

## Domain Background

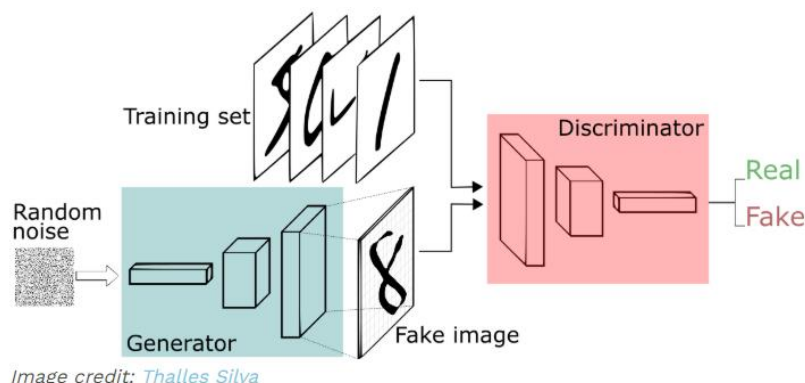
Content Marketing is the process of creating an organic growth channel for your Business content. Content can range from Cat Videos, blog talks or funny pictures, all created to attract new potential customers. Content Creation can be very time taxing so creating content efficiently would be very beneficial to a Business.

Chatbots' central purpose is to facilitate interactions between products or services and customers. They connect with customers, can provide content and even facilitate a purchase. Chatbots are very scalable and personable. Since Messaging apps have surpassed social networking apps in usage, Companies need to engage customers through this medium in a cost and sales effective way. Chatbots satisfy both of those requirements!

My personal motivation to work on this Chatbot problem stems from the opportunity I see in creating an engaging Chatbot for businesses. It would be very beneficial to my current company and I want to contribute something that helps us succeed.

## Problem Statement

Generative Adversarial Networks are deep neural net architectures comprised of two nets pitting one against the other (1). A Generator net which attempts to generate correct output is created by feeding in context (or random noise). A Discriminator is a neural network created to detect real input from fake input. When it comes to text generation though it is extremely hard to determine what an acceptable utterance vs fake one. Training a GAN is done by the two playing cat and mouse, if the discriminator catches the generator the generator updates itself to produce better outputs, if the generator is not caught the discriminator updates itself to better catch "fake" input.



The discriminator and thus the whole process of training a text generator, does better if it is pre-trained on a dataset. Currently, there is no publicly available pre-trained discriminator available.

## **Datasets and Inputs**

The Ubuntu Dialogue Corpus will be used for the dataset in the creation of my proposed chatbot. This dataset was developed and released to provide a large dataset for dialog systems to facilitate breakthroughs in AI dialog systems. It will be used to train the chatbot agent to select the best answer to questions posed by users.

The dataset is appropriate for the given problem because it includes good and bad responses given the context allowing a pre-trained model to be created. It includes "noise" responses so the model has to select the correct response out of the noise (2).

## **Solution Statement**

There are two main ways to build chatbots currently, generative and retrieval based. Generative has been proven to be very useful in providing answers to questions and since it "generates" responses character by character, word by word, or token by token it can provide answers never seen before. This however does not work in its favor here since the responses need to be measurable on how close they are to the correct answer. My solution to this problem is to create a pre-trained Discriminator for textual response generation using reinforcement learning. This will act as the other half in the GAN structure to discriminate against the generator. My solution will be based on "The Ubuntu Dialogue Corpus: A Large Dataset for Research in Unstructured Multi-Turn Dialogue Systems" paper (2).

## **Benchmark Model**

Term Frequency Inverse Document Frequency has been used widely to determine how important a word is in a document and information retrieval. I propose using this model as a benchmark to determine if my own is "successful". To test the successfulness of this model I will be using Recall@K and precision@K. Recall@K. I will measure the success of the created model against the success of this naïve model.

## **Evaluation Metrics**

The evaluation metrics to be used will be Recall@K and Precision@k. Measures the models ability to find the relevant documents while Precision@K measures the models

ability to reject non relevant documents. If the model can both choose the correct text given the context, and reject generated text given the context, then it is a great discriminator for our generator. Both measures will be used in determining the models usefulness since if only one is used, potentially a bias model can become the optimal choice. Using the F score I think is the optimal choice and combination of the two. The mathematical Representation is as follows:

$$\mathcal{F} = \frac{1}{\frac{1}{2} \left( \frac{1}{\mathcal{P}} + \frac{1}{\mathcal{R}} \right)} = \frac{2 \times \mathcal{P} \times \mathcal{R}}{\mathcal{P} + \mathcal{R}}$$

## Project Design

### Data Preprocessing

I intend to use the built in preprocessing features of the Ubuntu dataset. One way to alter the input is to apply a lemmatization function and a stemming function. "The goal of both stemming and lemmatization is to reduce inflectional forms and sometimes derivationally related forms of a word to a common base form" (4). This means that words such as organize, organizes and organizing is treated as one word as well as democracy, democratic and democratization will be treated as one word as well. A tokenizer will also be applied to the data to separate the words so the input can have lemmatization and stemming applied to it. The data is organized in (context, response, correct flag) so I will use the previous techniques on the context and response.

### Data Investigation

Data investigation will be important to this model creation for validation reasons. If the model is trained with thinking that unwanted utterances like "I don't know" or "goodluck with that" to questions are acceptable than this discriminator will fail to reject these unwanted responses. Therefore investigation into how many of these responses are in the training set and removal of them if necessary to improve the "accuracy" of the discriminator. If training data is removed it will also be removed from my naïve model to make sure it will be a true benchmark model.

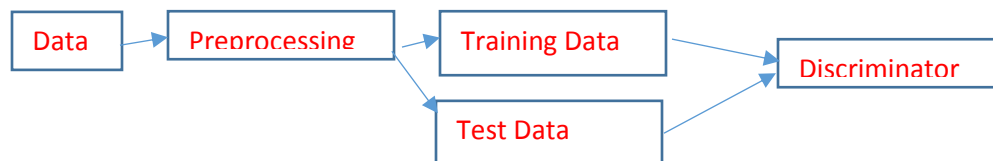
### Model

Since I am going to target industry chatbot training discriminator for my end product, the discriminator needs to be able to account for the context of the whole conversation to be able to determine the correct response. A general question and answer bot does

not need to keep track of previous interactions. A discriminator for a customer focused chatbot does need to take these into account to approve or reject responses. For example, if a user asks "What movie should I see?" and the bot replies "Die Hard" and the user then asks "What times?", if the bot does not take into account the previous context it will reply either a wrong answer or an I don't know answer. To solve this problem I propose using a Recurrent Neural Network which "which captures information about what has been calculated so far" (3). This allows the model to take into account what has already occurred in the conversation and equip the model to better accept or reject responses based on this.

I will try several different neural net architectures and select the one that returns the best results. There are also several different types of internals to use with recurrent neural networks so I am going to research different ways to alter a RNN. Different areas of ML will give insight on how an RNN is currently used in different fields and leverage these uses.

### Project Description



### Project References

- (1) <https://skymind.ai/wiki/generative-adversarial-network-gan>
- (2) <https://arxiv.org/pdf/1506.08909.pdf>
- (3) <http://www.wildml.com/2015/09/recurrent-neural-networks-tutorial-part-1-introduction-to-rnns/>
- (4) <https://nlp.stanford.edu/IR-book/html/htmledition/stemming-and-lemmatization-1.html>