Michigan Technological University

PSY 6990

Business Case Report;
Visual Search Application for Visually Impaired

By

Nisarg Dave, Erin Richie, Sivaramakrishnan Sriram

PSY 6990: Explanations in AI Systems

Project 1

Overview:

Artificial Intelligence is in existence for decades from now. The Main intuition of an Artificial Intelligence system will be to make the machines learn and think simultaneously in order to assist a task.

The Visual Search Application is an Artificial Intelligence system that helps the visually challenged people to "hear what they wanted to see". The research on making an image input to text and voice reception using Natural Language Processing began in 2012. But the working model Implementation was first implemented in "in the wild" youtube video where the video had text descriptions of what's the primary activity happening in the video.

Our System:

Our Visual Search Intelligence system will capture the environment in front of us as an image. This image is sent as an input to the Convolutional Layer of our deep network which gets processed in the mean pooling layer. Then the processed information goes through the LSTM Recurrent Neural Network Layer which gives the output of the features that got captured in the Neural Network as individual words in a sequence of one by one .

Technical Description:

Our System uses Deep learning to represent the environment in front of persons with disability. We are using deep learning networks for analyzing the scene and describing in natural language. We are explaining all technical details of our system in terms of data science project lifecycle.

Data munging:

The primary data is Video, real time video from head mounted device will get as an input to our deep learning network. The training data can be the labelled video files with appropriate description of each frame.

Feature engineering:

The most important step is to perform the transformation of features.We will transform each and every scene of video to static images and then we will feed those into convolution layer of CNN. The CNN will use transformed images and text input associated with each scene. The transformation will help because we can interconnect or relate between two sets of frames for getting better joining of events.

Predictive modeling:

We are creating stack of two different neural network architectures. First, we will use Convolution neural network. Each current scene will get transformed and will be send as a input image to convolution layers. Convolution layers will extract the important features of the image. The higher level layers will gather generic image information such as shape, boundaries, object type etc. The deep level layers will gather more important and descriptive information. We will use mean pooling so we can converge to the real representation with accurate measures. The mean pooling will combine all extracted feature knowledge and try to make some sense by interconnecting them. The knowledge representation is mostly in terms of vectors. All those vectors will get combined and mapped into separate space for finding correlation and similarity. After generating N such vectors, the algorithm will create the final vector which can fully represent the video in abstract. CNN can't represent the described image in form of natural language so this final knowledge will be the input to another LSTM Recurrent Neural network. The LSTM will get each internal representation and process in terms of natural language generation. Each and every output will get aligned in terms of occurrence so order of occurrence also matters when we are dealing with natural language processing. LSTM network will take each occurrence and its order into consideration. It can even store certain information and later it can compare and see whether it fits into current order or not? It's reinforcement learner with slight modification in the present state. The future states have significantly high importance thus it can even change the behavior of algorithm.

Overview of Architecture:

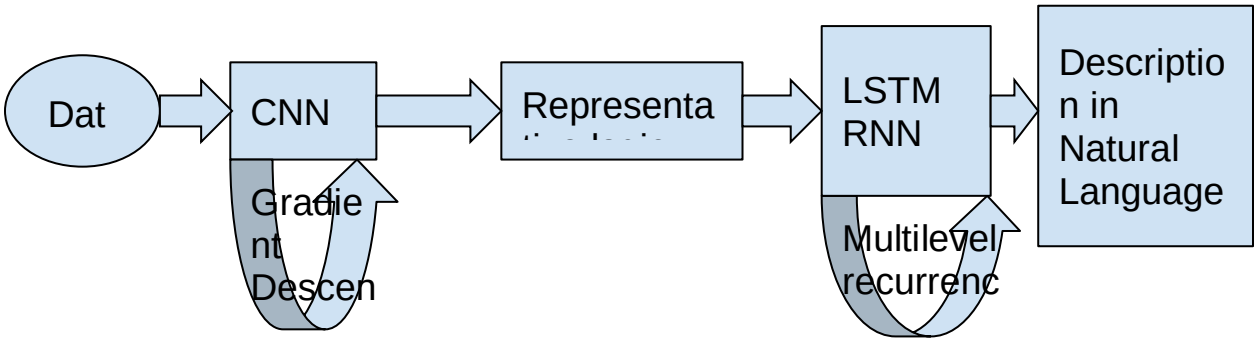

Fig 1.0: System Architecture

This is Overall abstract architecture of whole system as a one, the CNN and LSTM RNN is explained in deep below.

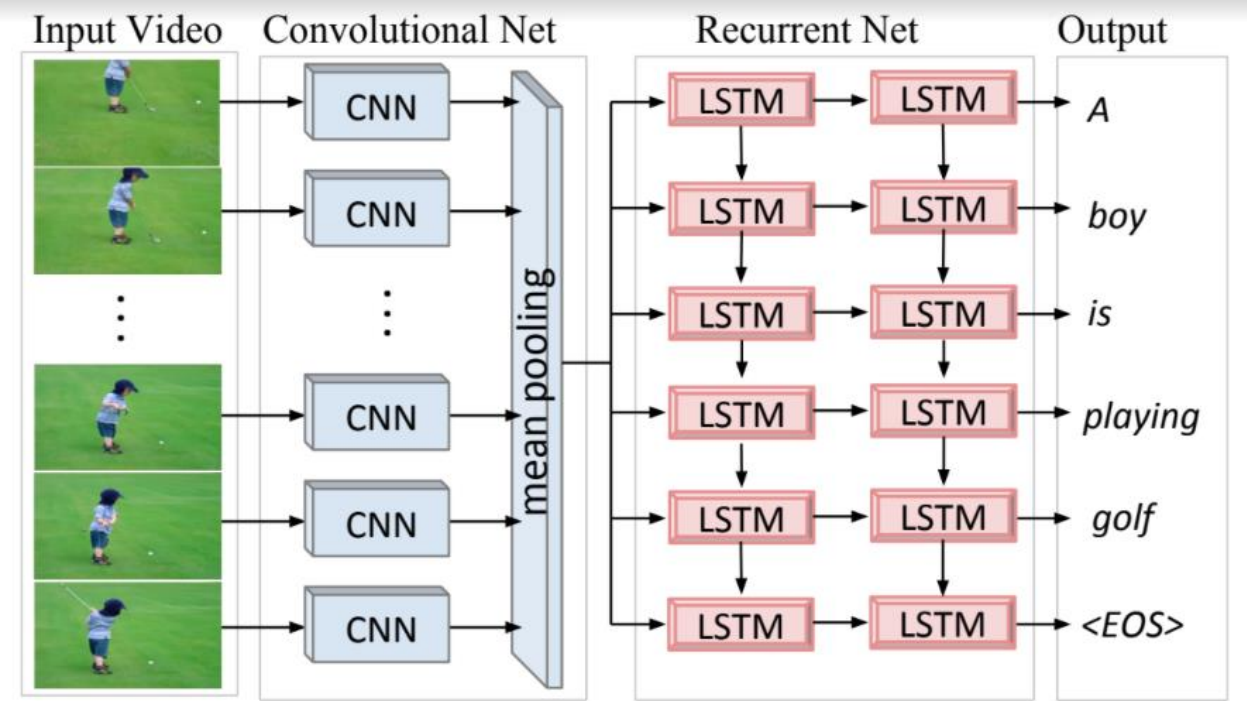

Fig 2.0 Neural network Architecture[1]

Optimization:

We are reducing dimensions for making it less computationally intensive. The Optimization is necessary step for both time and space complexity. We are considering stochastic gradient descent for better higher dimensional optimization in multimedia applications. The justifiable learning rate and learning through experience in real world environment makes much complex optimization task. Gradient descent finds the global minima for statistically computed error rate. It even has flexibility of slow/fast convergence. We can always adjust the learning rate according to need. Stochastic processing will help in using random samples and optimizing it independently with respect to others.

Explanations:

Explanation to CEO:

Goal:

To create a product that will assist the blind or sight impaired in busy environments through use of visual search technology.

Context:

The blind and sight impaired face more challenges than the sighted on a day to day basis. Walking to work or the store in a busy city like area can be challenging. We have the opportunity to create a device that may assist those who may not have access to a companion, seeing eye dog, or are unable to use current market sight enhancement technologies.
The current market is small and niche- but no other 'competitor' is using the same type of technology that we are proposing. Other products, such as e-sight [3], attempt to enhance vision for the visually impaired- but there is yet to be a product outside of the alternative of guide dogs and companions that seeks to explain the visual world using audio.
We feel that this product could prevent accidents, save lives, and allow our users more independence. Though be aware that we would need to consult a legal team to ensure that we are not accountable if the product is used incorrectly or in unrecommended environments.

Global System:

The system will use visual search. Visual search is an AI tool that recognizes features within a given image, interprets, those features, and works to explain certain aspects of the image to the user. We will be using this technology to search images and look for hazards within the user's path.

Local System:

The system will consist of a visual search eyepiece that can be attached to a pair of sunglasses, hat, or headband. The visual search piece will take frequent pictures of the environment and understand what obstacles may lay in the path of the consumer. When an obstacle is detected, the visual search piece will communicate with an earpiece that is worn in the ear like a hearing aid. Phrases such as "car ahead" or " open manhole, step right" will be spoken. The product could help users to cross the street and navigate busy city areas, potentially saving users from accidents and saving lives.

Explanation to Customer:

Goal:

To deliver a product that will help you to feel safe and independent, especially in busy city environments.

Context:

Our product seeks to assist you and act as your guide. We want to provide you with an option that can give you more information than a cane, but doesn't require the same amount of care as a guide dog. Imagine if you had a voice in your ear to gently alert you to hazards in your surroundings.
While we believe our product can assist you in your daily tasks, please be aware that it is meant to be used in a city environment, so it may not be as effective in rural areas. As well, our product is not made for those with hearing impairments. Those who have hearing impairments may find it difficult to carry out day to day activities with a speaker piece in their ear, as it can make it more difficult to hear the world around you.

Global System:

The product consists of a visual search eyepiece that can be attached to a pair of sunglasses, hat, or headband and a comfortable earpiece that rests inside of the ear much like a hearing aid.

Local System:

Our search and report tool rapidly takes pictures of the world around you and scans them for hazards, reporting warnings such as "car ahead, don't cross", so that you may navigate busy streets safely. Our search and report technology doesn't just detect for something in your path, but knows the difference between a car or a human, for example, and can let you know what to expect in front of you. In this way you know whether the obstacle is dangerous or not.

Explanation to Regulator:

Goal:

The Goal of this Explanation will be in order to get the AI system approved by any of the regulatory departments. In other case, if any Medical community or association is going to buy this system , the goal of explanation will be in order to satisfy their regulations and their standards. In the latter case, the recipient's goal for the explanation will be in order to calibrate the trust in the system before they actually buy it.

Context:

The intended use of the visual search application is to assist people with disabilities. People will be given a wearable device like a glass. This glass will be used to capture image of the environment in front of them. It will then get processed through the AI system and will generate the activity portrayed in the image using Natural Language Processing. The output will be heard as an audio output.

Further Explanation:

In order to explain this system to you, the intended use will not solve the purpose. You may want to know about the ethics involved, the risks associated with the system, the class associated with the system, the QMS report and as our target audience is on healthcare industry, we need to explain the Clinical evaluation Report.

As we said in the context, the intended use of this application is for the people with visual disabilities. This system will just tell what is the primary activity there in the image. The output will be a audio output. This system is not against any ethical regulations as it just converts an image information to an audio output.

As a regulator you will be willing to know more about the data with which the system is getting trained. This is very important from a regulator perspective the behavior of an Artificial Intelligence system is completely dependent on the data with which we are training. The data could be an easy factor with which the AI could go misleading.

The data with which we would like to train the AI system will be images we take on the environment surrounding us. These images will include basic objects like a car , a park ,a computer , a cell phone etc. The description for these images will also be the object name or a verbal reasoning term . Then the system will be trained on it's own to frame sentences based on the features that it gets extracted from the image.

The next part that is to be explained is the class of the system. As our system is related to healthcare we need to confront which class of risk does our system belongs.

Class 1 = the system which gives less risk. Something like thermometer or disability aid.
Class 2= the system with medium risk. Like things which makes direct contact to human body.
Class 3 = High Risk which involves life situations.

Our system belongs to class 2, as it makes a human body contact through the glass and the headphones. The risk associated with our system is medium as you can see this system behaves as an disability aid and will have minimum human body contact. As we mentioned earlier the chance of this system getting misleading is also very less.

The Quality Management report is one of the important factor in determining the approval. The quality of this system depends on the data with which we train the system and the how the weights and bias are applied in the Neural Network Layers. This will determine the quality of the

prediction that predicts the description of the image. So, the quality can be measured with the number of accurate predictions the system makes.

The clinical report will be more of reporting the effectiveness of the system . There is a fact which says no medical device is 100% accurate. But as this being cited as a Artificial intelligence system , and this being not more of technically a medical device. So, the clinical report can measure only the accuracy and ethics of the system.

The system is responsible to process data in real time so the data storage isn't necessary. Eavesdropping isn't possible but if we want to deploy the reinforcement learner for more intelligence then collecting data will make sense. In that case we can use the differential privacy for deploying privacy preserved machine learning. The regulator should verify and explain the possible privacy mechanism being used if data collection routines are deployed. Specifically, system isn't using the cloud infrastructure or server side computations so security issues are mitigated.

Contrastive Reasoning cases:

Though our system will have capacity to transform the image and describe the featured activity , it cannot be able to monitor all the time and cannot be a full-time assistant as the images can be noisy and the noisy data description can be easily misleading. So, if the user wants to know about a specific description only , he can send the image to the system and get the description. So this system is just a simple time assistant and not a full time assistant.

Failure Methods:

The only failure will be if the predictions go wrong. But though our system is going give the description of the activity and if the image is clear enough, the chances of getting a wrong prediction is very less.

Explanation to Engineer:

Goal:

To create the product which explains the environment in real world scenario.Creating dynamic and complex networks will help for making smarter and speedy decision accurately.
The goal here isn't limited to the explanation of core technical details but it is extended in technical details, troubleshooting and testing details and even in upgrade cycles. Making the network architectures self explainable at each and every level is expected as a final deliverable. The system should explain itself in a concise manner to each developer.Making the networks,

which reveals the internal working by themselves.In computer industry, source code is not easily interpretable so mentioning the comments and writing markdowns for future developers is necessary. Overall goal is to make whole system interpretable and easily understandable in each and every developing phases including background study, development, testing, troubleshooting and update iterations.

Context:

Intelligent visual search is taking real time video as a input, processes it within neural network stack and generates the live description in natural language. The core system is the stack of two neural networks. The stack of CNN and LSTM-RNN is core of whole system. The video transformed and processed by CNN will have final core representation logic. Logic will be input for LSTM-RNN, RNN will act as a logic decoder and it will decode the meaning out of that final knowledge representation. RNN will describe the video scenario in natural language. The system will have several phases such as data cleaning, feature engineering, predictive modelling & training-testing. All the phases will be executed and managed by engineering team. Engineers are key to service routines, too. The update schedules and maintenance cycles are managed by them.

Further Explanation:

1. **CNN Working**
2. **LSTM-RNN Working**
3. **Data handling and munging**
4. **testing**
5. **Troubleshooting**
6. **Core functionality**
7. **System Services**

Primary thing is the core system working. Complex systems like neural networks have lot of hidden layers with plenty of nodes. All the nodes have their own weights and activation function. All nodes are responsible for different tasks. Each node contributes its own logical inference in final representation. Generally it's black box kind of system and so you can't explain any level in such precise detail so it's very necessary to find method through which each level can explain its working.Testing and maintenance is necessary for such a complex system. Engineer need to explain the troubleshooting, testing and maintenance guide for all developers. So explaining the core functionality and system services is very very important.

Global System:

The global system works with external headset and earpiece. The headset camera takes the environment scenario and processes it. The processing contains understanding the scene and creating the audio description for the same.

Local System:

First local system is the feature transformation system which reduces dimensions and transforms the current data into new plane. The second prominent local system is convolutions which identifies knowledge and extracts it into important feature vectors which ultimately converges to knowledge. The third most important thing is recurrent network which generates the audio description out of knowledge generated from CNN. The last important local system is to optimize the mathematical model for the better performance. Explaining the optimization mechanism is necessary too. This is important for the troubleshooting and testing needs.

Failure Methods: The Major failure here is over generalizing the model. The system should not be hard bounded by the parameters. Using the optimization might mislead the overall system response time. This is real time system so real time processing is necessary. Any failure or delay may lead to incorrect actions or consequences. The system should be able to avoid the significant delays.

Conclusion:

Considering the fact that 'Singularity is near' we are considering reasonable systems which can explain themselves. The convergence of natural intelligence and artificial intelligence probes the serious issues of ambiguity and uncertainty. Today's black box logic systems can be dangerous and complex in long term, if they are failed in explaining themselves to others. We tried to focus on possible explanations of applied AI System for modern AI age.

Our Visual Search AI system concept was explained in terms of four different stakeholders. The concept is targeted for visually impaired people. For each new stakeholder, different levels of explanation are expected. We balanced necessary information about how the product would operate, with pros and cons for each differing level of understanding.

References

1. Translating Videos to Natural Language Using Deep Recurrent Neural Networks by Subhashini Venugopalan, Huijuan Xu, Jeff Donahue, Marcus Rohrbach, Raymond Mooney, and Kate Saenko

2. https://towardsdatascience.com/how-to-get-clinical-ai-tech-approved-by-regulators-fa16dfa1983b
3. “How ESight Works -.” *How ESight Works* |, lowvisionmd.org/esight/how-esight-works/.