What is Data Augmentation?

Definition

Data augmentation (DA) encompasses methods of increasing training data diversity without directly collecting more data. Most strategies either add slightly modified copies of existing data or create synthetic data, aiming for the augmented data to act as a regularizer and reduce overfitting when training ML models.

Trades off: As DA aims to provide an alternative to collecting more data, an ideal DA technique should be both easy to implement and improve model performance. Most offer trade-offs between these two.

Techniques

- **Rule-based technique**

  DA primitives which use easy-to-compute, predetermined transforms sans model components. Feature space DA approaches generate augmented examples in the model's feature space rather than input data.

  - Synonym Replacement

    Original sentence:
    "The product is excellent."

    Augmented sentences:
    "The product is outstanding."
    "The product is superb."

  - Random Insertion

    Original sentence:
    "The delivery was fast."

    Augmented sentences:
    "The quick delivery was fast."
    "The delivery was super fast."

  - Random Deletion

    Original sentence:
    "The customer service was helpful and responsive."

    Augmented sentences:
    "The customer service was and responsive."
    "The customer service was helpful and."

  - Sentence Swap

    Original sentence:
    "I would highly recommend this product. It's worth the price."

Augmented sentences:

"It's worth the price. I would highly recommend this product."
"This product is worth the price. I would highly recommend it."

- **Example Interpolation Techniques**

Interpolates the inputs and labels of two or more real examples. This class of techniques is also sometimes referred to as Mixed Sample Data Augmentation (MSDA). It involves combining pairs of examples from the original dataset to create new examples.

Let's say you have a dataset of customer reviews for a product, and you want to augment the data using Example Interpolation:

Select two random reviews from the original dataset that have similar sentiments or ratings.

Original reviews:

Review 1: "This product exceeded my expectations. It's amazing!"

Review 2: "I'm really impressed with this product. It's fantastic!"

Interpolate the sentences from the two reviews to create a new review. This can be done by combining corresponding sentences from the two reviews.

Interpolated review:

New Review: "This product exceeded my expectations. I'm really impressed with this product. It's fantastic!"

Repeat this process multiple times with different pairs of reviews having similar sentiments or ratings.

By applying Example Interpolation to the text data, you can create new reviews that retain the overall sentiment and characteristics of the original reviews while introducing some variations in the specific phrases and expressions used.

Again, note that the example provided is a simplified representation, and the actual implementation may require additional preprocessing and consideration of the specific structure and characteristics of your text data.

- **Model-Based Techniques**

Involve leveraging pre-trained models to generate augmented examples.

Example:

Select a pre-trained language model, such as GPT-3, that has been trained on a large corpus of text data.

Choose a prompt or a partial sentence from your original dataset.

Original prompt: "The weather is"

Use the pre-trained language model to generate completions for the prompt. You can sample from the model's output to introduce variation.

Augmented examples:

"The weather is beautiful today."

"The weather is unpredictable lately."

"The weather is cold and rainy."

Repeat this process with different prompts from your original dataset or create new prompts to generate additional augmented examples.

By utilizing a pre-trained language model, you can generate new text examples that are coherent and maintain the linguistic style and patterns of the original data. This technique can be particularly useful when you need to generate large amounts of diverse text data or when you want to explore different variations of the given prompts.

Keep in mind that the availability of pre-trained models may vary, and you may need to fine-tune or adapt the model to your specific task or domain for optimal results.

Applications

Below are some NLP applications that can be solved using DA methods.

NLP application refers to a broader system or software that utilizes one or more NLP tasks to address real-world problems or provide value-added functionality. NLP applications integrate multiple NLP tasks and techniques to solve complex language-related challenges.

- **Low-Resource Language:** Refers to a language for which there is limited availability of linguistic resources, such as annotated corpora, lexicons, and pre-trained models. In a much simpler way, it is a language for which there is very little data and limited tools available for building language models and performing natural language processing tasks. These languages are often spoken by smaller populations and have not received as much attention in terms of research and development compared to widely spoken languages.
- **Mitigating Bias:** Refers to the process of identifying and reducing or eliminating biases that may exist within systems, processes, or algorithms. In the context of AI and machine learning, bias can arise from various sources, including biased training data, biased features, or biased algorithmic decisions. Mitigating bias is crucial to ensure fairness, equity, and ethical considerations in AI applications.

- **Fixing Class Imbalance:** Refers to addressing the unequal distribution of classes in a dataset, particularly when one class is significantly underrepresented compared to the other(s). Class imbalance can pose challenges in machine learning tasks, as models tend to be biased towards the majority class, resulting in poor performance in the minority class.

- **Few-Shot learning:** Refers to a machine learning paradigm that focuses on training models to learn from a limited amount of labeled examples (few shots) for novel or unseen classes. Unlike traditional machine learning approaches that require a large amount of labeled data for each class, few-shot learning aims to generalize from a few labeled examples and adapt quickly to new tasks or classes.

- **Adversarial Examples:** Refer to specially crafted inputs that are designed to mislead machine learning models. These inputs are typically created by introducing imperceptible perturbations to the original inputs, which can cause the models to make incorrect predictions or classifications.

Tasks

NLP task refers to a specific problem or objective that involves processing and understanding natural language text. NLP tasks focus on specific aspects of language analysis and aim to accomplish a particular linguistic or computational goal.

- **Summarization:** Refers to a technique where a given piece of text, such as a document or a paragraph, is summarized or condensed to create a shorter version of the original text. This technique can be used as a form of data augmentation to generate new training examples by summarizing the existing ones.
    - Original Text: "A study was conducted to investigate the effects of exercise on cardiovascular health. The results showed that regular exercise can significantly reduce the risk of heart disease and improve overall heart function."
    - Summary: "Regular exercise lowers heart disease risk and improves heart function, according to a study."

- **Question Answering:** Involves generating new training examples by transforming the original data into question-answer pairs. This technique can be used to expand the training dataset and introduce additional variations to improve the performance of QA models.
    - Original Text: "The capital of France is Paris. It is known for its beautiful architecture, rich history, and vibrant culture."
    - Question: "What is the capital of France?"
    - Answer: "Paris”

- **Sequence Tagging Task:** This involves assigning labels or tags to each element in a sequence of tokens, such as words, characters, or subwords, within a given text. The objective is to identify and label specific entities, parts of speech, or other linguistic elements in the text.
    - Example: POS tagging

- Input Text: "I like to eat pizza."
- Output Tags: ["PRON", "VERB", "PART", "VERB", "NOUN", "PUNCT"]
- "I" is labeled as "PRON" (pronoun) since it represents a personal pronoun.
- "like" is labeled as "VERB" since it is a verb.
- "to" is labeled as "PART" since it is a particle or a preposition.
- "eat" is labeled as "VERB" since it is a verb.
- "pizza" is labeled as "NOUN" since it is a noun.
- "." is labeled as "PUNCT" since it is a punctuation mark.

- **Parsing Tasks:** Involve analyzing the syntactic structure of a sentence or text and building a formal representation of its grammatical structure. Parsing aims to determine the relationships between words and phrases in a sentence, identifying the role of each word and how they relate to each other.
  - Input Sentence: "The cat chased the mouse."

```
            Sentence
               |
          +----+----+
          |         |
         NP        VP
          |         |
       +--+--+   +--+--+
       |   |   |   |
      Det  Noun  Verb  NP
       |   |   |   |
       |   cat  chased the mouse
```

- **Grammatical Error Correction:** NLP task that aims to automatically identify and correct grammatical errors in the text. The goal of GEC is to improve the grammatical accuracy and fluency of written text by detecting and rectifying errors in grammar, syntax, punctuation, or usage.
  - Input Sentence: "He is going to the park yesterday."
  - Corrected Sentence: "He went to the park yesterday."

- **Neural Machine Translation:** Utilizes artificial neural networks to automatically translate text from one language to another. It has revolutionized the field of machine translation and has become the dominant paradigm in recent years.

  The back translation method is an effective data augmentation technique for NMT, as it enables the model to learn from both the original parallel data and the generated synthetic data. This approach has been shown to enhance the performance of NMT models, especially in low-resource language pairs where the availability of parallel training data is limited.

- **Data-to-Text NLG:** A subfield of natural language processing (NLP) that focuses on automatically generating human-readable text from structured data or information. The goal is to transform structured data, such as databases, spreadsheets, or knowledge graphs, into coherent and understandable textual descriptions.

- **Open-Ended & Conditional Generation:** Open-ended generation in natural language generation (NLG) refers to generating text without specific constraints or predefined structure, allowing for creative and varied output. It enables systems to freely generate responses or text based on learned patterns and knowledge. On the other hand, conditional generation involves generating text based on specific conditions or constraints, allowing for more controlled and targeted output that adheres to predefined templates or criteria. Conditional generation provides control and structure, while open-ended generation promotes creativity and diversity. The choice of approach depends on the desired output and purpose of the NLG application.

- **Dialogue:** Involve understanding and generating human-like responses in a conversational setting. These tasks aim to enable machines to engage in interactive and contextually relevant conversations with users.

- **Multimodal Tasks:** Involve the integration and analysis of multiple modalities, such as text, images, audio, video, or other forms of data, to extract meaningful information and enable a deeper understanding of the content. These tasks aim to leverage the complementary nature of different modalities to enhance the accuracy and richness of NLP models.

References

https://arxiv.org/pdf/2105.03075.pdf

What is Back Translation?

Definition

Back translation, also known as reverse translation or dual translation, involves translating content, whether it is a query or paragraph, from one language to another and retranslating it to the original language. This method provides several options for the owner to make a decision that makes the most sense based on the task at hand.

Sometimes back translation may be confused by double translation. Double translation refers to the process of translating a text twice using two different translators or translation systems.
The process is as follows:

- Take a source text in language A.
- Translate the source text to language B using translator 1.
- Translate the translated text from language B to language A using translator 2.
- Compare the double-translated text with the original source text and make any necessary corrections.

The goal of back translation is to evaluate the translation models' performance and make content persuasive, effective, and relevant. However, double translation is used to enhance the quality of translations.

Other types of content might require back translations. Industries with important, high-risk content that requires perfection — for example, in the medical field, where mistranslations could mean the difference between life and death. For certain industries, regulatory requirements, ethics committees, and institutional review boards mean back translations must be a part of any translation process.

References

https://www.smartling.com/resources/101/what-is-back-translation-and-why-is-it-important/#:~:text=Back%20translation%20is%20often%20used,the%20meaning%20of%20the%20translation
https://lokalise.com/blog/back-translation-best-practices/
https://www.pactranz.com/back-translation/#:~:text=Fields%20where%20back%20translation%20is%20common%3A%201%20pharmaceutical,packaging%20for%20export%208%20food%20products%20More%20items

Query Refinement

https://queryunderstanding.com/query-rewriting-an-overview-d7916eb94b83
https://www.wordstream.com/query-refinement