AI Case Study

Language-specific, industry-specific, encoding-specific voice recognition system

Project description: There are a number of neural networks out there capable of parsing clearly articulated English as long as high quality microphone is used and common words are spoken. And there are a number of cases where non-English speakers need to communicate non-trivial terms in a specific audio encoding. We built a system for those cases and trained several neural networks to understand what no other AI can understand.

Outcome: Filled a gap in generic AI capabilities with a purpose - trained AI.

Background

Our client specializes in serving healthcare institutions, one of the most regulated industries. At the same time, it has endless opportunities for innovation and, of course, usage of Artificial Intelligence.

Some healthcare operations must be heavily and accurately documented. Our client is building a system that automates note taking. While normally done by a live person, this is something that could and should happen automatically.

Workflow

The project involves **building a speech corpus for a Scandinavian language** and evaluating the effectiveness of using highly compressed speech data using the u-law codec. **The DeepSpeech v1 library** is used for this purpose. The goal is to train and evaluate the neural network performance for the Coqui library over the English speech corpus.

The project also explores the use of **DeepSpeech v2 library** for the same task. The hardest challenge was to fine-tune models using GPU hardware cost-effectively. This was achieved by optimizing the training process and selecting the most efficient hardware for the job.

The speech corpus is a collection of speech samples from speakers of the Scandinavian language. These samples are recorded in a highly compressed format using the u-law codec. The DeepSpeech v1 library is used to preprocess and transform the speech data into a format that can be used for training neural networks.

The neural network performance is evaluated using the **Coqui library** over the English speech corpus. **The Coqui library is a speech recognition system based on the DeepSpeech architecture.** It is designed to work with large amounts of speech data and can be fine-tuned for specific languages.
To fine-tune the Coqui library for the Scandinavian language, the DeepSpeech v2 library is also explored. This library is an improvement over DeepSpeech v1 and offers better performance and accuracy. It is used to optimize the training process and fine-tune the neural network models.

Approach we used

We defined a limited dataset of words, and had a number of people read through it, so that it is easy to identify the beginning and end of every word in the audio files. Then we had those people read out sentences that included the words of interest.

We gathered our own training dataset, as big as it could reasonably be, and combined it with a few 3rd party datasets we have found.

We built our own hardware setup to minimize the training costs.

Then there were rounds and rounds of training and testing the outcome.

Challenges we encountered

While the task of understanding human voice seems trivial nowadays, there were significant complications:

- **Most importantly, we needed to understand the words in Danish**. The amount of source materials is much, much lower compared to English.
- The words we needed to understand were very specific and rare. Generic language models have no idea these words even exist.
- The audio feed went through a legacy audio encoder, a somewhat obsolete format that made it practically impossible to use a generic neural network.

Another challenge in the project was fine-tuning models using GPU hardware cost-effectively. This was achieved by optimizing the training process and selecting the most efficient hardware for the job. The models were trained on GPUs with high computational power, which significantly reduced the training time and cost.

The lessond learned are

Building AI projects from scratch is more challenging than using existing components, datasets, and APIs, but we still managed to achieve our goals and produce a unique audio recognition system.

Wrapping up

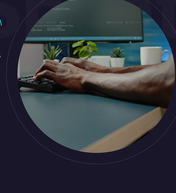
Our client got the technology to increase reliability of healthcare communications and procedures.

While generic AI is omnipresent nowadays, there are numerous gaps and niches that need to be filled.

Team & Spent Time

1	1	1	1
Project Manager	AI Specialist	Back End Developer	DevOps

800	4000+
Hours of work	Hours of neural network training



Let's get started

To coordinate next steps please contact:

Mail: contact@zfort.com

Tel: +1 202 9602900

LinkedIn: zfort-group

ZFort Group - Your reliable partner