An Approach for Devising Stenography Application Using Cross Modal Attention

Shanthalakshmi M.[1,*], Susmita Mishra[1], LincyJemina S.[1] ,Raashmi P.[1], Mannuru Shalin[1], Jananeee V.[1]

[1]Rajalakshmi Engineering College, Panimalar Institute of Technology, India
Emails: shanthalakshmi.m@rajalakshmi.edu.in; susmitamishra12@gmail.com; lincypit@gmail.com; raashmi.p.2018.cse@rajalakshmi.edu.in; mannuru.shalini.2018.cse@rajalakshmi.edu.in; jananee.v@rajalakshmi.edu.in

Abstract

This paper focuses on providing a solution to the direct conversion of speech to shorthand. Since shorthand is not understood by many but is used for writing quick transcripts, a product is developed that converts the speech to its appropriate Gregg shorthand. A website that will be used as a front end, will use a speech-to-text API to record the speech in real-time. The converted text will then be fed into a text-to-image retrieval model that derives its corresponding Gregg shorthand for the text. The text will then be displayed to the user in real-time. By achieving this, the model reduces the need to depend upon stenographers for transcribing scripts. The resulting model achieves a good result.

Keywords: Devising Stenography; Cross Modal Attention; Speech shorthand; Speech conversion

1.Introduction

Stenography is derived from the Greek words “steno” and “graphie” which means “narrow” and “writing” respectively. Stenography is an art of writing that enables spoken words to be written down quickly using symbols and abbreviations. It is more commonly known as “shorthand writing”. Stenography is widely used to record oral speech in written scripts. It is most commonly used in courtrooms where there is a necessity to record all the dictation, and in hospitals by doctors to write down long and legal medical terms in the shortest way possible. Shorthand usually consists of symbols, abbreviations, and patterns to depict words. Each word has its own individual pattern to recognize them uniquely. There usually exists a stenographer where short handwriting is followed. A Stenographer is the one who is responsible for transcribing the oral speech into written speech and for translating the shorthand into normal text. If a stenographer does not exist, it is hard to translate it into shorthand. So there is a need to convert speech to shorthand without an intermediary in between.

To prevent this, we introduced our model that converts live speech to Gregg shorthand. Initially, to record the narration that requires conversion, we use pre-existing APIs like Google API, Azure’s speech-to-text services that are highly effective in capturing words accurately. Then, we use our pre-trained models. We will use any neural network model which will provide semantic consistency for image-to text matching. The text that was generated using speech will be stored in a text file. The words are then fed into the models one by one. Since Gregg shorthand encompasses various lines and strokes to denote the words, the models then find the closest Gregg shorthand image that it could find and will give the output. The set of images will then be displayed on the screen as sentences. Our work not only focuses on converting text to its shorthand but instead provides a solution that converts speech that can be recorded in real-time, which is stored to generate its appropriate shorthand

Doi : https://doi.org/10.54216/JCHCI.030105
Received: January 15, 2022 Accepted: May 26, 2022

2.Related work

The existing models provide solutions that convert shorthand to English text. The thickness and length of various shorthand languages vary significantly. So the current existing model [1] uses OCR (Optical Character Recognition) to recognize written characters and legal terminologies. It has a recognition system that recognizes Gregg shorthand using a Raspberry Pi and a transcription system that converts them to English by using the Inception-v3 Convolutional Neural Network algorithm in TensorFlow. The Raspberry Pi model uses a camera to capture images which are then sent to Raspberry Pi 4 for processing and transcribing. The output is then generated and displayed on a screen, while the user also has the option to save it in a .csv file. However, there seems to be no existing model that converts text directly to shorthand in order to reduce the need for having a stenographer whenever an interpretation is needed. Hence we propose a speech-to-shorthand system that converts real-time speech into Gregg shorthand. The model records the speech, converts it into text in real-time, and then finally processes the text into Gregg shorthand.

3.Step-Wise Hierarchical Alignment Network for Image-Text Matching (SCAN)

Advanced work on image-text matching has shown that by using a cross-attention mechanism for matching text-image pairs, we can provide enough hints to the model for matching image-text pairs. The work done employed by implementing this mechanism [2] shows that using a stepwise hierarchical alignment network, it is possible to more accurately output an appropriate image for a text. This is done by executing cross-modal alignment using three stages of alignment(local-to-local, local-to-global, global-to-global) on two semantic levels, namely fragment level and context level. The word features are extracted by using any word embedding model for producing vectors and a bidirectional GRU to construct information from text. This model provides more textual information by exploring locally and globally, thereby leading to the precise retrieval of images.

4.Cross-Modal Attention with Semantic Consistency (CASC)

Great progress has been made in text-to image mapping. Cross-Modal Attention with Semantic Consistency (CASC), a hybrid approach for image-text mapping, has been proposed[3]. In the CASC method, local alignment is between the image regions and text words mutually from both text-image and image-text directions. Rather than global features, it uses multi-label classification for the images. The goal of the CASC method is to measure the similarity between the image and text and match the relevant instances in a more comprehensive representation, whereas the previous methods represented them in a common subspace. The features of the images were extracted using pre-trained Faster R-CNN with ResNet-101 on the Visual Gnomes dataset. RNN model is employed to extract the word features to embed the words along with the context. Bidirectional GRU uses semantic vector space for mapping image-text vectors. The previous SCAN method is effective and maps the image and text with a local stacked cross attention mechanism. But this CASC method achieves local alignment between image patches and words to achieve similarity between image-text pairs. This CASC method can be discussed in two parts. The first is an image to text attention. In this approach, the image region is compared to the corresponding word vector. The second is the text to image attention; in this approach the importance of each word is determined to each image vector. The main aim of this method is to incorporate local alignment between images and words. In this CASC model, the label vocabulary is built where meaningful words are extracted as labels for images. The vectors of images and words are used to generate global features. Usually, testing is done in the local alignment of words and images, but since global semantic consistency is not preserved, this measure is considered inaccurate.

5.Other Methods

A speech-to-text system is needed to convert spoken words to text that will be used as input for our model. The most commonly used speech systems initially derive various features by separating useful audio waves such as MFCC's(Mel Frequency Cepstral Coefficient).They can also be converted into spectrogram to measure frequencies of various ranges or signal strength. The conversion system[4] converts the text exactly as spoken by the user using a bidirectional Kalman filter. This non-stationary filter reduces noise and optimally estimates speech. Stacked Cross Attention Network (SCAN) provides similarity between image - text mapping by inferring latent alignment using both image regions and sentences.

6.Proposed Method

We propose a model that is based on the CASC model for efficient implementation of the speech-to Gregg conversion method. The Cross-Modal Attention with Semantic Consistence for Image-text Matching (CASC)

Doi : https://doi.org/10.54216/JCHCI.030105
Received: January 15, 2022 Accepted: May 26, 2022

is a method that exploits global semantics and cross-model attention to accomplish mapping text to image and image to text. It removes the inability to achieve local alignment as there may be a possibility of inaccurate detection or misinterpretation. This hybrid framework deals with both local alignment and global semantics. It also addresses the fact that by only using fine-grained matching, it is inaccurate to represent global semantic consistency as it may miss interrelated regions that represent the words present. This leads to misrepresentation or defective matching of sentences leading to less accuracy of the overall model. This weakens the overall logical workings of the model. Hence, it proposes a hybrid approach that utilizes cross-modal attention and multi-label prediction. It delivers a solution by working out a similarity, i.e., local alignment from both directions(image-to-text and text-to-image). The multi-label classification also maintains consistency of semantics. All these qualities lead to a greater model of image-text matching that can be used as a basis to implement our model where each word has its equivalent Gregg shorthand.

Based on CASC, our model will be redesigned to translate text into shorthand. The cross-modal attention will greatly assist in identifying the shorthand and the local alignment from both directions will lead to less misrepresentation of words. The redesigned model will be trained with our dataset which will lead to approximate results.

Workflow of Speech to Gregg Shorthand Converter

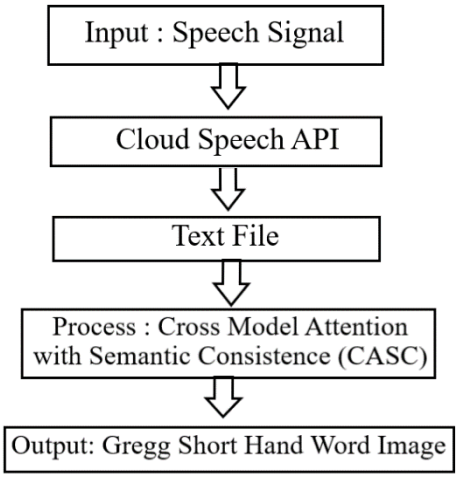

Figure 1:The above denotes the architecture of speech-to-Gregg shorthand converter.

7.Experimental Setup

7.1.Datasets

Currently, there are no publicly available datasets that have massive collections of Gregg transcriptions with their corresponding texts. The existing transcriptions of Gregg shorthand can be used to produce appropriate datasets for our experiment. The dataset provided for this experiment was created by using 25000 images of Gregg shorthand with their matching text. This also includes legal terminologies that will be used in courts. Over 250 words are included in the dataset that was created. Each word consists of 5 images making the total count around 2500.

7.2.Platform and other details of the model

The model will be implemented in a machine learning platform, namely TensorFlow for timely processing. Since our main focus is generating shorthand, we use any pre-existing Speech-to-text APIs for extracting speech features and producing text. In our case, we use Microsoft's cloud service's pre-existing cognitive service for speech-to text service. The API is then later imported into TensorFlow. The designed model is

Doi : https://doi.org/10.54216/JCHCI.030105
Received: January 15, 2022 Accepted: May 26, 2022

executed and tested with the imported dataset. The entire model is then integrated into a web application for providing an interface to the user.

7.3User interface

The web application consists of a record button to record the speech. It serves as the tool that provides input for speech-to-text conversion. The record is then sent to the API for conversion into text. This converted text will be fed into our model for extracting output images. The images will be displayed simultaneously as when it is produced. The shorthand words will be displayed side by side like a sentence on the screen.

8.Implementation

In this section, we depict our approach in three steps: 1.The speech input taken is converted into text 2. The text is then fed into and converted to produce Gregg. 3. The output is displayed to the user.

Initially, the dataset is loaded and pre-trained. The model was initially trained with the Gregg shorthand dataset. The learning rates are adjusted and the dataset is validated. Later, during the execution of our model, the generated text from the speech is fed into the algorithm. The text-image matching is done similarly to SCAN [5] where words and images are mapped into a single space. Unlike CASC, feature extraction is not necessary since we only need to redeem an image(i.e., Gregg shorthand) based on the word (i.e., text). Python torch is used to iterate over the dataset and load objects. For faster processing, Pytorch is used to efficiently generate data that utilizes the full power of the GPU. The tensors are formed from the samples and the collate_fn is configured. Data loaders are configured for both the testing set and the training set to manage the batches split and to iterate over the dataset. The sentences are to be first split into separate words. They are tokenized and converted as a NumPy array.

The model's sub-package includes model definitions for a variety of applications, such as picture classification and pixel-wise semantic segmentation. Autograd is a Python library that allows you to differentiate between all Tensor operations. Starting from a variable, it does backpropagation. A dictionary subclass called OrderedDict remembers the order in which the keys were first placed. The PyTorchtorch.abs () method computes the input tensor's element-wise absolute value.

Our model uses CASC's many functions for identifying and separating words. Since Gregg doesn't provide transcription for numbers and punctuations, they are removed. The corpus vocabulary uses the raw text corpus and converts it into processed text. It holds the metadata of the corpus while also serving as a storage location. All this is done with the help of a natural language toolkit. Dictionaries are created to list and store the mapped text to image pairs. Whenever the model comes across a new word, after retrieving the suitable shorthand, the pair is stored in the dictionary. If the word already exists in the dictionary, then the pair is automatically retrieved and loaded to the output. The dictionary which was serialized is then loaded as a vocabulary wrapper in its native data type. The text, which was stored in a text file, is read and loaded as a Python object for manipulation. The appropriate image for the corresponding word is taken and displayed to the user. The Gregg shorthand for the words in a sentence will be displayed as and when the model retrieves it, and the shorthand images will be displayed adjacent to each so as to view it as a sentence.

Evaluation of the test set is conducted using an evaluation tool that conducts various tasks such as scaling and encoding. The model is evaluated and the validation set is tested. Checkpoints are used for inference and to save the latest epoch that was run.

Doi : https://doi.org/10.54216/JCHCI.030105
Received: January 15, 2022 Accepted: May 26, 2022

9.Execution of Gregg Converter

(a) (b)

Figure 2**:** (a) The speech recorded is stored internally as a text file. (b) The resulting Gregg shorthand is generated.

10.Conclusion

In this article, we present an approach that produces the Gregg shorthand for the spoken narration. We will use the existing image-texting model known as Cross-Modal Attention with Semantic Consistence i.e., CASC to implement the Gregg converter.The recorded audio format is initially processed. The model is redesigned and trained using the newly created datasets.The deep learning model will retrieve the shorthand images of its corresponding text and will display the output to the user.This study can be further improved so as to produce the expected shorthand results simultaneously while being recorded.

Screenshot of Sample UI

Figure 3:The diagram shows user interface of our proposed Gregg converter. The record button is clicked to recorder the speech. Once recorded, the converter button is enabled and when clicked, displays the shorthand in the text box as shown above.

References

[1] DionisA. Padilla, Nicole Kim U. Vitug and Julius Benito S. Marquez., "Deep learning approach in Gregg shorthand word to English word conversion" (2020)
[2] ZhongJi and Kexin Chen, "Step-Wise Hierarchical Alignment Network for Image-Text Matching '' (2021)
[3] Xing Xu, Tan Wang, Yang Yang, Lin Zuo, FuminShen, and Heng Tao Shen, "Cross Model Attention with Semantic Consistence for Image Text Matching'' (2020)
[4] Neha Sharma andShipraSardana, "A Real-Time Speech to Text Conversion system using Bidirectional Kalman Filter Matlab''(2016)
[5] Kuang-Huei Lee, Xi Chen, Gang Hua, Houdong Hu and Xiaodong He, "Stacked Cross Attention for Image-Text Matching" (2018)

Doi : https://doi.org/10.54216/JCHCI.030105
Received: January 15, 2022 Accepted: May 26, 2022

[6] K. R. Abhinand and H. K. AnasuyaDevi,"An Approach for Generating Pattern-Based Shorthand Using Speech-to-Text Conversion and Machine Learning '' (2013)
[7] R.Rajasekaran , K.Ramar, "Handwritten Gregg Shorthand Recognition'' in *International Journal of Computer Applications* (2012)
[8] Zihao Wang , Xihui Liu , Hongsheng Li , Lu Sheng , JunjieYan , Xiaogang Wang and Jing Shao, "CAMP: Cross-Modal Adaptive Message Passing for Text-Image Retrieval'' in *IEEE/CVF International Conference on Computer Vision (ICCV*) (2019)
[9] StanislavFrolov , Tobias Hinz , Federico Raue , J¨ornHees and Andreas Dengel, "Adversarial Text-to-Image Synthesis: A Review" (Neural Networks Journal,2021)
[10] SaifuddinHitawala, "Comparative Study on Generative Adversarial Networks''(2018)
[11] Cheng Wang, Haojin Yang, Christian Bartz and ChristophMeinel, "Image Captioning with Deep Bidirectional LSTMs'' (2016)
[12] Daniela Onita , Adriana Birlutiu and Liviu P. Dinu, "Towards Mapping Images to Text Using Deep-Learning Architectures'' (2020)
[13] Christine Dewi , Rung-Ching Chen , Yan-Ting Liu and Hui Yu , " Various Generative Adversarial Networks Model for Synthetic Prohibitory Sign Image Generation" , (2021)
[14] Hao Wu , Jiayuan Mao , Yufeng Zhang, Yuning Jiang, Lei Li, Weiwei Sun, and Wei-Ying Ma., "Unified Visual-Semantic Embeddings: Bridging Vision and Language with Structured Meaning Representations" , (2019)
[15] Scott Reed, ZeynepAkata, Xinchen Yan, LajanugenLogeswaran , BerntSchiele and Honglak Lee, "Generative Adversarial Text to Image Synthesis" , (2016)

Doi : https://doi.org/10.54216/JCHCI.030105
Received: January 15, 2022 Accepted: May 26, 2022