

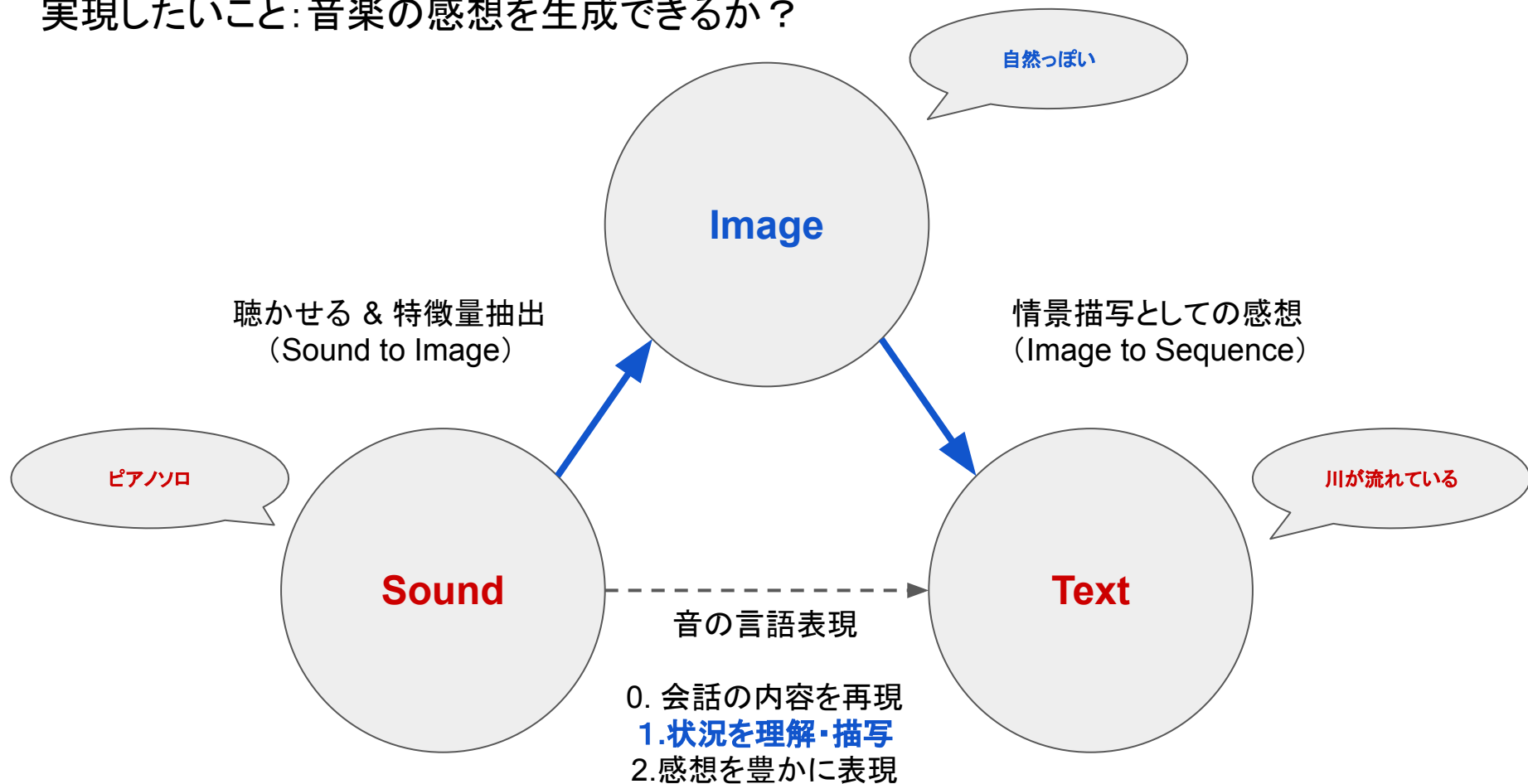
Oto no “Sommelier”

Sound 2 Image 2 Sequence

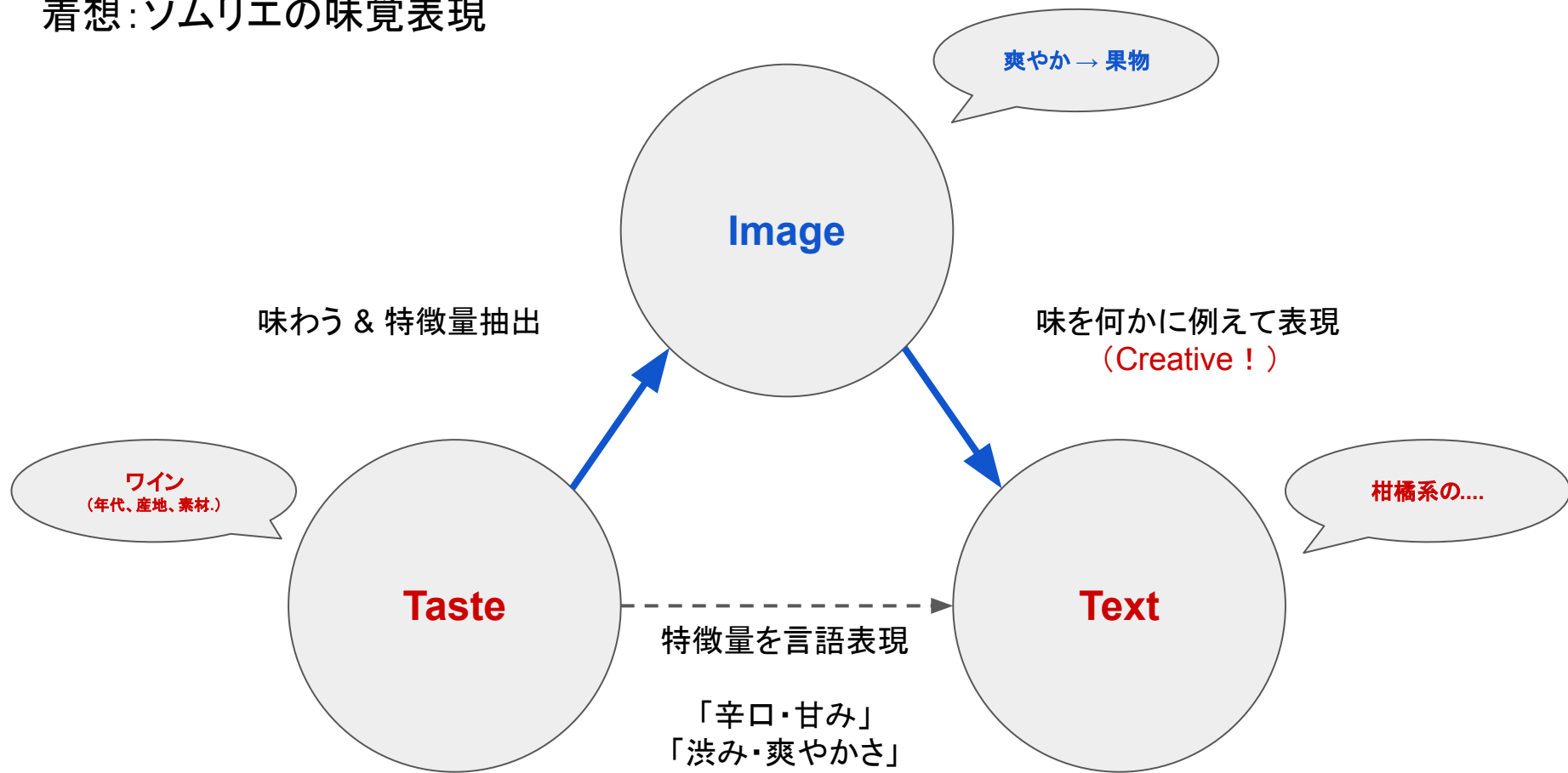
DIVE INTO CODE
Term 2019.4

Makoto Shirasu

実現したいこと: 音楽の感想を生成できるか？



着想:ソムリエの味覚表現



なぜやるのか？

1. 機械学習の基礎技術(音声認識 & 画像認識 & 自然言語処理)をフルパッケージで使ってみたい。

=> end2end(音声認識 → 自然言語)よりも、画像(=例え)を経由することで表現力が増すのでは？

2. 音楽の感想を述べることは、なかなか難しい。

=> 音楽評論家：職業柄、専門的な知識に裏付けられたコメント。一般の人は

=> 一般の人(TwitterやFacebookのSNS)：「よかった！」「上手い！」=> もう少し豊かにできる？

3. 感想表現は人間らしい創造的な営み(正解はない)。物事の感じ方の違いにこそ個性が現れる。

=> 問題意識：現代社会は客観性偏重？(その裏返しにアート、クラフトに対する関心が高まっている)

=> 少しだけ言葉豊かに感想を表現し、共有することを通じて、主観を取り戻すきっかけになればいい。

Sprintで一番感動した瞬間

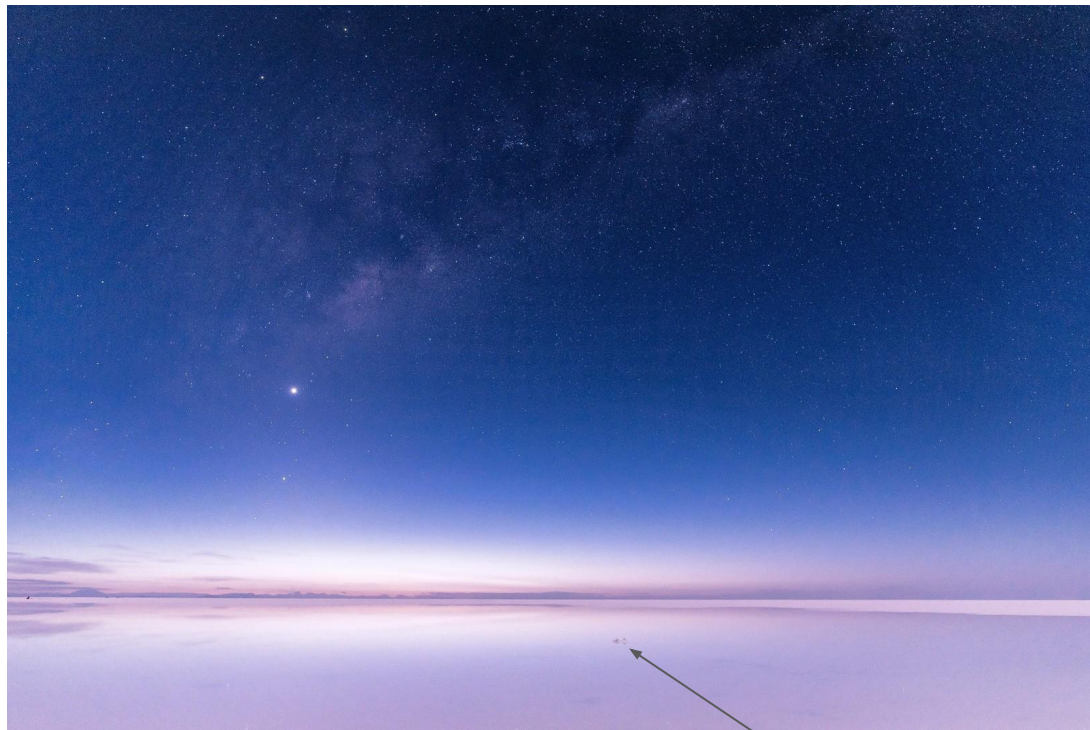


Image:

ウユニ塩湖

Caption:

A view of a lake with a boat in the background.

Tecnology:

Image to sequence

pytorch-tutorial/tutorials/03-advanced/image_captioning_at_master · yunjey/pytorch-tutorial

ボート?? (そうは見えないけど ...)

要件定義:

- ・選択肢は? : ①入力した音から状況を推測する、②豊かな感想を生成する

=> 今回は、まずシンプルに①の実現を目指す。

- ・重要な評価指標は?

=> 感想なので、正確性よりも**類似度(=〇〇っぽい)**が重要

- ・どこに力を入れる?

①モデルのつなぎこみ(音声 => 画像 & 画像 => シーケンス)

②音声データの取り扱い(前処理、モデル)

③データ収集

プロジェクトの全体感(構想)

1. **データ収集**: どのようなデータ(音声・画像・動画)をどこから取得する？

=> Youtube ? Kaggle ? Free Sample ?

2. **前処理**: 音声データの前処理、学習データ(画像・音声の対応関係)の生成

=> Kaggle (FreeSound主催) のコンペから学ぶ。

3. **モデル構築・予測・評価**: 特に、どのようなモデルを構築する？

=> Image to Sequenceをアレンジして、"Sound => Image => Sequence"

4. **その他**: 参考になりそうな論文を読む & まとめて、ヒントを得る。

Expansion (Not to Do)

1. 言語表現の豊かさの追求

=> 言葉の表現力を高めようとする、生成モデルを用いる必要があるかも。

2. リアルタイム性の追求(マイクで拾った音声をリアルタイムで変換する)

=> まずは、音声ファイルを読み込ませて出力すること。

おわり