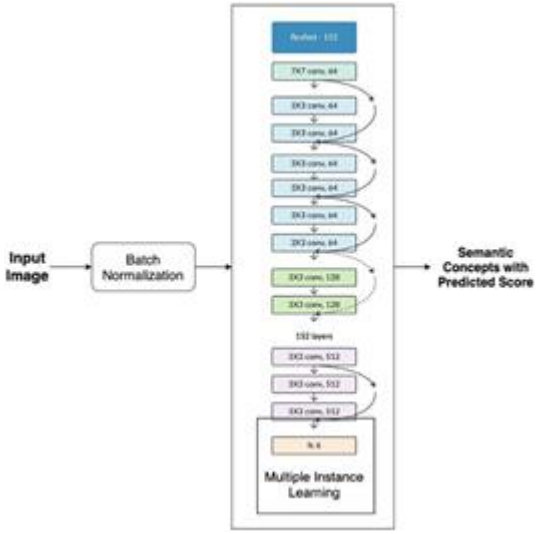

Fig. 3. Semantic feature extraction architecture of the model

C. GLOBAL SCENE VISUALISER

This module selects and sorts the concepts, in the level of concept-related image regions. The semantic concepts are used in a more "soft" manner, by correlating them with the original image to find related regions. In this way, the incorrectly predicted concepts can also find the right image regions due to their similar appearances, so that the model can largely tolerate the potential errors.

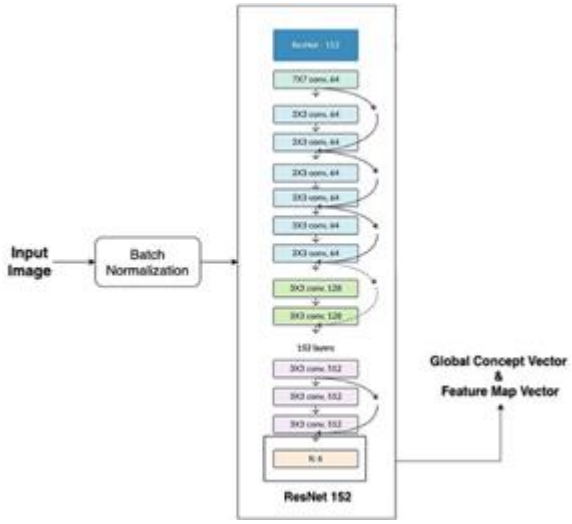

Fig. 4. Global scene visualizer architecture diagram of the model

D. GATED FUSION

Gated fusion is a gated combination of semantic concepts and global scene. There are two main reasons for introducing the global scene as follows. 1) It is uneasy to decide the semantic order only from separated semantic concepts during attention, since the order involves not only the hypernym relations between concepts, but also the textual entailment among phrases in high levels of semantic hierarchy. Gated Fusion Unit is the core of context modulation is the proposed gated fusion of semantic concepts and global scene, so it would be interesting to qualitatively analyze which images focus more on concepts or scene.

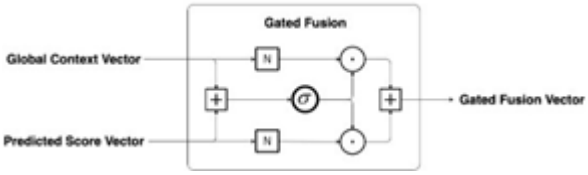

Fig. 5. Gated fusion architecture diagram of the model

E . CONTEXT MOUDLATED ATTENTION

This module computes the weighted sum representations to adaptively describe the attended image regions. To aggregate these regions for the whole image representation, LSTM network is used to sequentially take them as inputs, where the hidden states dynamically propagate the representations of image regions until the end. The LSTM includes various gated mechanisms which can well suit the complex nature of semantic order. The hidden state at the last timestep can be regarded as the desired image representation with semantic order. the context-modulated attentional LSTM, which se- lectively attends to multiple image regions by predicting a sequence of concept-related attention maps. It then explicitly aggregates their representations in the sequential manner of LSTM, in which the sequential order can be regarded as the desired semantic order for concepts. context-modulated atten- tional procedure, the information in the initial attention map is additively modulated by the image context and subtractively modulated by the previous hidden state, to finally produce the concept related attention map.

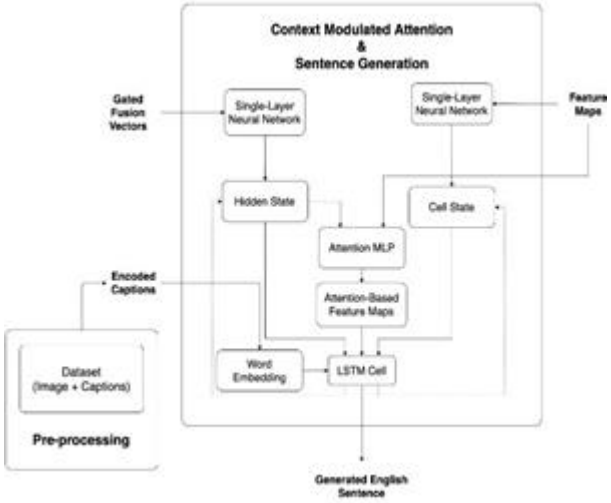

Fig. 6. Context modulated attention architecture of the model

F . NEURAL MACHINE TRANSLATION

The generated English sentence is received by this module . This module translates the sentence to desired Tamil language by using NMT. Attention mechanism is used to improve the performance of Neural Machine Translation by selectively focusing on sub-parts of the sentence during translation.It outputs the translated sentence.

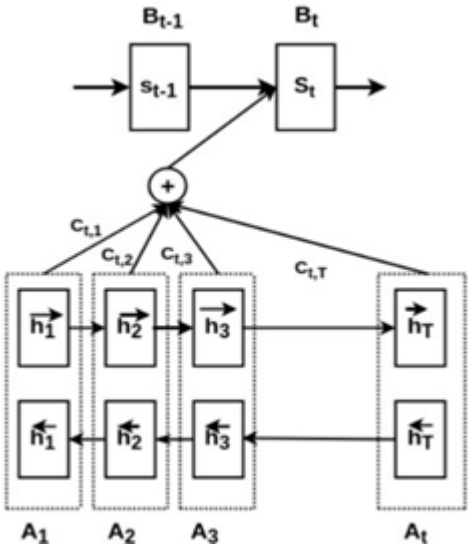

Fig. 7. Neural Machine Translation architecture of the model

IV. RELATED WORK

Image Annotation is a well-known problem over the years and various different methods have been suggested for the same.All these methods have tried to improve the BLEU - 4 score with improved accuracy. However,Neural Machine Translation for English-Tamil is relatively new and has fewer resources.Following section covers in detail the different works performed in these domains. matching as well. A neural image captioning generator and show the effectiveness of the image and sentence matching. However, these models are originally designed to predict grammatically-complete sentences, so their performance on measuring the image-sentence similarity is not very good. Different from them, the project's work focuses on the similarity measurement, which is especially suitable for the task of image and sentence matching.

A. VISUAL - SEMANTICEMBEDDINGMETHODS

The work proposes a semantic-enhanced image and sentence matching framework [1] , where semantic concepts and order can be effectively learnt by multi-regional multi-label CNN and context-modulated attentional LSTM [1] , respectively. A neural machine translation system [2] is proposed with an attention based approach to translate sentences to Tamil language. The first visual-semantic embedding framework [3], in which ranking loss, CNN and Skip-Gram are used as the objective , image and word encoders, respectively. Under the similar framework, replace the Skip-Gram [3] with LSTM for sentence representation learning, use a new objective that can preserve the order structure of visual- semantic hierarchy additionally consider within-view constraints to learn structure-preserving representations. The image and sentence using deep canonical correlation analysis [5] as the objective, where the matched image-sentence pairs have high correlation. Based on the similar framework. use Fisher Vectors (FV) [4] to learn more discriminative representations for sentences, and alternatively use RNN to aggregate FV and further improve the performance.In addition to the global matching methods, make the first attempt to perform local similarity learning between fragments of images and sentences with a structured objective. Exploit a multimodal CNN [4] for matching image and sentence, which can be regarded as an end-to-end framework for similarity score prediction. In contrast to them, semantic concepts are extracted from local image regions and then learn their semantic order.

B. IMAGE CAPTIONINGBASED METHODS

Image captioning methods can also be extended to deal with image-sentence matching, by first generating the sentence given an image and then comparing the generated sentence with ground truth one. A multimodal auto-encoder [5] for bidirectional mapping, and measure the similarity using the cross-modal likelihood and reconstruction error. A multimodal RNN [4] model to generate sentences from images, in which the perplexity of generating a sentence is used as the similarity. A long-term recurrent convolutional network for image captioning, which is also extended to image and sentence

C. DEEP ATTENTION BASED METHODS

The proposed model is also related to some methods simulat- ing visual attention . RNNs [5] and differentiable Gaussian filters to simulate the attention mechanism, and apply it to handwriting synthesis. The deep recurrent attentive writer for image generation, which develops a novel spatial attention mechanism based on 2- dimensional Gaussian filters [7] to mimic the foveation of human eyes. Present a recurrent attention model that can attend to some label-relevant image regions of an image for multiple objects recognition. A neural machine translation which can search for relevant parts of a source sentence to predict a target word. An attention-based model [7] which can automatically learn to fix gazes on salient objects in an image and generate the corresponding annotated words. Different from these models, this work focuses more on the modelling of context information during attention to compensate for the lack of semantic information, and proposes context-modulated attention to find concept-related image regions for semantic order learning.

D. FROM CAPTIONS TO VISUAL CONCEPTS

The word detector outputs serve as conditional inputs to a maximum-entropy language model [3] . The language model learns from a set of over 400,000 image descriptions to capture the statistics of word usage. Global semantics is captured by re-ranking caption candidates using sentence-level features and a deep multimodal similarity model [7] . The system is producing a BLEU-4 score of 29.1%. When human judges compare the system captions to ones written by other people on a held-out test set, the system captions have equal or better quality 34% of the time.

E. LONG SHORT-TERM MEMORY

Learning to store information over extended time intervals by recurrent backpropagation takes a very long time, mostly because of insufficient,decaying error backflow. Hochreiter's analysis [7]of this problem is briefly reviewed, then addressed by introducing a novel, efficient, gradient based method called long short-term memory (LSTM). Truncating the gradient where this does not do harm, LSTM [7] can learn to bridge minimal time lags in excess of 1000 discrete-time steps by enforcing constant error flow through constant error carousels within special units. Multiplicative gate units learn to open and close access to the constant error flow. LSTM is local in space and time; its computational complexity per time step and weight is O. 1. The experiments with artificial data involve local, distributed, real- valued, and noisy pattern representations. In comparisons with real-time recurrent learning, backpropagation through time, recurrent cascade correlation, Elman nets, and neural sequence chunking, LSTM [7] leads to many more successful runs, and learns much faster. LSTM also solves complex, artificial long-time-lag tasks that have never been solved by previous recurrent network algorithms
[5] .

F . MACHINE TRANSLATION

The chapter reviews the current state of research, development, and use of machine translation (MT) systems [8] . The empirical paradigms of example-based MT [2] and statistical MT [9] are described and contrasted with the traditional rule-based approach. Hybrid systems [9] involving several approaches are discussed. Two recent developments within the rule-based paradigm are discussed, namely, anaphora resolution for MT and interlingual and knowledge-based MT. As a major new application, spoken language MT [8] is introduced. The prospect of MT systems [2] for minority and less developed languages is discussed, along with the use of MT [8] on the Internet, notably for web-page translation. Finally, tools for translators are described, particularly those which exploit bilingual parallel corpora [2] (translation memories, bilingual concordances), as well as translator oriented word-processing tools.

V. IMPLEMENTATION

MSCOCO - Microsoft Common Objects in Context actually is used as one of the datasets for this objective in a particular major way. MSCOCO consists of 82783 training and 40504 validation images, each of which is associated with 5 sentences, which really is quite significant. 1000 images for testing are used, and generally perform 5- essentially fold cross-validation and report the averaged results. In general, the MSCOCO dataset basically is a JSON file and it is not sort of possible to literally carry it for the further processes. So, this JSON file is fed as an input for the COCO API, which essentially is a fairly large image dataset designed for object detection, segmentation, person keypoints detection, stuff segmentation, and caption generation in a subtle way. Finally, a file with desired annotations is retrieved and NLTK is implemented.Here, the tokenizer actually is implemented to separate the auxiliary verbs like ‘is’, ‘are’, ‘was’, ‘were’ from the main semantic concepts. For all intents and purposes, auxiliary verbs for the most part are removed and the POS tag is implemented. POS Tag - Part Of Speech tagger, processes a sequence of words, and attaches a part of speech tag to each word, or so they specifically thought. They particularly are used to categorize the nouns, verbs, adjectives and numbers from the sentences and tag themselves with the respective words in the sentences in a generally major way. Once the stemming and cleansing process essentially is done, the output produces a PICKLE file, containing a set of dictionary-like structures. NLTK- Natural Language Toolkit is a suite of libraries and programs for symbolic and statistical natural language processing for English written in the Python programming language, which particularly is quite significant. NLTK generally is used for two purposes : Tokenizing and implementing POS Tag in a fairly big way. The tokenizer in NLTK is used to break up a piece of text into smaller parts, fairly such as sentences and words.

The use of the PICKLE file is to actually retrieve the word count from it. This file contains thousands and thousands of words, sort of clubbed together in a messy arrangement, not matching the words to their particularly corresponding images. Thus, the PICKLE file is sorted and retrieves the top 1000 most occurring words and uses those words for the vocabulary training for producing a high accuracy rate while testing and predicting the model in a subtle way. Then the Pytorch is used to implement ResNet 152 to retrieve the desired semantic concepts in a generally major way. ResNet 152 basically is a convolutional neural network having 152 deep layers. A pre-trained version of the network is loaded and trained on more

than a million images from the ImageNet database. The pretrained network can generally classify images into 1000 object categories. In addition to this, MIL is introduced in this model, which for all intents and purposes is quite significant. MIL - Multiple Instance Learning particularly is a type of supervised learning which receives a set of instances that kind of are individually labeled, and the learner receives a set of labeled bags, each containing many instances. It generally divides the multiple segments and distinguishes them individually in a subtle way. ResNet 152 basically uses Softmax function and generates the next corresponding hidden layer. But in this case, the Sigmoid function is used to generate the next corresponding hidden layer in a kind of major way. 1000 pre- trained images from the ImageNet database are loaded again, replacing the existing ones and written to a HDF5 file (Hierarchical Data Format Version 5) for training the model.

Finally, the HDF5 file, the corresponding dataset images and pre-trained weight file is fed to the pytorch model, producing a weight file showing weights of each semantic concept with corresponding precisions. This weight file cannot be used for predicting the semantic concepts for the desired image, since they are to be tested. So this file is tested by feeding into the pytorch model again, with the corresponding dataset images, to make another PICKLE file, and this new file is used for predicting processes.Thus, when the desired image essentially is loaded into the model, the actual corresponding semantic concepts for all intents and purposes are displayed with the respective prediction scores in a fairly major way. To extract the global scene concepts and feature map of the image, another ResNet model pre-trained on ImageNet is used . Specifically, the last fully connected layer is used to extract the global scene concept of the image, and the feature map of the image. After obtaining the global scene vector x and concept score vector p, their gated fusion can be formulated as:

$$\hat{\mathbf{x}} = \|W_x\mathbf{x}\|_2,\ \hat{\mathbf{p}} = \left\|W_p\mathbf{p}\right\|_2,\ \mathbf{t} = \sigma(U_x\mathbf{x} + U_p\mathbf{p})$$
$$g(\mathbf{x},\mathbf{p}) = \mathbf{t}\odot\hat{\mathbf{x}} + (\mathbf{1}-\mathbf{t})\odot\hat{\mathbf{p}},$$

The Attention network is composed of only linear layers and a couple of activations. Separate linear layers transform both the encoded image (flattened to 4096, 14 * 14, 512) and the hidden state to the same dimension, viz. the Attention size. They are then added and ReLU activated. A third linear layer transforms this result to a dimension of 1, whereupon softmax is applied to generate the weights. A long short-term memory (LSTM) network is implemented that produces a caption by generating one word at every time step conditioned on a context vector, the previous hidden state and the previously generated words. English-Tamil parallel corpus is taken. Using OpenNMT-py, the corpus is cleaned and tokenized . Byte-Pair encoding and word- embedding techniques are employed to obtain a final dataset. The corpus is trained using an encoder decoder architecture with attention to obtain a model. This model receives the english caption and translates to corresponding Tamil annotation of the image.

VI. CONCLUSION

Thus an Image annotation model is built which can understand semantic concepts in a better way and knows what concepts to attend at different intervals of time. It knows how much global information must be taken to generate a particular sentence. An Neural Machine Translation system is built based on Encoder- Decoder Architecture with attention. Byte Pair Encoding and Word Embedding techniques are employed to improve the accuracy of translation. This model is used to translate sentences from English to Tamil Language. In order to handle out of vocabulary concepts, a feedback system can be implemented. The network asks the user to rate the translation and provide suggestions in case of error in sentence generation. This can help the system to improve its accuracy and also can make it generate personalised captions for the user.

REFERENCES

[1] Choudhar Himanshu, Pathak Aditya Kumar, Saha Rajiv Ratan and Kumaraguru Ponnurangam., “Neural Machine Translation for English-Tamil,” in Proceedings of the Third Conference on Machine Translation (WMT), Volume 2: Shared Task Papers, pages 770–775 Belgium, Brussels, October 31 - November 1, 2018. https://www .aclweb.org/anthology/ W18-6459/

[2] Fang H, S. Gupta, F. Iandola, R. K. Srivastava, L. Deng, P. Dollar, J. Gao, X. He, M. Mitchell, J. C. Platt, et al., “From captions to visual concepts and back,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2015, pp. 1473–1482. htt ps:// www.researchgate.net/publication/30 8809117_From_cap tion- s_to_visual_concepts_and_back

[3] Hochreiter S and J. Schmidhuber, “Long shortterm memory,” Neural Compute., vol. 9, no. 8, pp. 1735–1780, 1997.
https://dl.acm.org/doi/10.1162/neco.1 997.9.8.1735

[4] Nadeem Jadoon Khan, Waqas Anwar, and Nadir Durrani. 2017. Machine translation approaches and surveys for indian languages. arXiv preprint arXiv:1701.04290 https://arx iv.org/abs/1701.04290

[5] Parth Shah , Vishvajit Bakrola, IEEE., “Neural Machine Translation System of Tamil Languages - An Attention based Approach,” in Proc. IEEE Conf. Comput. Vis.

Pattern Recognit., 2019 https://ieeexplore.ieee.org/document/8882969

[6] Simonyan K and A. Zisserman, "Very deep convolutional net- works for large-scale image recognition," in Proc. Int. Conf. Learn. Representations, 2014. Available at , https:// arxiv.org/pdf/1409.1556.pd f Somers H, "Machine translation: latest developments," in The Oxford handbook of computational linguistics. https://personalpages.manchester.ac .uk/staff/harold. somers/Mitkov-book-chapter.pdf

[7] Wei Y, W. Xia, J. Huang, B. Ni, J. Dong, Y. Zhao, and S. Yan, "CNN: Single-label to multi-label," arXiv:1406.5726, 2014. Available at , https://arxiv.org/ abs/1406.5726 Yan Huang , Qi Wu , Wei Wang, and Liang Wang, IEEE., "Image and Sentence Matching via Semantic Concepts and Order Learning," in Proc. IEEE Transactions on Pattern Analysis and Machine Intelligence (Volume: 42, Issue: 3, March 1 2020) https://ieeexplore.ieee.org/document/8550752