August 2022

Simulated Alternatives for Data Augmentation in Machine Learning

Vikram Aggarwal

Yuri Vasilevski

Sukhdeep Singh Sodhi

Ambarish Jash

Simulated Alternatives for Data Augmentation in Machine Learning

ABSTRACT

In domains such as automated speech recognition and search, pre-collected data is often used to understand user input and map it to an intention. In such systems, the context of the user query is important in arriving at the correct response. However, in many use cases, the amount of context to resolve the concept is limited. This disclosure describes data augmentation techniques that increase the amount of data available to understand user input. The techniques described herein can be used to automatically generate data that is adjacent to known correct observations and distinct from most incorrect observations. The techniques described in this disclosure enable growing a seed data set into a much larger corpus.

KEYWORDS

- Data augmentation
- Data Annotation
- Training dataset
- Phonetic distance
- Query understanding
- Virtual assistant
- Automated speech recognition (ASR)

BACKGROUND

In domains such as automated speech recognition (ASR) and search, pre-collected data is often used to understand user input and map it to an intention. Example use cases in automated speech recognition include a user engaging in conversation with a virtual assistant via a smart

speaker or other device. Typical use cases for search include spelling correction, query understanding, or query correction.

In most such systems, user understanding relies on previously collected, labeled data, obtained with specific user permission. For example, such data may be obtained from paid participants or volunteers that provide spoken utterances of known words that are recorded and used for training. In the case of speech recognition, the labeled data includes speech utterances (observations) and the corresponding transcription (labels). In the case of search, the labels are the concepts specified in an internal representation (names of personalities, names of songs, technical terms used in sports etc.), and the words used for these concepts are the observations.

For example, a label X may correspond to the observation "Amitabh Bachchan" and also the observation "अमिताभ बच्चन" (English and Devanagari spellings of the name of the Indian actor), and perhaps also "Amitji," "बच्चन जी" or other known variants of his name. The quality of these mappings determines the eventual quality of the end outcome for downstream applications. If the mappings are too tight, an application may recognize "Amitabh Bachchan" alone, but fail to recognize the similar, but looser variant "Amitav Bachan". On the other hand, if the mappings are too loose, the application might misclassify incorrect alternatives into the label. For example, "Amit Shah Bhashan", which is a completely different concept (a phrase which refers to the concept "speech by Amit Shah") rather than personality mentioned earlier, may get classified into the label for "Amitabh Bachchan."

In such systems, the query context is useful in arriving at the correct response. However, when the context is small, slightly looser mappings may be beneficial. When providing spoken queries, different users speak in different ways (accent, intonation & prosody) while in search, users might be unsure of how to spell something (e.g., a non-English speaker may spell the term

"Coronavirus" as "karona wairus" matching its local pronunciation). In both cases, the amount of context to resolve the concept is limited.

DESCRIPTION

This disclosure describes techniques that can be used to augment data for such systems in an extensible way. The techniques described herein can be used to generate data that is adjacent to correct observations, and far from most incorrect observations.

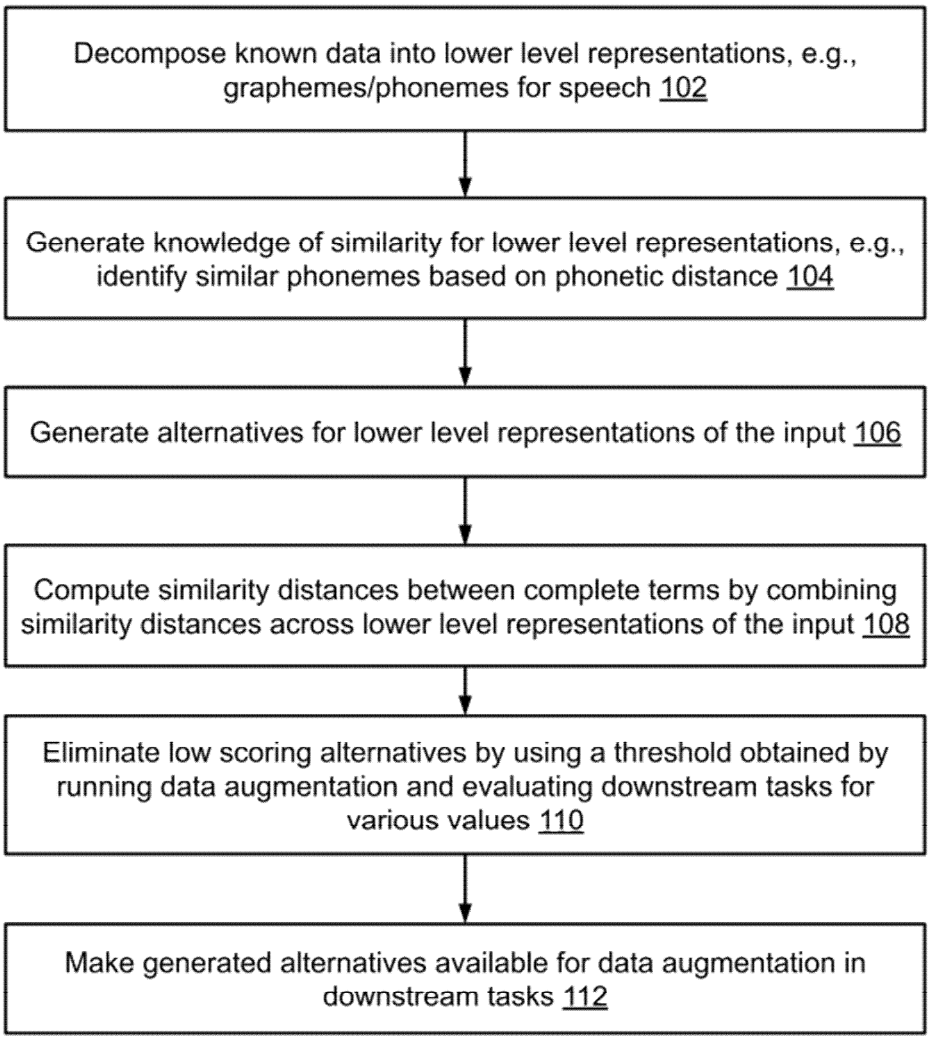

Fig 1. Example process to generate alternatives for data augmentation

Fig. 1 illustrates an example process to generate alternatives for data augmentation, per techniques of this disclosure. First, known data is decomposed (102) into lower-level representations suitable for the application domain. For example, for the speech domain, the observations are broken into graphemes or phonemes, which represent the way that humans speak the content. Next, knowledge of similarity for the lower-level representations is generated (104). For each domain, there are many alternatives that can be used as a measure of the distance between lower-level representations. For example, in the speech domain, the distance can be measured using phonetic distance, accent etc. In most cases, the distance between the representations leads to a natural probability distribution of co-occurrence of the alternatives. Alternatives that are closer to each other appear together more frequently, while alternatives that are far apart do not appear as frequently.

The knowledge of similarity is utilized to generate alternatives for the lower-level representations are generated (106). A broad set of alternatives can be obtained by using a number of different techniques. A combination of some or all of the techniques can be used. Similarity distances for complete terms are computed by combining the similarity distance for the lower-level representations (108).

This enables generating a score that indicates the frequency of each alternative is seen for the phrase, e.g., the relative frequencies with which the alternatives “Amit Shah Bhashan” and “Amitav Bacchan” are seen for the phrase “Amitabh Bachchan.” Alternatives that are associated with less than a threshold score are eliminated (110). The threshold can be obtained by running data augmentation and evaluating downstream tasks on various values. The generated alternatives are made available for data augmentation for downstream tasks (112).

A number of different options are available as a measure of distance among the

alternatives generated for data augmentation. For speech, phonetic distance can be used to identify the phonemes that are similar to others. In languages spoken by many different sets of speakers, similarity can be derived from different accents (e.g., Brazilian Portuguese versus Portuguese from Portugal; different dialects of Hindi; etc.). For speech in noisy situations (outdoor noise, road noise), simulated addition of a specific kind of noise can be performed to observe the corruptions caused in input audio. In case of text, the proximity of keys on the layout on a keyboard can be used to identify similar lower-level representations. When processing handwritten words, visual similarity between letters and numbers could be utilized as a proxy for distance (e.g., 0 and 6 look alike, l and i look alike, e and c look alike, while z and m typically look distinct in handwritten text).

The described techniques can be used for speech recognition in applications such as virtual assistants, speech transcription, or other applications with spoken input. The techniques are also valuable for text queries and to train language understanding systems. Some advantages of the techniques described in this disclosure include:

- Data augmentation is done in advance of the task of query interpretation or other downstream tasks. As a result, query processing does not suffer from increased latency, since data augmentation is often the time-consuming step.
- The downstream task can be evaluated for quality improvement over the un-augmented data. This allows choosing the correct tradeoff between increase of data through augmentation and the improvement in quality achieved through augmentation.
- The generated alternatives can be utilized by any downstream systems where data quality is poor, or the cost of data acquisition is high. For example, audio data is expensive to gather. The techniques described in this disclosure list a way to take a seed data set and

grow it to a much larger corpus.

- The alternatives are generated as necessary. Instead of storing the alternatives, they are created at training time.

CONCLUSION

This disclosure describes data augmentation techniques that increase the amount of data available to understand user input. The techniques described herein can be used to automatically generate data that is adjacent to known correct observations and distinct from most incorrect observations. The techniques described in this disclosure enable growing a seed data set into a much larger corpus.

REFERENCES

1. Ouyang, Yu and Shumin Zhai. “Alternative hypothesis error correction for gesture typing” U.S. Patent Application Publication No. 20180074698, filed November 9, 2017.
2. Siohan, Olivia and Pedro J. Moreno Mengibar. “Speech recognition using associative mapping” U.S. Patent Application Publication No. 20160171977, filed February 22, 2016.