International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878 (Online), Volume-8 Issue-4, November 2019

Evolving Artificial Perception by Integrating Nlp and Cv

M. Prabu, Mohammed Saif, P Naveen Barathi, Kailash Venthan

Abstract: ***This paper discusses the concept of integrating artificial perception of an artificial intelligence by integrating NLP and CV, this should be able to solve 50% of problems where the data is usually in a raw format and not understandable by the machine. This method helps in the automatic labelling and understanding the data so it is easier for the machine to understand and help in our day to day tasks. "Perception is the ability to become aware of something which is internal or in the external environment through the use of the 5 senses" this is a natural capability of humans but has never properly been achieved in a machine. In the past five years massive strides have taken place in both natural language processing and computer vision but none of these advancements have increased the intelligence and perception of computer systems in the dramatic way that was expected. This difference in what was expected and what has finally been delivered is due to the fact that both these fields have evolved separately whereas perception requires these two dimensions of hearing (Natural Language Processing) and vision (Computer Vision) to be integrated.***

Keywords: Computer Vision. Data Labelling, Image Captioning, Natural Language Processing, YOLO (You Only Look Once) Algorithm.

I. INTRODUCTION

Intuitively, readers can grab the gist of the particular event more easily simply by looking a photo or the particular video than by simply reading news document, plus thus we believe of which the multi-modal data can also reduce the problems for machine to recognize a new servant. While just about all summarization systems focus in only natural language Processing (NLP), the opportunity to be able to jointly optimize the top quality of the summary using the aid of Automatic speech recognition (ASR) and computer vision (CV) handling systems is widely dismissed.Natural Language Processing plays an essential role inside our daily lifestyle and has been researched for several decades. By information retrieval to data mining, we are usually exposed to text summarization.

Manuscript published on November 30, 2019.
* Correspondence Author
M. Prabu*, assistant professor in SRM Institute of Science and Technology, Ramapuram
Mohammed Saif, B. Tech degree in Computer Science from SRM Institute of Science and Technology, Ramapuram.
P Naveen Barathi, B. Tech degree in Computer Science from SRM Institute of Science and Technology, Ramapuram
Kailash Venthan, B. Tech degree in Computer Science, SRM Institute of Science and Technology, Ramapuram.

With the coming in the information age and typically the emergence of multimedia technological innovation, multimedia data (including text message, image, audio and video) have increased dramatically. Media data have greatly altered how people live and even make hard for consumers to obtain important
info successfully.The vision stream learns deep representations from the original visual information via deep convolutional neural network. The Language part of the processing system is used to encode the different visual aspect of the methods in which the image is provided. Since the two fields are complementary, combining the two streams can achieve better classification accuracyLet us assume that we would like to train data and we have images which are not labeled, it become hard or nearly impossible and would definitely generate wrong results

The goal of this work would be to unite the NLP and Computer Vision to make a new platform for mining the useful information within multi-modal data to boost the grade of multimedia media summarization for the machines.

II. RELATED WORK

Our work is inspired by the methods of XMC and YOLO for labelling and classification of data respectively.

A. ABC-CNN Technique

This technique contains four major components, which focuses on the information region of an image where it is based on question guided attention. Section 1 of the component will extract the features of the image. Section 2 understands the question. Section 3 gets the data from the image. Section 4 generates the text based on the image or results. An answer will be generated in a single sentence or sometime even a word.

B. End-to-End Module Network Technique

This model is employed to handle integrative reasoning form of visual question respondent. Section A is where the model is divided to get different neural modules. Section B is where the layout policy is enforced with RNN to predict layout expression for every question. Section C is used to predict the outcome, for queries and applies to the input image to urge the solution.

C. Attention Mechanism

This method solves only question-based pictures. Attention model can specialize in question specific region of a picture instead of all the other options in a picture.
There are some ways to tackle these issues.

One approach to use attention to the present illustration is by suppressing or up the options at numerous spacial areas. Utilizing the question like options with these native image options like solely umbrella region, a coefficient issue for every lattice space may be patterned that decides the spacial areas to the questions, which might then be ready to be used to method image weighted options.

D. Ask Me Anything (AMA) Technique

This technique mixes the generated description of a picture with external existing object to produce a solution of a general question answer pairs. There is a major element during this model. First, the data from an object is extracted, read and merged with other data. Secondly, this element is used to predict a set of attributes of the image through CNN. The third part, suggests and gets the element a VQA model with multiple inputs. During this element encoded attributes, captions and the info area unit is taken as single input to get the appropriate data.

E. Ask Your Neurons Technique

These techniques have neurons to answer the question which are post using several RNN and CNN to make the required changes to convolutional neural network. It individually searches for the content.

III. ALGORITHM USED

The method used to create our project is by combining NLP and CV i.e. You Only Look Once and Image Captioning respectively.

A. You Only Look Once (YOLO)

The previous predictors of image detection like R-CNN, are a bit slow when compared to this method.The YOLO methodology applies many neural networks to the complete image. This network divides the image into regions and predicts by bounding boxes and scoring for every region. These bounding boxes are weighed by the different scoring techniques.The higher the score the better accuracy for the production of results. The model uses a file called as weights which store the information of the trained data. And a configuration file which store the list of objects that are stored.During training we optimize the following data by using the multi-part loss function with the formula:

$$\lambda_{coord}\sum_{i=0}^{S^2}\sum_{j=0}^{B}\left[(x_i-\hat{x}i)^2+\quad(y_i-y_i)^2\right]+$$
$$\lambda_{coord}\sum_{i=0}^{S^2}\sum_{j=0}^{B}\left[(\sqrt{w_i}-\sqrt{w_i})^2+\quad(\sqrt{h_i}-\sqrt{h_i})^2\right] \quad (1)$$
$$+\sum_{i=0}^{S^2}\sum_{j=0}^{B}\left[C_i-C_i\right]^2-\lambda_{coord}\sum_{i=0}^{S^2}\sum_{j=0}^{B}\left[C_i-C_i\right]^2$$

where 1 obj i denotes if object appears in cell i and 1 obj ij denotes that the jth bounding box predictor in cell i is "responsible" for that prediction.Equation (1) known as Multi-Part Loss Function the loss function is used to find the number of mistakes and inaccuracy of the results. This also gives us the data about where the mistakes are made and where to change the values to be corrected.YOLO has the limitation that the boxes that are implied and recognized cannot have two same objects, for example, this method cannot recognize that if there are two cats it can assign the object only to one cat which has the higher possibility of the attribute of cat. Since the model learns to predict bounding boxes from data, it become hard to generalize object with never seen/ different ratio. The YOLO application is used to find constantly the images in the screen. This method is faster than other R-CNN and other neural networks.

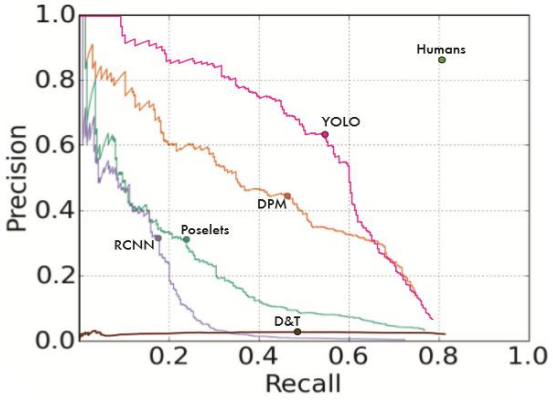

Fig.1: Comparison of different methods w.r.t. Speed and precision

Here we take an example of Picasso datasets and compare the speed with other neural nets.

B. Image Captioning

Image captioning is a way of reading the image which is given and describing the events which are happening in the image.For eg, when there is a flock of birds sitting on a branch of a tree, the image captioning technique is used to give a sentence such as "A flock of birds sitting on a tree branch".Up to this point we've always generated training data ahead of time and fit the neural network to it. The memory demands can be considerable. If the training data can be generated, as the neural network needs it, it is possible to use a Keras generator. The generator will create new data, as it is needed. The generator provided here creates the training data for the caption neural network, as it is needed.The neural network accepts two objects (which are mapped to the input neurons). The first is the photo. The second is an ever growing caption. The caption begins with just the starting token. The neural network's output is the prediction of the next word in the caption. This continues until an end token is predicted or we reach the maximum length of a caption.In many applications of daily use, natural language processing machine learning is widely adopted– voice support, chat bots, question answer system, automated email responses, etc. It is up for a discussion of that your application case if you are interested or are looking for a operationto apply natural language treatment to a company or a product.

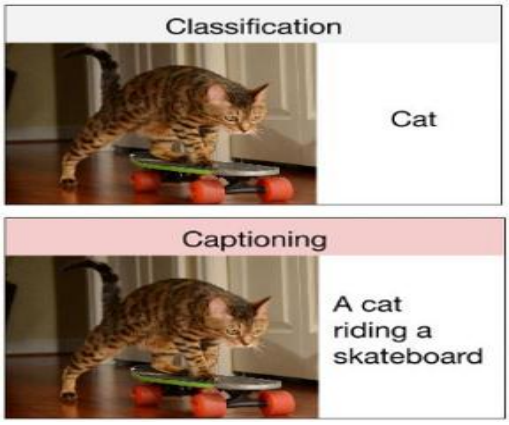

Fig. 2: Example of Image Captioning

The word which has the highest probability (from the neural network) is chosen which suits the best for the model. Each time a new word is predicted for the caption. It is not always necessary to generate all of the training data ahead of time.

IV. PROPOSED SYSTEM

A. Voice/Language stream

The data collected shows us the readability with a cummilated score of 5, these are given below shows us in different languages Chinese and English about the readability of our system

We therefore compare manually that the readability and understandability of Fig. 2 by the statistical data given above with the scale between 10.Currently there are only systems such as google assistant, alexa and siri which rely heavily on nlp technologies to deliver a voice assistant as a service this is not comprehensive in nature as the voice assistants have no sense perception or understanding of the environment partially due to the fact that computer vision or vision in general has not become a part of what these systems are designed to do.The proposed system could be implemented in any current computing device such a phone or pc with a camera and mic to give it capabilities of perception and action creating fully capable assistants for humanity to benefit from.There is currently the need to develop systems which can perceive their environment as it changes and act within these environments without any delay or training time, this need arises out of the fact that the only way to progress machine learning technology is to make it a general case technology the problem then is to integrate technologies which understand sound (NLP) and vison (CV) to form a general case perception system.The above-mentioned methods are implemented with NLP to give us applications which can be applied to various devices.The sole purpose of this paper is to make ease of access for devices, for differently abled people with special needs.

B. Image Stream

Table I : Summary of Quality Evaluation

	Method	Readability	Informativeness
English	Text only	3.72	3.28
	Text + audio	3.08	3.44
	Text + audio + guide	3.68	3.64
	Image match frame	3.67	3.83
	Image topic IR	3.8	4.1
	Reference	4.52	4.36
Chinese	Text only	3.64	3.4
	Text + audio	3.16	3.48
	Text + audio + guide	3.6	3.72
	Image match frame	3.62	3.92
	Image topic IR	3.73	4
	Reference	4.88	4.84

Text-image matching is the most challenging module of our framework. Although we use a state-of-the-art approach to match text and images, the performance is far from satisfactory. To determine a somewhat strong upper bound for the task.

We define two a-weighted objective functions to give different weights to textual and visual information. We define an a-weighted matching-based objective

As we have discussed in fig. 2 about the quality evaluation which has been retrieved by using the following formulae:

$$F(S) = \frac{\alpha}{M_A} F_S(S) + \frac{1-\alpha}{M_e} F_m(S)$$

$$\text{where } F_S \text{ and } F_m \text{ are image cluster values}$$

$$\text{where } \alpha \text{ is the weight for image pi} \qquad (2)$$

$$\text{where } M_e \text{ and } M_A \text{ is the speech transcription by the salient score}$$

The Equation (2) is called Weighted Classification Distribution which gives us a perspective of the speech transition and this Equation (2) is used to transcribe the image and help us save the data in a manner that its JSON file can be used and imported in a csv.

This can further be used in Python or other languages to perform analysis, after each epoch run the neural net gets trained better and serves the purpose of applying it to the Machine Learning curve.

C. Architecture

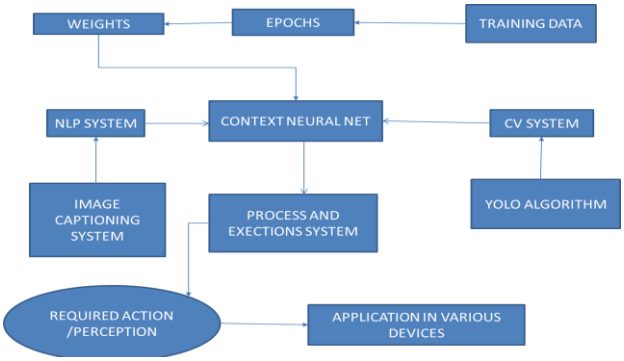

Fig.3: Architecture Flowchart

This flowchart depicts the implementation and the structure of all the algorithm used to create an structure to apply it in various devices.

IV. RESULT ANALYSIS

A. Accessibility features

Fig.4: Grid Pattern

The implementation part is just as shown above, just like the algorithm is used made to study the maps of Japan, this implementation part is used with an example which is integrated with a PC.Now, in this model we can just ask the PC to go to the particular grid. For ex. “Go to 2A and select paint”, should automatically do the command and give us the output. i.e. Open Paint application.The same can be used, but using number labeling to select options and fill out forms easily. Only a voice command is used to do our job along with vision.This method is mainly useful for people with special needs and are not able to use their hands. This drastically increases the accessibility options for many users which encourage them to use these systems.

B. Auto-Data Labeling

There is a data scarcity in the field of Machine Learning and OpenCV, even when a Web Crawler is used to retrieve information it only fetches us RAW data/ Unlabeled data.

Our project implements that whenever we take photos, the data should be labeled automatically, hence, reducing the labor of labelling each individual data.

The applications are endless, we can make applications like Google Map Navigation, but for people who are blind by helping them recognize the objects in front of them and aid them.By using, the AR-Kit and AR-Core in modern devices to augment the data and recognize the object along with its features.

VI. CONCLUSION

The proposed system could be implemented in any current computing device such a phone or PC with a camera and a microphone to give it capabilities of perception and action creating fully capable assistants for humanity to benefit from.The system could aid both disabled and the elderly in tasks which involve sense perception such as seeing things around them and hearing things around them and interacting with these external items effectively.Mundane daily tasks which involve only sensing things such as sorting items and labeling items would no longer require training a specific neural network and could be accomplished on the go. This is a quantum leap in machine learning and would hugely reduce the time wasted by humans in doing these tasks daily.

REFERENCES

1. E. Cambria and B. White, "Jumping NLP Curves: A Review of Natural Language Processing Research [Review Article]," in IEEE Computational Intelligence Magazine, vol. 9, no. 2, pp. 48-57, May 2014.
2. R. S. Dudhabaware and M. S. Madankar, "Review on natural language processing tasks for text documents," Computational Intelligence and Computing Research (ICCIC), 2014 IEEE International Conference, pp. 1-5, 2014.
3. Q. Wu, D. Teney, P. Wang, C. Shen, A. Dick and A. van den Hengel, "Visual question answering: A survey of methods and datasets", Computer Vision and Image Understanding, vol. 163, pp. 21-40, 2017.
4. K. Kafle and C. Kanan, "Visual question answering: Datasets, algorithms, and future challenges", Computer Vision and Image Understanding, vol. 163, pp. 3-20, 2017.
5. S Antol, A Agrawal, J Lu and M Mitchell, “Vqa: Visual question answering” In IEEE International Conference on Computer Vision (ICCV), pp. 2425-2433, 2015.
6. H. Li, J. Zhu, C. Ma, J. Zhang, and C. Zong, “Multi-modal summarization for asynchronous collection of text, image, audio and video,” in Proc. Conf. Empirical Methods Natural Language Process., 2017, pp. 1092–1102.
7. B. Erol, D.-S. Lee, and J. Hull, “Multimodal summarization of meeting recordings,” in Proc. IEEE Int. Conf. Multimedia Expo, 2003, pp. III-25–III-28.
8. T. Hasan, H. Boril, A. Sangwan, and J. H. Hansen, “Multi-modal highlight generation for sports videos using an informationtheoretic excitability measure,” EURASIP J. Advances Signal Process., vol.2013,no.1,2013,Art.no.173.
9. “Data scarcity, robustness and extreme multi-label classification” Rohit Babbar1 ·Bernhard Schölkopf2
10. “Read, Watch, Listen, and Summarize: Multi-Modal Summarization for Asynchronous Text, Image, Audio and Video” BY Haoran Li , Junnan Zhu, Cong Ma, Jiajun Zhang , and Chengqing Zong
11. “VQAR: Review on Information Retrieval Techniques based on Computer Vision and Natural Language Processing” BY Prof. DhatriPandya
12. “Natural Language Requirements Processing” By Alessio Ferrari, Felice Dell’Orletta, Andrea Esuli, Vincenzo Gervasi, University of Pisa, Stefania Gnesi.

AUTHORS PROFILE

M. PRABU received the M.E degree in Computer Science from the department of computer science. He is currently working as an assistant professor in SRM Institute of Science and Technology, Ramapuram and is pursuing his Ph.D degree in the field of Image Processing.

MOHAMMED SAIF is currently pursuing his B. Tech degree in Computer Science from SRM Institute of Science and Technology, Ramapuram. He is currently working on projects related to Machine Learning, Data science and Artificial Intelligence field.

P NAVEEN BARATHI is currently pursuing his B. Tech degree in Computer Science from SRM Institute of Science and Technology, Ramapuram. He is currently working on projects related to Machine Learning and IOT technologies.

KAILASH VENTHAN is currently pursuing his B. Tech degree in Computer Science, SRM Institute of Science and Technology, Ramapuram. He is currently working on projects related to Machine Learning and is aspiring to open a startup.