公立はこだて未来大学 2023 年度 システム情報科学実習
グループ報告書

Future University Hakodate 2023 Systems Information Science Practice
Group Report

プロジェクト名

脳をつくるプロジェクト 2023

Project Name

Make Brain Project 2023

グループ名

グループ 13

Group Name

Group 13

プロジェクト番号/Project No.

13-C

プロジェクトリーダ/Project Leader

太田怜志　Reiji Ota

グループリーダ/Group Leader

井戸智斗志　Satoshi Ido

グループメンバ/Group Member

今野光琉　Hikaru Konno
井戸智斗志　Satoshi Ido
小齋友里菜　Yurina Kosai
稲井嵐堂　Rando Inai

指導教員

香取勇一　栗川知己　加藤譲　佐々木博昭　冨永敦子　ヴラジミールリアボフ　佐藤直行

Advisor

Yuichi Katori　Tomoki Kurikawa　Yuzuru Kato　Hiroaki Sasaki　Astuko Tominaga
RIABOV, B Volodymyr　Naoyuki Sato

提出日

2024 年 1 月 17 日

Date of Submission

January 17, 2024

概要

本プロジェクトでは、長い動画から短い動画を生成する人工知能の開発に取り組んだ。最近、倍速視聴をする若者が増えている。若者は講義動画だけでなくアニメや映画などの映像作品でも倍速視聴を行う傾向がみられる。倍速視聴をする要因として、限られた時間内でより多くの情報を得ようとするタイムパフォーマンスを求めるためである [1]。情報の伝達方法が新聞や本だけではなく、ニュースや YouTube と映像化した。映像化したことにより、最終的に情報が伝えたいことを把握しにくくなった。倍速視聴することにより、結末を早く知るため、倍速視聴を行っている。倍速視聴の需要に着目し、長い動画を短く編集し、効率よく情報を得ることが可能となるようなシステムを開発することにした。

まず動画を短いものにするのにあたって、特徴物や前後で変化した以外のものを消すことにした。検証しやすさとプロジェクト学習の目的ををを考え、編集される動画は観光動画とすることとした。観光動画であれば、中にある観光物を切り出して集めることで全体が把握できると考えたからである。

次に動画内にある物体をコンピュータに特徴物を認識してもらうためには元となる画像を収集する必要がある。そのためグループメンバーで市内の観光物を撮影して画像データを収集し、加工し、画像認識を可能とする学習データを作成した。このデータを用いて学習を行ったところ、模擬の画像で認識していることを確認した。システムに動画内の音声による説明を文章に変換し、要約を行ったものを短く編集した動画の適切な箇所に貼り付けるシステムを加えた。このシステムが一般人が利用できるサービスであるかを確かめるために比較実験を行った。元動画と短く編集した動画を視聴してもらい、情報量に違いはあるのか、見やすさはどうかなどのアンケートを実施した。

結果として動画内にある物体の画像だけを抜き出し集めたため、情報量が多い動画となってしまった。特に異なる画像が出力される上に、文字情報まで加わるため一度の処理に対する負荷が大きいと評価された。また動画としての良さを失っているとも評価された。しかし動画内の観光物を抜き出し、短く編集出来ているため目標は達成できている。そのため情報量を減らした上で動画の良さを出すために、画像を出力するのではなく５秒程度の動画をつなぎ合わせることで、改善できるのではないかと考えている。

本グループは画像認識を一般人が利用できるサービスの実現に向けて活動した。その結果、函館市内の観光動画に関しては画像認識を行い、全体が把握できる動画の作成に成功した。見やすさが欠けているため改善が必要ではあるが、変更を加えることで一般人が利用できるサービスが実現できると考えている。また今後函館市内だけではなく、他の観光名所の画像データを収集することで汎用的に利用可能になっていくと考えている

キーワード　倍速視聴, タイムパフォーマンス, 画像認識

（※文責: 井戸智斗志）

Abstract

In this project, we worked on the development of artificial intelligence to generate short videos from long videos. Recently, the number of young people who watch videos at double-speed has been increasing. It has been observed that young people tend to watch not only lecture videos but also animations, movies, and other video works at double-speed. One of the reasons for double-speed viewing is the desire for time performance, in which more information is obtained in a limited amount of time [1]. The method of information transmission is no longer limited to newspapers and books, but has shifted to news and YouTube. The visualization of information has made it harder to grasp what the information is ultimately trying to convey. The information is viewed at double speed so that the reader can quickly grasp the end result. Focusing on the demand for double-speed viewing, we decided to develop a system that would enable people to edit long videos into shorter ones and obtain information more efficiently.

First, in making the video short, we decided to delete features and anything else that had not changed before and after the video. Considering the ease of verification and the purpose of the project study, it was decided that the video to be edited would be a sightseeing video. We thought that if the video was a sightseeing video, we would be able to grasp the whole picture by cutting out and collecting the sightseeing objects in the video.

Next, it was necessary to collect the original images in order to have the computer recognize the features of the objects in the video. Therefore, group members took pictures of tourist attractions in the city, collected and processed the image data, and created training data that would enable image recognition. When training was conducted using this data, it was confirmed that the system was able to recognize the simulated images. A system was added to the system that converts the audio description in the video into text, summarizes it, and pastes it into the appropriate section of the shortened and edited video. A comparison experiment was conducted to see if this system is a service that can be used by the general public. We asked the participants to watch the original video and the shortened video, and conducted a questionnaire to find out if there was any difference in the amount of information and how easy it was to watch.

As a result, only the images of objects in the video were extracted and collected, resulting in a video with a large amount of information. The result is a video with a large amount of information. In particular, since different images are output and textual information is also added, the processing load for one time was evaluated to be too large. The video was evaluated as having a heavy burden for a single processing. It was also evaluated that the video lost some of its quality as a moving image. However, if the tourist attractions in the video were extracted and edited shortly The goal was achieved, however, because the tourist attractions in the video were extracted and edited in a short time. In order to reduce the amount of information Therefore, in order to reduce the amount of information and bring out the quality of the video, we thought it might be possible to improve the quality of the video by joining together 5-second video clips instead of outputting images. We believe that we can improve the quality of the video by joining five-second video clips instead of outputting images.

This group worked toward the realization of a service that allows the general public to use image recognition. As a result, the group succeeded in creating a sightseeing video of Hakodate city that can be grasped in its entirety through image recognition. However, the video lacks visibility and needs to be improved, but with some modifications, the group believes that a service that can be used by the general public will be realized. In addition to Hakodate City, we believe that the service can be used for general purposes by collecting image data from other tourist attractions in the future.

Keyword Speed Watching, Time Performance, Image recognition

（※文責: 井戸智斗志）

目次

第 1 章　背景

近年、過程よりも早く結果を求めたがる若者が多い。例えば、近畿大学において実施された授業評価アンケートの結果によると、講義動画を視聴する際に倍速機能を使用して視聴していた学生は約半数を占めていた [2]。しかし、倍速視聴では重要な情報を見逃しやすくなり、効果的な学習や情報の習得において問題がある。また学生は講義動画を倍速視聴するだけではなく映画やアニメなどでも倍速視聴が行われている。このような傾向はタイムパフォーマンスが良いという理由から行われる。つまり若者は映像を楽しみたいというよりもタイムパフォーマンスを求めていることが分かる。そして、倍速視聴を行う原因は、情報を映像化したことで情報を受け取る側が全体像を把握できなくなったためである。例えば、文字によって情報を伝達する新聞では、一覧表や見出しにより全体像が把握しやすい作りとなっている。それに対して、動画にはそれらがないため把握が困難である。また情報を受け取る速さも媒体に依存するようになった。例えばテレビのニュースなどはテレビ局の放送する速さに依存し、自分で速さを制御したり、不要な情報をスキップということが不可能である。これらへの解決策として若者は倍速視聴を行っている。映像から特徴があるところだけを画像情報として取り出して短い映像にすれば全体像を短期間で把握できる上に映像的な特徴も活かすことができる。

（※文責: 井戸智斗志）

1.1　該当分野の現状と従来例

NEC では長時間の映像からユーザーの要望に沿ったシーンのみを抽出し、さらにそのシーンを説明する要約文まで自動生成できる技術を開発した。この技術はコンピュータが映像内で起きていることを認識し、ユーザが要求する箇所を取り出す。加えて要約文章までを自動的に生成するものである [3]。活用先としては保険調査です。ドライブレコーダーの映像から事故発生時のシーンを抜き出した上に、事故当時の説明文を自動生成することが出来る。これを事故調査報告書の活用に利用することで、調査員は少しの変更だけで報告書を作成することが可能となる。また製造業に応用することで品質管理から書類の作成を一括で行うことが可能となり、映像をすべて確認する必要がなくなる。また看護や介護に応用することで日報の作成時間を削減することが可能となり、サービスの向上に努めることができる。この技術の最大の特徴はチャット形式で行えることにある。AI と会話するかのように書類作成ができるため、変更が容易となる。

（※文責: 井戸智斗志）

1.2　現状における問題点

NEC が開発した技術は幅広い分野では応用可能ではあるが、一般人が利用できるサービスではないと考えている。理由は業務の効率化が主眼に置かれているからである。業務の効率化ゆえに、チャットをしながら資料を作成する業種に限られてしまう。つまり従来は人間が確認をして手入力をしていた作業を、AI が自動で行い人間が修正することで短時間で終わらせることがこの技術の

魅力である。しかし広く一般の方に利用してもらうためには、作品的面白味が欠けていると考えている。サービスを利用することで画像の抽出ではなく、短い動画へ編集することで利用レベルを下げることが出来ると考えている。

（※文責: 井戸智斗志）

1.3 課題の概要

まず、入力する動画が、観光地の動画や試合の動画、講義動画と様々なジャンルがあるため、ジャンルごとに抽出することが問題である。これは、ジャンルによって、動画の内容や特徴が異なるためである。例えば、観光地の動画では、風景や建物などの物体が重要であるが、試合の動画では、選手や得点などの動きが重要である。また、講義動画では、音声やテキストなどの言語が重要である。このように、ジャンルによって、動画要約に必要な要素が変わるため、ジャンルごとに抽出することが必要である。

次に、元の長い動画と、要約した短い動画を見た理解度を一致させることが挙げられる。これは、動画要約の目的は、動画の内容を簡潔に伝えることであるためである。理解度を一致させるために、動画の内容を要約する際に、例えば、動画の内容に関連しない物体や音声を除外するという工夫が必要である。この工夫は、動画の内容を理解するために、不要な情報を削ることが重要であるためである。動画要約によって、動画の情報量は減るが、動画の価値は減らないようにすることが重要である。

さらに、入力した元の動画と出力された要約動画の解釈の差を埋めるためには、不要な部分である切り抜きの工程が重要となる。切り抜きとは、動画の一部を切り取ることである。切り抜きによって、動画の長さを短くすることができる。しかし、切り抜きには、注意が必要である。切り抜きによって、動画の内容が変わってしまう場合があるからである。例えば、試合の動画では、試合の流れや結果を変えるようなシーンを切り抜いてしまうと、動画の内容が正しく伝わらない可能性がある。また、講義動画では、講義の内容や構成を変えるようなシーンを切り抜いてしまうと、動画の内容が分かりにくくなる可能性がある。このように、切り抜きによって、動画の内容が損なわれる場合があるため、切り抜きの工程は重要となる。本グループは、切り抜きの工程において、動画の内容を損なわないように、切り抜きのタイミングや範囲を工夫する必要があった。

動画の音声や抽出した画像をどこの部分に貼るかといった構成も重要となる。構成とは、動画の要素をどのように配置するかということである。構成によって、動画の見やすさや魅力が変わる。例えば、観光地の動画では、音声や画像を動画の雰囲気に合わせて配置することで、観光地の魅力を高めることができる。また、講義動画では、音声や画像を動画の内容に合わせて配置することで、講義の理解度を高めることができる。このように、構成によって、動画の効果を最大化することができる。構成の工程において、動画の目的やジャンルに応じて、音声や画像の選択や配置を工夫することが重要である。

最後に、動画要約をする AI に学習させる過程で、画像と動画間の関連性も学習させることが挙げられる。画像と動画間の関連性とは、画像が表す物体や場面が、どのように関係しているかということである。画像と動画間の関連性を学習させることで、動画要約において、より適切な画像を選択することができる。例えば、観光地の動画では、同じ観光地の画像を選択することで、観光地の雰囲気を伝えることができる。また、試合の動画では、試合の流れに沿った画像を選択することで、試合の展開を伝えることができる。さらに、講義動画では、講義の内容に関連した画像を選択

することで、講義のポイントを伝えることができる。このように、画像と動画間の関連性を学習させることで、動画要約の質を向上させることができる。

（※文責: 今野光琉）

第 2 章　到達目標

2.1　本プロジェクトにおける目的

1.3 節の課題を達成するために本グループでは目標を 2 段階で分けて活動を行った。まず、1 つ目の目標は 3 章のシステムを早期に実現することであった。具体的には、入力した動画から独自に作成したデータセットを用いてコンピュータに学習させることにより特徴物だけを切り抜き、短い動画を出力させる。不格好な動画が出力されることを許容し、短時間で動くモノを作り出すことを目指した。2 つ目の目標は、1 つ目の目標で作成したシステムを修正することである。具体的には切り抜いた画像の配置場所を信頼度にしたがって決定し、視聴者が見やすい動画を作成する。また添付するテキストは 5 秒以内で読み終えることが可能な要約文とする。そうすることで図 2.1 で示すように 10 分の動画を 30 秒で概要把握出来る動画を作成できるシステム開発を行う。

（※文責: 井戸智斗志）

図 2.1　目標

2.1.1　通常の授業ではなく、プロジェクト学習で行う利点

グループ C のプロジェクトでは、データセットを一から作る必要があり、機械学習モデルに学習させるため、膨大な時間が必要となる。そのため、通常授業で行うことは向いていない。そして、現代の倍速視聴を行う需要から、倍速視聴から生じる問題をグループ C 自ら見出し、1 年間で問題を解決すべくシステムを開発するため、プロジェクト学習で行う意義がある。また、函館の観光動画を入力し、出力で函館の観光スポットが短時間で把握できれば、函館の観光業を活性化できるという点で、函館への貢献が見込める。つまり、プロジェクト学習の動機づけに記載されている、社会的に意味のある環境と活動の中でより強く学ぶことと一致する [4]。

（※文責: 小齋友里菜）

2.2　具体的な手順・課題設定

動画要約を行うアーキテクチャを構築することを目的として、以下のように手順を設定した。

1. 映像の要約化手法の調査
   まず、動画要約に関する過去の研究や関連文献を調査する。具体的には、動画要約には、どのような手法やアルゴリズムが提案されているのか、そして、その中でも代表的なものはどんなものなのかを把握する。また、動画要約に必要な技術として、要約画像を生成する技術、音声をテキストに変換する技術、テキストから重要なキーワードを抽出する技術などを検討する。
2. データセットの作成
   次に、要約対象となるさまざまなジャンルの映像データを収集する。本プロジェクトでは、函館の観光動画の要約することをテーマとした。そのため、函館の観光物を撮影環境などが異なるように撮影し、データとして用いる。これは、動画要約の汎用性や頑健性を検証するためである。収集したデータに対して、VoTT と呼ばれるアプリケーションを用いて、画像内の物体やテキストにアノテーションを付けた。アノテーションとは、データに対してラベルやメタデータなどの情報を付与することである。アノテーションを行うことで、学習データとして使用できるように準備する。
3. 物体検出モデルの学習
   次に、物体検出のための機械学習モデルを訓練する。物体検出とは、画像や動画の中に存在する物体の位置と種類を同定する技術である。物体検出は、動画要約において、動画の内容を理解するための重要な手がかりとなる。収集したデータセットを使用して、物体検出のための機械学習モデルを訓練する。機械学習モデルとは、学習したパターンや規則を表現する数学的なモデルである。本学における、画像認識やデータサイエンス入門、データサイエンス演習の講義で学んだ技術や、YOLO[5] などの物体検出アルゴリズムを適用し、特徴的な物体やテキストを検出できるようにする。
4. アーキテクチャの構築
   次に、調査した手法、検討した技術、訓練した機械学習モデルを組み合わせ、動画要約を行うアーキテクチャを構築する。アーキテクチャとは、システムの構造や機能を表す図や仕様のことである。動画要約を行うアーキテクチャは、おおまかに分けて、動画の前処理、動画の要約、動画の後処理の 3 つの部分から構成される。動画の前処理とは、動画を要約するために必要な情報を抽出する部分である。例えば、動画をフレームに分割し、物体やテキストを検出し、音声をテキストに変換するなどの処理を行う。動画の要約とは、動画の内容を短時間に圧縮する部分である。例えば、各フレームで検出を行い、重要な物体を選択し、要約画像を生成するなどの処理を行う。動画の後処理とは、要約した動画を整形する部分である。例えば、要約画像を連結し、テキストを付加するなどの処理を行う。
5. 要約動画の評価
   要約動画の評価を行う。評価には、被験者に対して、評価用の映像と構築したアーキテクチャを用いて要約した評価用の映像を比較してもらい、元の動画と要約動画の間で、どれだけ同じ情報を伝えられるかという指標によって評価する方法を採用する。そして、質的・量的な評価を行い、評価の結果をもとに、改善点や要約の効果を把握する。
6. 改善とバリエーションの検討
   次に、改善とバリエーションの検討を行う。改善とは、評価の結果やフィードバックを元に、要約手法や生成結果の問題点を解決することである。例えば、物体検出の精度を向上させることや、要約画像の品質を改善することなどが挙げられる。バリエーションとは、要約パターンの多様性を増やすことである。例えば、要約の長さやスタイルを変えることや、要

約に含める要素を変えることなどが挙げられる。改善とバリエーションを行うことで、動画要約の柔軟性や魅力を高めることができる。

7. バリデーションと最終調整
   最後に、バリデーションと最終調整を行う。バリデーションには、動画要約の評価と同様の内容のアンケートを実施し、要約動画の改善点を見つける方法を採用する。要約動画の好評評価が一定以上に達するまで、改善と評価の繰り返しを行いながら最終的なシステムを調整する。

これらの作業を行うことで、動画要約を行うアーキテクチャの構築を完了する。

（※文責: 今野光琉）

2.3 課題の割り当て

各人の得意分野及び関連性、時間軸のスケジュールを基準に以下のように割り当てた。

井戸智斗志　リーダーシップを発揮できることからメンバーをまとめ、プロジェクトを進めた。函館駅周辺のアノテーション作業、システム評価のための動画制作を行った。

今野光琉　プログラム作成が得意なことから手法の調査、プログラム及びデータセットの制作に取り組んだ。

小齋友里菜　データセットに必要な素材の収集を行った。五稜郭駅周辺のアノテーション作業を行った。システム評価のための動画制作を行った。

稲井嵐堂　国語能力が高いことから発表原稿と報告書の添削に取り組んだ。のアノテーション作業を行った。システム評価のための動画制作を行った。

（※文責: 井戸智斗志）

第 3 章　課題解決のプロセスの概要

2.2 節で具体化した各小課題の解決のプロセスの概要を、各々記述する。

1. 映像の要約化手法の調査
   解決過程：インターネットから関連する資料やライブラリを検索した。
2. データセットの作成
   解決過程：インターネットから画像をダウンロードしてアノテーションを行った。しかしインターネットの画像では種類が不足したため、グループメンバーで函館の観光地で写真を撮り、データセットを作成した。
3. 物体検出モデルの学習
   解決過程：グループメンバーで作成したデータセットを YOLOv8 を用いてトレーニングさせた。
4. アーキテクチャの構造
   解決過程：動画内の物体やテキスト本プロジェクト用に改変した物体検出用の機械学習モデルである YOLO を適用して特徴的な物体やテキストを検出する。 そして抽出した画像を画像処理ライブラリである rembg を用いて、背景をトリミングし、一枚の要約画像を生成するにあたって十分な量の物体やテキストを検出したら、それぞれの物体やテキストが重複しないように 1 枚の画像にまとめる。その後、FFmpeg を用いて検出が終わるまでの範囲の動画を音声ファイルに変換し、テキスト化のライブラリである Speech Recognition を用いて音声ファイルをテキストに変換する。その際に変換したテキストを用いて、動画がポジティブであるかネガティブであるかを判断し、要約動画に用いる BGM を選択する。そして、選択した BGM と要約したテキストと画像を組み合わせて、5 秒程度の要約動画を生成する。これらの処理を分割したフレーム全てに対して検出が終わるまで繰り返し、生成した順序に従い動画を組み合わせ出力する。
5. 動画要約の評価
   解決過程：被験者に元映像と要約映像を比較してもらい、２つの情報の受け取り量の近似性を５段階で評価してもらう。種類は観光、講義、ニュースの３つを予定している。またなぜそのような評価をしたのかを記入してもらい、質的・量的の両面で評価し、今後の改善点を見つけていく。
6. 改善とバリエーションの検討
   解決過程：評価を受けて修正すべき点を洗い出し、変更を行う。そして複数の動画ジャンルにも対応できるようにするために、改善のどこを上限とし、プロジェクトを終了するかを決定する。具体的には観光だけに特化したシステムにするのか、講義やニュース以外にも対応出来るようにするかなどを決める。
7. バリテーションと最終調整
   解決過程：評価とシステム修正を繰り返して５割の被験者が好評価となるまで修正を行い続ける。

（※文責: 井戸智斗志）

第 4 章　課題解決のプロセスの詳細

4.1　各人の課題の概要とプロジェクト内における位置づけ

井戸智斗志の担当課題は以下のとおりである。主にチームリーダーとして、プロジェクトのタスク管理や、役割分けを行った。

5 月　案出し、プロジェクトリーダーを務める。どんな AI カーを作るか話し合い。
6 月　プロジェクトの方針を決める、中間発表の原稿作成、ポスター作り、YOLO の実装、データセット作り開始。
7 月　データセット完成、中間報告書の作成。
8 月　函館市内の観光物の写真撮影。
9 月　収集した画像データをアノテーションする。
10 月　アノテーションやり直し。
11 月　アンケート用の動画制作。
12 月　アンケート作成、評価実験。
1 月　期末報告書の作成。

（※文責: 井戸智斗志）

今野洸琉の担当課題は以下のとおりである。

5 月　案出し、案出しの時にでた案実際に実装できるかの検討。
6 月　中間発表の原稿作成、ポスター作り、YOLO の実装、データセット作り開始。
7 月　データセット完成、中間報告書の作成、オンライン作業できるように PC を設定。
8 月　機械学習モデルに学習させる、システム仮実装。
9 月　デザイン性を向上させる。
10 月　システム完成。
11 月　アンケート作成、アンケート結果からシステム改良。
12 月　期末報告書の作成。

小齋友里菜の担当課題は以下のとおりである。

5 月　案出し。
6 月　中間発表の原稿作成、ポスター作り、YOLO の実装、データセット作り開始。
7 月　データセット完成、中間報告書の作成。
8 月　機械学習モデルに学習させる、システム仮実装。
9 月　デザイン性を向上させる。
10 月　システム完成。
11 月　アンケート作成、アンケート結果からシステム改良。
12 月　期末報告書の作成。

稲井嵐堂の担当課題は以下のとおりである。

5 月 案出し。

6 月 中間発表の原稿作成、ポスター作り、YOLO の実装、データセット作り開始。

7 月 データセット完成、中間報告書の作成。

8 月 機械学習モデルに学習させる、システム仮実装。

9 月 アノテーション作業。

10 月 システム完成。

11 月 サンプル動画作成、アンケート作成、アンケート結果をまとめる、アンケート結果からシステム改良。

12 月 期末報告書の作成。

（※文責: 小齋友里菜）

4.2 担当課題解決過程の詳細

井戸智斗志

5 月 プロジェクトを円滑に進めるために、自己紹介を行った。さらに、このプロジェクトで何をやりたいかを話し合った。5 月末までグループリーダーを務め、その期間はプロジェクト内でのコミュニケーションでの舵をとった。そして、個々でやってみたいことを発表した後、類似する内容によって機械学習班と AI カー班の 2 つに分かれた。AI カー班ではなにをやりたいのかで意見が分かれたため、さらに 2 グループに分かれ、再度個々で、プロジェクトでやりたいことを発表した。そして、グループ C のテーマを決めた。大まかに活動予定表を作った。

6 月 5 月に決めたテーマから、Few shot が実際に動かせるかといった検討をしたり、行うシステム内容を具体的に話し合った。そして、中間発表の練習を通し、教授たちからもらったアドバイスをもとに、システムの内容を再度検討し、動画を 1 枚の画像に要約するのではなく、音楽の要素も取り入れようと話し合った。そして、観光地や、講義動画の特徴物を抽出するデータセットがないため、データセット作りも開始した。7 月 7 日の中間発表に向け、原稿づくり、ポスター作り、発表練習を行った。全体発表と、前半発表の背景、今後の予定と評価方法についての発表を担当した。

7 月 7 月 7 日中間発表。中間報告書作りを行った。

8 月 データセット作りに励んだ。データセットは、SNS（インスタ等）やインターネットから収集していく予定だった。しかし、函館という比較的マイナーな市で集められるデータは限られていたため、ネットから画像を引っ張って人工知能に学習させることは不可能だとわかった。よってデータセットを自分たちで作る必要があると考えた。実際に、函館の観光物の写真を撮影しに行った。1 つの観光物につき昼間と夜間の 2 種類、それぞれ 500〜600 枚程度の写真を撮った。撮影担当場所は、函館の駅周辺を行った。撮影箇所は函館駅、ハセガワストア駅前店、ラッキーピエロ駅前店、はこだてビール、立待岬、赤レンガ倉庫、函館西波止場、最古の電柱、八幡坂、旧北海道庁函館支庁庁舎、旧函館区公会堂、ラッキーピエロマリーナ末広店である。

9 月 データセットを完成させたが、30GB のデータセットを紛失させたため、作るシステム内容について話し合った。函館の観光を昼と夜でアノテーションのデータセットを作っていたが、昼だけに絞ることにより、データ量を半分にして実装していくことに進めた。

10 月　残っていた画像データから学習データセットを完成させた。その際、バックアップの重要性を学び、容量が大きい PC でバックアップをとろうとしたところ、プロジェクト PC がクラッシュしてしまった。そのため大学の Google アカウントにアップロードすることになった。ここで大学のアカウントのクラウド保存容量が事実上無限であることを知った。作成したデータセットを用いて、学習させたところ認識出来ていない観光物があることが判明した。そのため再度担当区域内の観光物の撮影を行った。

11 月　動画要約システムの性能を確認するために函館の紹介動画を作成した。予定では１０分を超える動画と５分程度の短い動画の 2 本を作成する予定であった。しかしメンバーへの連絡ミスと確認ミスで要求通りの動画を集めることが出来なかった。そのため２本の動画で共通の動画を入れることで、予定していた１０分の動画と６分の動画を作成した。紹介動画の担当箇所は函館駅、ハセガワストア駅前店、ラッキーピエロ駅前店、はこだてビール、赤レンガ倉庫、旧北海道庁函館支庁庁舎、旧函館区公会堂、ラッキーピエロマリーナ末広店である。またメンバーが各自作成した紹介動画を繋ぎ合わせて、BGM をつける作業を行った。

12 月　動画要約システムの性能を評価してもらう評価テストを google フォームで作成した。最終発表まで１週間しかなかったため、プロジェクトメンバーに回答をお願いするのに加えて、情報システムコースの discord サーバの管理人に許可を取り、協力を仰いだ。12 月 8 日に最終発表を行った。メンバーが発表を行いやすいように発表原稿を作成した。

1 月　期末報告書の構成を話し合い、担当箇所を決めた。できるだけ開発に関わった箇所に近い項目に割り当てられるようにした。締切までに提出したが文字数が不足しているメンバーがいたため、どの部分にどのような内容を書くのかを具体例とともにアドバイスを行った。

（※文責: 小齋友里菜）

今野洸琉

5 月　プロジェクトのテーマについて忘却学習モデルについて考えた。案出しの時にでた、忘却学習モデルについて調べたところ、既に存在していることが分かり、再度テーマを出しを行った。そして、新しく出た案「動画倍速」に Few shot が実装できるか確認できるか調べた。

6 月　5 月に決めたテーマから、システム内容を具体的に話し合った。そして、中間発表の練習を通し、教授がたのアドバイスをもとにシステムの内容を再度検討し、動画を 1 枚の画像に要約するのではなく、音楽の要素も取り入れようと話し合った。システム内に用いるライブラリについても調べた。また、完成した人工知能の評価方法を決定した。YOLO の実装を行い、メンバーに共有した。そして、観光地や、講義動画の特徴物を抽出するために適したデータセットがないため、データセット作りも開始した。データセットを作る作業に関しては、プロジェクトの時間外に集まって行った。7 月 7 日の中間発表に向け、原稿づくり、ポスター作り、発表練習を行った。システム内容と目標、現段階での状況についての発表を担当した。

7 月　7 月 7 日中間発表。中間報告書作りを行った。オンラインで夏休み作業できるように、PC を遠隔で行えるように Team Viewer をプロジェクト用の PC にインストールした。

8 月　学習データに必要な画像の収集を行った。担当箇所はトラピスチヌ修道院、函館市熱帯植物園、函館空港、ラッキーピエロ戸倉店、ハセガワストア湯の川店である。夏休み明けにメンバーがアノテーションの作業に入れるように、インストールするアプリの確認、操作方法を

習得した。

9 月 メンバーに対してアノテーションの説明を行い、システム作成に取り掛かった。しかし画像が回転してしまうというバグが発生した。またアノテーションをしたはずなのにデータが保存されないというバグも発生した。このままではデータセットの作成が予定通りに行かないため、アノテーションを行うアプリを変更した。変更に仮想環境が必要であり、口頭での説明では環境構築ができないと判断し、３人分の環境構築を行った。９月２９日にパソコンが起動しなくなるという緊急事態に落ちいた。再起動を試みたが上手くいかず今後の活動が危ぶまれた。しかし CUI での操作が可能な状態まで復旧したため、作成したコードの取り出しを行い、プロジェクトリーダーと相談することになる。

10 月 何度試みても再起動する見込みがないと判断し、学習データなどを犠牲にして初期化することにした。しかしデータを引き継ぐことができ、環境構築をし直すだけで元通りに戻すことができた。しかしすでに学習したあったデータは消えてしまったため、学習し直すこととなる。このころ動画からトリミングした画像をどのように貼るのかの議論が行ったが、納得のいくものがなかなか出来上がらなかった。

11 月 メンバーが作成したデータセットが順次完成して、PC に学習させていくことになった。物体を認識するために必要な画像枚数は当初５００～６００枚と想定していたが、３００枚程度で学習精度が上がらないケースが多々あった。学習が終了するたびにインターネット上から学習させた観光物の画像を探し、PC が認識するかを確認した。認識出来なかった観光物に関しては担当のメンバーに連絡して、画像収集からやり直すように指示を行った。

12 月 作成したシステムは順調に出来かがり、あとは発表するだけの状態となっていた。しかしプロジェクト学習の日に動かしてみると画像が順番に貼れていないという問題が発生した。すぐに修正して１２月８日の最終発表会に間に合わせることが出来た。

1 月 期末報告書ではシステム内容や技術面について担当した。自分の担当箇所を自分から確認し、１月１０日の年始の授業では全て書き終えていた。そのため２月１４日の外部発表会の準備に着手した。外部発表ではグループ C 以外に A と B についても発表することになるため、A と B に発表してほしい箇所を確認してスライド作成を行った。

（※文責: 小齋友里菜）

小齋友里菜

5 月 プロジェクトのテーマについて忘却学習モデルについて考えた。案出しの時にでた、忘却学習モデルについて調べたところ、既に存在していることが分かり、再度テーマを出しを行った。そして、新しく出た案「動画倍速」に Few shot が実装できるか確認できるか調べた。

6 月 5 月に決めたテーマから、システム内容を具体的に話し合った。そして、中間発表の練習を通し、教授たちからもらったアドバイスをもとに、システムの内容を再度検討し、動画を 1 枚の画像に要約するのではなく、音楽の要素も取り入れようと話し合った。システム内に用いるライブラリについても調べた。YOLO の実装を行い、メンバーに共有した。そして、観光地や、講義動画の特徴物を抽出するデータセットがないため、データセット作りも開始した。7 月 7 日の中間発表に向け、原稿づくり、ポスター作り、発表練習を行った。システム内容と目標、現段階での状況についての発表を担当した。

7 月 7 月 7 日中間発表。中間報告書作りを行った。

8 月　データセット用の画像収集を行った。担当箇所は五稜郭、五稜郭タワー、ラッキーピエロ五稜郭公園前店、ラッキーピエロ人見店、ラッキーピエロ本町店、ハセガワストア五稜郭店である。

9 月　不足分の画像データを収集していたところ夜ということもあり、周りの視線が辛かった。また変に絡んでくる人がいたことから画像収集を中止することとなった。この段階でグループリーダーがアノテーションデータを飛ばすというハプニングが発生し、夜の画像は収集しないこととなった。またアノテーションは仮想環境が必要で、パソコンの容量を大きく必要とした。しかしその容量は現状ないためデータ整理から始まった。

10 月　システムの性能確認のための動画収集として、立待岬、五稜郭、五稜郭タワー、ラッキーピエロ五稜郭公園前店、ハセガワストア五稜郭店を担当した。動画編集ではグループメンバーで唯一テロップをつけることにした。１箇所につき２分程度の説明があるため、テロップをつける作業は多くの時間を要した。

11 月　システム評価用の動画は個々で完成してあとは繋げるだけとなった時に問題が発生した。予定していた箇所の動画がなく、今から作成していては間に合わなくなってしまった。このままでは予定していた評価用の動画が完成しないため、どうするか話し合った。しかし全くグループメンバーと合意を得ることが出来なかった。特にリーダーと認識が合わずたびたび言い合いになった。最終的には自分が折れることで話し合いは終了し、評価用動画に関しては２つの動画に同じものも混ぜることで従来通りの尺に合わせることになった。

12 月　１２月８日に最終発表会を行った。練習では原稿を覚えられず相方に任せていた箇所も、当日までに説明できるまでに仕上げた。そのため発表当日は相方に頼ることなく発表することが出来た。

1 月　期末報告書では担当課題解決過程の詳細を主に担当した。プロジェクト学習のある日は早めに来て挨拶をして、メンバーとの交流を深めた。特にメンバー間でのすれ違いによって雰囲気が悪くなった時には仲裁に入り、最悪の状態にならないようにした。そのためメンバーの特性をよく知っている。加えて勉強会をやろうと自ら主催したりなど、グループの空気を良くする働きをした。

（※文責: 小齋友里菜）

稲井嵐堂

5 月　画像認識や、人工知能の基礎知識の習得を行った。案出しを行った。

6 月　5 月に決めたテーマから、システム内容を具体的に話し合った。そして、中間発表の練習を通し、教授たちからもらったアドバイスをもとに、システムの内容を再度検討し、動画を 1 枚の画像に要約するのではなく、音楽の要素も取り入れようと話し合った。システム内に用いるライブラリについても調べた。7 月 7 日の中間発表に向け、原稿づくり、ポスター作り、発表練習を行った。システム内容と目標、現段階での状況を担当した。発表原稿では自身の国語力を活かし、文章全体の推敲を行った。

7 月　7 月 7 日中間発表。中間報告書作りを行った。

8 月　データセット用の画像収集を行った。担当箇所はハセガワストア中道店、ラッキーピエロ港北大前店、ラッキーピエロ昭和店、ラッキーピエロ美原店である。

9 月　今野君にアノテーションの環境構築を行ってもらい順調に作業を進めた。この時期にグルー

プリーダーのデータセットが消えるというハプニングが発生し、援助することになる。具体的には画像データをもらい、アノテーションの作業を代行するものである。特に函館駅周辺は観光物が多いためアノテーションを行う画像の枚数が１万枚を超えていた。９月中には自分の担当区域のデータセットは完成して、チェックを待つことになった。

10 月　パソコンに学習させたところ目標通りの認識をすることが出来なかったため、作り直すことになる。再度画像を収集した。前回の反省として駐車場に車がたくさん停まっていたため、正しく認識出来なかったと推定して朝方の車が少ない時間帯に撮影を試みた。その後アノテーションの作業を行い、データセットを完成させチェックを行った。その結果推定は正しく認識させることが可能となった。

11 月　システム確認用の動画を撮影に行った。担当場所はハセガワストア中道店、ラッキーピエロ港北大前店、ラッキーピエロ昭和店、ラッキーピエロ美原店である。グループメンバーの中で唯一食べ物の紹介を行っている。カレーとチャイニーズチキンバーガーを注文して、どのような商品なのかをわかりやすくした。そのため外見の建物だけではなく、内装や商品もあるため華やかな動画に仕上げることが出来た。

12 月　12 月８日に最終発表会を行った。練習の時にグループメンバーが休む日があった。その時に discord で発表練習を中継することでどんな感じで行うのかを伝え、優しさを感じる一面であった。説明では指を使って説明するなど、聞き手に伝わりやすく、飽きない工夫が見られた。

1 月　期末報告書では結果と評価の箇所を担当した。自分から分からない所は質問をして、グループ全体として止まることがないように心がけていた。また報告書を作成する時は黙々と作業に集中して、指定された文字数を書き終えた。

（※文責: 小齋友里菜）

第 5 章　結果

5.1　前期の結果

前期の結果として、主に、テーマの選定、要約動画の画像部分の試作品作成を行った。

5.1.1　テーマの選定

我々のグループの目的は、人工知能を利用し人間の作業効率の向上を可能にするシステムの構築であった。メンバーそれぞれが人工知能に関する様々な文献の調査、学習を行った。その結果、人工知能の能力の１つである画像認識に着目し、既存の技術を複合し新たなシステムを構築することを目標とした。意見としてグループ内から「AI カー」と「動画を短く解釈する人工知能」の作成が挙がった。それぞれのメンバーが気に入ったテーマごとにグループを組み、プロジェクトに取り組んだ。

（※文責: 稲井嵐堂）

5.1.2　試作品作成

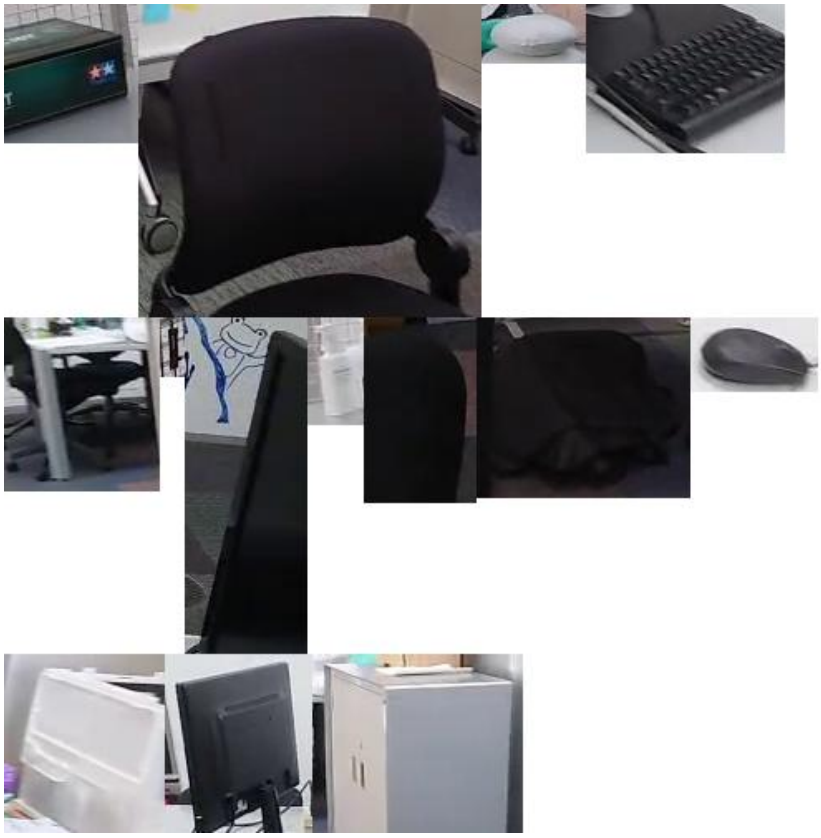

図 5.1　試作システムの出力画像

試作品の作成を行った。以下のシステムは Python により実行した。また、データセットはデ

フォルトのもののみを使用し、そのデータセットに合わせた映像を撮影した。まずは、撮影した映像を本プロジェクト用に改変した物体検出用の機械学習モデルである YOLO を適用し、フレーム分割した。それぞれのフレームごとに、rembg ライブラリにより不要な部分の背景をトリミングし、図 5.1 のように 1 枚の画像を生成するに十分な数の物体を検出した後、pillow ライブラリを用いて物体が重ならないように 1 枚の画像への貼り付けを行った。

（※文責: 稲井嵐堂）

5.2　後期の結果

図 5.2　完成したシステムの出力画像

5.2.1　システム製作

試作品を基に完成品の作成を行った。試作品と異なる点としては、試作品の時点では「1 枚の画像を生成するに十分な数の物体」とあるが、図 5.2 の通り 1 枚の画像を作成する物体の数を 4 つに決定しシステムの作成を続けた。また、rinna と呼ばれる日本語に特化した大規模言語モデルを利用し、動画内の音声の要約された内容やキーワードを文として要約後のスライドショーとともに提示するアルゴリズムの製作も並行して行った。また、後述する動画内で撮影された観光地（主に建造物）のデータセットを作成するために、数十か所の観光地の写真、総数約 20000 枚を撮影した後、それらにアノテーションと呼ばれる、画像内の各オブジェクトを枠で囲みラベルを付ける作業をほぼすべての画像に対して手作業で行った。

（※文責: 稲井嵐堂）

5.2.2　評価実験用動画の製作

後に行う評価用実験のため、評価用の 10 分程度の動画が必要であったが、本校のプロジェクト学習の規則により、グループ自ら動画を製作することとなった。評価用の動画のテーマの中から「観光」を選択し、函館に存在するめぼしい観光地を対象に動画を撮影し、機械音声とともに函館観光地紹介動画を製作した。　評価実験にて利用した 2 本の作成動画には、以下の構成を施した。

前提として、２本の動画に分けた理由は、システムによって抜き出されるオブジェクトの数が多いケースと少ないケースで、システムが被験者に及ぼす知覚的支援の大きさが変化する可能性を考慮したためである。具体的には、第一の動画では、オブジェクトの数は比較的少なめに、オブジェクト数に反比例してそれぞれの観光物に関する説明を多めになるよう、編集を行った。また、第一の動画に登場するオブジェクトは以下の通りである。

- 五稜郭タワー
- ラッキーピエロ
- 旧北海道庁
- トラピスチヌ修道院
- 金森赤レンガ倉庫

また、ここに観光物に関しての説明の例を示す。

（例）五稜郭タワー

「五稜郭公園は、北海道函館市にある公園で、国の特別史跡に指定されています。五稜郭は、江戸時代末期に築造された稜堡式の城郭で、日本初の西洋式城塞として知られています。五稜郭公園は、広大な敷地内に約 1,500 本もの桜が咲き誇る、北海道でも有数の桜の名所です。公園内には、五稜郭タワーという展望塔もあり、五稜郭の星形の眺望や函館市街や津軽海峡を一望できます。また、公園内には箱館奉行所という江戸幕府の役所が復元されており、幕末の歴史を学ぶことができます。最近は、映画「燃えよ剣（もえよけん）」、大河ドラマ「青天を衝け（せいてんをつけ）」、コミック「ゴールデンカムイ」などの舞台としても、注目が高まっています。」

第二の動画では、オブジェクト数は多く、説明が少なめになるよう編集を行った。また、第二の動画に登場するオブジェクトは以下のとおりである。

- 函館駅
- 立待岬
- ハセガワストア（各店）
- ラッキーピエロ
- はこだてビール

（※文責: 稲井嵐堂）

5.2.3　評価実験の実施

本プロジェクトは被験者に動画と、それをシステムで処理を行ったものの 2 つを比較してもらい、本システムの評価として記録した。また、それらのフィードバックを今後の開発方針の参考とするため、被験者に２つの動画の感想を記してもらった。具体的な実験手順を以下に記す。

1. 被験者に、グループで製作した 10 分程度の動画（以下、元動画と呼ぶ）を見てもらう。
2. 動画についての感想と、写されたオブジェクトを可能な限り多く書き出してもらう。
3. 次に本システムにて処理を行った動画を見てもらう。
4. 処理後の動画に対して、感想を書いてもらう。

処理後の動画についての被験者の感想を以下に示す。

- 短い時間で４枚もの写真を見せられてそれを認識するというのはもはやフラッシュ暗算と同じようなものであると感じる。そもそも内容が要約されていて欠落しているのにあんなに一度に表示されてしまうと必要である情報も認識されなくなってしまうと思う。同時に写真と文字を見せられると人間はより情報量のおおい写真の方に目が行ってしまうので文字はほとんど読めなかった。よってこのような評価となった。
- 誤解はないと思うが、動画の方が魅力的だと思う。
- 元動画は流れがありわかり易かったが要約ではその流れがなかったので誤解が生じるかもしれないと感じた。
- 概要は把握できるが、内容は把握できるかわからない。
- 情報が多く入っていたが少し速くて見にくかった、色々な場所の画像が多く出てきたので少し見えにくかった。

また、実際にシステムが選定したオブジェクトと、被験者の書き出したオブジェクトの一致率は平均して５割であった。

（※文責: 稲井嵐堂）

第 6 章　外部評価

6.1　中間発表

中間発表の評価

7 月 7 日に、プロジェクト学習の中間発表を行った。グループで発表用ポスターを作成し、口頭での説明を行った。発表内容は主に、プロジェクトの目的、システム内容、開発状況、今後の展望、により構成し、５分強の発表ののち質疑応答を行った。中間発表では、Google フォームでのアンケートを用いて、評価を記入してもらった。評価は発表技術と発表内容についての 2 項目で行う。それぞれに 10 段階での評価とその理由や感想を自由記述にて書いてもらった。結果的に 39 人の方々からフィードバックを得ることができた。アンケートに回答してもらった 9 割が学生で、1 割が教員であった。まず、発表技術についての評価を示す。39 人の平均評価点は 8.05 点であった。発表内容の平均評価点は 8.4 点であった。(どちらも小数点以下第二位で四捨五入) また、評価理由の記述を一部以下に示す。

発表技術

- グループごとにポスターを見て回ることが見やすかった。
- 初めに、全体お説明があってよかった。
- 話が分かりやすく、質問にもしっかり答えていた。
- 話の内容が、初めて聞く人でも理解できる内容で、かつ興味をひかれた。
- 後半が読んでるだけ感があって中身が入ってこなかった。
- 声が大きくて聞き取りやすかったです。聞きたいグループの発表を長く聞けるシステムが面白いと思いました。
- 視覚的に、よくまとめられていたと思います。ポスターが文字よりも画像をベースに作ってあるので視覚的に分かりやすい。

発表内容

- アイディアは面白かった。今後の活動としてはまだ検討できる所があるかも。
- 需要のある分野の中で現実可能な目標を選択していて良いと思いました。
- どの程度の性能のものが出来上がるのが予想解かないので、現時点では評価がしにくい。
- 現代の動画を倍速するだけだと、全体の概要が掴めず結局普通の速度に戻してしまうことがあるので、全体の概要を捉えつつ、手軽に見れるのは良いと思った。
- 本当はスライドショーだけで飽きないような工夫も必要なのかと思った。
- 「タイパ」という言葉が出始めた今において、時事を捉えたテーマであると思った。
- 講義動画でスライドと異なる内容を話すものがあるのでそういうものの理解がしやすいのかなと思いました。

総評としては得られた評価は 7 点以上が全体を占めており、1〜4 点などの低い評価は少なかった。発表技術については、ポスターが見やすく理解しやすいもので、ポスターについてもいい評価

を得られた。発表内容については、内容や最終成果物がどんなものになるかわかりづらいといった意見が多く見受けられた。前期活動は、どんなものを作るかといった話し合いが主に行われていたため、成果物の制作に着手できなかったことが原因として考えられる。また、質疑応答の際、ポスターは分かりやすいが、システムの内容がいまいち理解できないといった声が挙げられた。実際に動かしてみたりすることの重要さを実感した。スライドを作成しなかったため、説明不足な部分が多く、聴講者に分かりやすく説明できなかった点が、反省点に挙げられる。反省点を踏まえて後期からの活動に生かしていきたい。

（※文責: 稲井嵐堂）

6.2 成果発表

6.2.1 成果発表の評価

12 月 8 日にプロジェクト学習の最終成果発表を行った。発表用スライドを作成し、中間発表と同じく口頭での説明をおこなった。2 週間程前からプロジェクト内にて発表練習を繰り返し行い、先生方に修正をしていただいたのちに成果発表会本番に臨んだ。また、評価理由の記述を以下に記す。

発表技術

- 実験映像などがあり分かりやすくまとめられていて良かったと思った。
- 目的が理解しやすかった。
- できた成果と今後の課題が示されていて良かった。
- やりたいことと成果物が一貫していて聞きやすかった
- どういったアルゴリズムで生成しているのかの説明がわかりやすかった。
- スライドのまとめ方がよく、プロジェクトの内容の理解しやすい発表だった。
- 活動内で多くのリソースを割いていたというアノテーション付けについて、気になる点を質問させていただき、満足いく回答をいただけた。
- 非常に見やすいスライドだった。それぞれのグループについてもう少し知れると良いと思った。ページ切り替えのアニメーションを使うページと使わないページがあると重要なページの印象がより強く残るのではないかと思った。
- もう少し要約の効果を比較して見せるといいのではないかと思う。(同じ画面内で対比させるなど)

発表内容

- 実験結果の反省点と改善点をしっかりまとめられていて良かったと思った。
- 見やすいスライドだった。
- とても良かったです！
- 目標として提示されていたものがきちんと形になっており、その上で問題点などが論じられていた。一貫しており今後の発展も期待できそうだと思った。
- 今後の展望が具体的だったのでプロジェクトのゴールがはっきり伝わった。
- 現時点の成果物の問題点もしっかり分かっていて、全体的なプロジェクトの成果がよかったと思った。

- 話題となったファスト動画に関して、よく潮流を追った活動がなされていたと感じた。他の質問にも挙がったように、データセットのような既存情報をさらに活用されるとよいなと感じた。
- 動画出力と言っていたが画像のスライドショーのような動画になってしまっていたが、今後の展望で触れられていた成果物を用いるパソコンにある程度のスペックがないといけないのは大変そうだと思った。
- マシンスペックや時間などの制約の中で最大限のことをしているなと感じた。

スライドに対しては、見やすく内容が理解しやすいという評価が多く見受けられた。特に指摘されていた点としては、実験用動画と要約後の動画を実際に見てもらったことが挙げられていた。また、スライドの作成においては、難解な単語の使用を可能な限り抑え、専門外の聴講者に見てもらっても理解できるようなスライド作成を心掛けた。反省点として挙げられるものとしては、上記の２本の動画を別のページで再生したのだが、同じページで再生した方が見る側にとって対比を比較的簡単に行うことができるのではないかと考えられる。

（※文責: 稲井嵐堂）

第 7 章　今後の課題と展望

今後の課題は短く編集した画像の出力方法を変更することである。現在はトリミングをした画像を４枚まとめて出力して、音声による説明を要約した文章を中央に配置しているため、情報量が多い動画となっている。また評価アンケートからも視聴するのが大変という意見を頂いた。そのため情報量を減らすように変更する必要がある。まずトリミングした画像を出力するのは動画的良さがなくなっていますため、トリミングした箇所の前後３秒程度の動画を撮り出し、繋げていくことで動画的良さを残すことが可能となる。また４枚ではなく１枚にすることで情報量を減らすことができる。しかし音声による説明がなくなるため、切り出した数秒の動画内で音声により説明を文章化して要約したものを合成音声を使い、説明させることで音声を残すことが可能となる。そうすることで文字情報をなくすことができ、情報量をさらに減らすことができる。以上の変更を加えることで短時間で概要を理解できる動画をより分かりやすくすることが可能と考える。また動画要約という非専門職の方でも利用できるサービスとなり、画像認識をより低レベルで実装可能となると考える。

（※文責: 井戸智斗志）

第 8 章　技術面

8.1　開発環境

8.1.1　プログラミング言語

本グループは、函館の観光物を検出することに特化した物体検出モデルを開発するために、Python というプログラミング言語を用いて、実験や評価に必要なプログラムを実装した。プログラムを実装する際には、Python の標準ライブラリのほかに、さまざまな外部ライブラリを使用した。本グループが使用した外部ライブラリには、主に以下のものがある。

- ultralytics: これは、ビジョン AI と呼ばれる、画像や動画などの視覚的なデータを扱うための人工知能の分野に関するライブラリである。このライブラリには、YOLOv8 という最先端の物体検出アルゴリズムが含まれており、本グループは、このアルゴリズムを用いて、函館の観光物を検出するモデルを学習させた。YOLOv8 は、画像や動画の中に存在する物体の位置と種類を、高速かつ高精度に検出することができるアルゴリズムである。
- moviepy: これは、Python で動画編集を行うためのライブラリである。このライブラリには、動画の切り取りや結合、トリミングや回転、フィルターやエフェクトなどの機能が提供されている。本グループは、このライブラリを用いて、検出モデルの出力として生成された画像を、動画ファイルに変換する処理を行った。また、動画ファイルから音声を抽出する処理も、このライブラリを用いて行った。
- PIL: これは、Python Imaging Library の略で、Python で画像ファイルを扱うためのライブラリである。このライブラリには、画像ファイルの読み込みや保存、形式の変換、サイズの変更、色の調整、回転や反転、切り抜きや貼り付けなどの機能が提供されている。本グループは、このライブラリを用いて、検出モデルの出力として生成された画像を、一枚の画像にまとめる処理を行った。また、画像にテキストや図形を描画する処理も、このライブラリを用いて行った。
- speech_recognition: これは、Python で音声認識を行うためのライブラリである。このライブラリには、音声ファイルやマイク入力から音声を取得し、テキストに変換する機能が提供されている。本グループは、このライブラリを用いて、動画ファイルから抽出した音声を、テキストに変換する処理を行った。このテキストは、検出モデルの出力として生成された画像に、音声の内容を表すテキストとして描画した。

また、rinna と呼ばれる言語モデルを使用した。rinna とは、rinna 株式会社が開発した日本語に特化した 36 億パラメータを持つ GPT 言語モデルである。GPT とは、Generative Pre-trained Transformer の略で、大量のテキストを自己教師あり学習により学習した汎用言語モデルである。rinna は、日本語の Wikipedia やインターネット上のテキストなどを学習データとして利用し、日本語のテキスト生成において高い性能を発揮する言語モデルである [6]。本グループは、rinna と speech_recognition を組み合わせて、音声から重要なキーワードを抽出するプログラムを作成した。まず、speech_recognition で音声をテキストに変換する。次に、rinna でテキストを解析し、重要なキーワードを検出する。このようにして、音声の内容を理解し、関連するキーワードを抽出

することができる。

（※文責: 今野光琉）

8.1.2 実行環境

本グループで実装・実験を行うために使用した PC の仕様は次のとおりである。

- CPU: AMD Ryzen 7 3700X 8-Core Processor
- GPU: NVIDIA RTX 4070 ti
- RAM: DDR4 64GB

OS には、Linux ベースのオープンソースのオペレーティングシステムである Ubuntu 20.04 LTS を用いた。ubuntu は、多くの研究者や開発者に利用されており、様々なプログラミング言語やツールがサポートされている OS でああり、本プロジェクトが利用したプログラムの、安定した動作と高い互換性を持った OS であった。

この CPU は 8 コア 16 スレッドで高速な処理が可能である。また、GPU は最新のグラフィックスカードで高性能な演算が可能であり、この GPU は CUDA コアや tensor コアなどのハードウェア備えているため、並列計算や機械学習などの処理において高い効率と性能を発揮する。そして、RAM は大容量でかつ DDR4 を採用したため、多くのデータを扱え、データの読み書き速度も速い。これらの仕様は、本グループが用いた、言語モデルのような自然言語処理のタスクを行う際や、物体検出アルゴリズムのようなコンピュータービジョンのタスクにおいて、使用に適していると考えたため、この仕様の PC を用いている。

（※文責: 今野光琉）

8.2 実験方法の詳細

8.2.1 物体検出モデル

私たちは、YOLOv8 という最新の物体検出アルゴリズムを用いて、函館の観光物を検出することに特化した物体検出モデルを作成した。このモデルは、函館の観光物として定義される、函館の歴史や文化を反映したさまざまな観光スポットや建造物を、高い精度と速度で検出することができる。

函館の観光物とは、函館市内や周辺にある、函館の魅力を象徴するような観光スポットや建造物のことである。例えば、函館山は、函館の夜景を一望できる絶景スポットである。五稜郭は、幕末の戦跡を残す歴史的な要塞である。金森赤レンガ倉庫は、明治時代の貿易の拠点として建てられたレトロな建物である。これらの観光物は、函館のシンボルとして、多くの観光客や写真家に人気がある。

私たちは、これらの観光物の画像を、インターネットから収集するのではなく、自分たちで実際に現地に行って撮影することにした。これは、インターネット上の画像は、一般的な角度や条件で撮影されたものが多く、モデルの汎化性能を向上させるには不十分だと考えたからである。私たちは、それぞれの観光物について、時間帯や距離、角度など、さまざまな条件で撮影を行い、約 600 枚ずつの画像を得た。全体として、約 19200 枚の画像を収集した。

収集した画像に対して、私たちは、YOLOv8 の入力として必要なアノテーションを行った。アノテーションとは、画像内に存在する観光物の位置と種類を、バウンディングボックスとラベルという形式で表すことである。私たちは、画像ごとに、観光物のバウンディングボックスの座標とサイズを手動で指定し、ラベルとして観光物の名前を付けた。この作業は、非常に時間と労力がかかるものであったが、モデルの性能を向上させるためには必要なものであった。

アノテーションを行った画像データを、訓練データとテストデータに分割し、YOLOv8 に学習させた。訓練データとは、モデルが学習するために使用するデータであり、テストデータとは、モデルの性能を評価するために使用するデータである。私たちは、それぞれの観光物ごとに、訓練データを約 7 割、テストデータを約 3 割の割合で分割した。これは、訓練データとテストデータの分布が、実際の画像の分布に近くなるようにするためである。私たちは、YOLOv8 のハイパーパラメータを調整し、訓練データを用いてモデルを学習させた。

8.2.2　アーキテクチャの詳細

本グループは、観光用の動画を入力として、その動画の内容を要約した動画を出力するシステムを開発した。このシステムは、物体検出や音声認識、テキスト解析などの技術を組み合わせて、動画の中に存在する特徴的な物体や音声の内容を抽出し、それらを一枚の画像にまとめて、要約された動画を出力するアーキテクチャとなっている。概要図は図 8.1 の通りである。

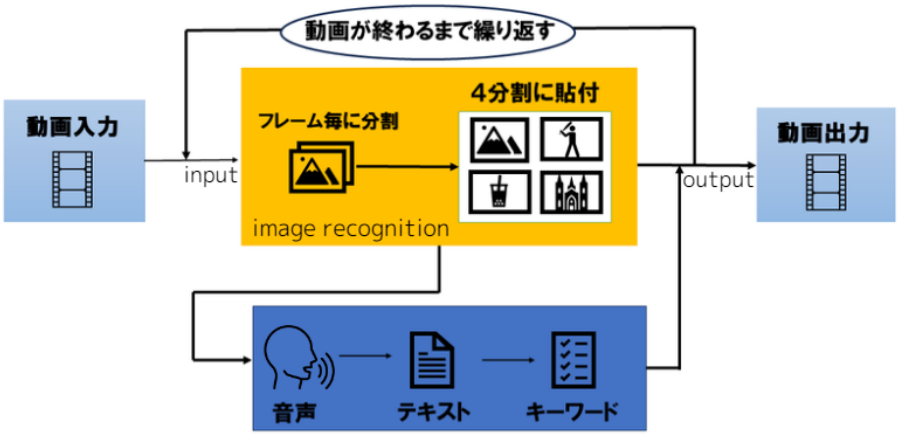

図 8.1　おおまかなシステム概要

このシステムの具体的なアーキテクチャは、次のようになっている。

まず、システムに観光用の動画を入力する。観光用の動画とは、函館の観光スポットや建造物などを撮影した動画のことである。この動画は、本グループが自分たちで実際に現地に行って撮影し、編集したものである。この動画は、システムによってフレームに分割され、各フレームに対して以下の処理が行われる。

動画の入力後、本プロジェクト用に改変した物体検出用の機械学習モデルである YOLOv8 を適用して、入力された動画の特徴的な物体をフレーム毎に検出する。物体検出とは、画像や動画の中に存在する物体の位置と種類を検出することである。前述したように、本グループは、YOLOv8 を函館の観光物に特化させるために、函館の観光物の画像を大量に収集し、アノテーションを行い、学習させた。また、学習には、CPU のみだと膨大な時間がかかるため、このような機械学習モデルの学習に適した NVIDIA RTX 4070 ti の CUDA コアの利用し、高速で学習を行った。また、

検出した物体が何フレーム目にあるか、どのディレクトリに検出した物体が格納されるかなど、検出時の情報を取り出すような処理を YOLOv8 のプログラムに書き加えた。このようにして、本プロジェクト用に改変した YOLOv8 を作成した。このモデルを用いて、各フレームに対して物体検出を行う。この際に、同じ名前の物体を複数回検出した場合は、それが正しい物体である確率が最も高い物体のみを用いるようにした。これは、物体検出の誤りを減らすためである。

その後、4 種類の物体を検出したら、PIL と呼ばれるライブラリを用いて、それぞれの物体が重複しないように、均等な大きさに変換して 1 枚の画像にまとめる。具体的には、それぞれを均等な大きさに変換後、検出した順に左上、右上、左下、右下にそれぞれの物体を何もない白い画像に張り付ける。この画像は、動画の要約として使用される。

その後、moviepy と呼ばれるライブラリを用いて、1 つ目の物体を検出したときのフレームから、4 つ目の物体を検出した地点までのフレームの範囲の動画を音声ファイルに変換する。

そのテキスト化のライブラリである speech_recognition を用いて音声ファイルをテキストに変換する。この際、speech_recognition の動作が不安定であり、音声ファイルをテキストに変換できないという状況がたまに起きるという問題があった。そのため、エラーが起きて変換に失敗した場合は、変換が成功するまでこの処理を繰り返す必要があり、このために処理時間がかかってしまった。

その後、rinna と呼ばれる言語モデルを用いて、変換したテキストから重要なキーワードを抽出する。この際、「次の文章の中から、キーワードだけを 3 つ抽出してください。」というプロンプトを与えて、言語モデルから動画の内容を要約するキーワードである出力を得た。この際、キーワードではなく文章を出力することがたまにあるという問題があった。そのため、単語が 3 つ「・」や「、」で区切られたものが出力されるまでこの処理を繰り返す必要があり、このために処理時間がかかってしまった。

そして、moviepy を用いて、4 つの物体をまとめた画像の中央に抽出したキーワードを張り付け、この画像を 5 秒程度の動画に変換する。

これらの処理をフレーム全てに対して検出が終わるまで繰り返し、作成した順序に従い動画を組み合わせ出力する。最終的に出力された動画が、観光用の動画の内容を簡潔に伝えることができる要約動画となる。

このようなアーキテクチャによって、要約動画を出力するシステムを作成した。

（※文責: 今野光琉）

参考文献

[1] 損害保険ジャパン株式会社. ”損保ジャパン『Z 世代映像研究課』設立！【若者の動画視聴実態】を調査 Z 世代の“快適”な視聴速度は 1.5 倍速、他世代と比べて約 1.2 倍のセリフ量をストレスなく理解していることが判明”. PR TIMES. 2022. https://prtimes.jp/main/html/rd/p/000000188.000078307.html, (参照 2023-07-14).

[2] 近畿大学. ”16,227 人のオンデマンド授業視聴データを徹底分析　視聴速度を適宜切り替えて受講するのが大学生のトレンド”. NEWS RELEASE. 2023. https://www.kindai.ac.jp/news-pr/news-release/images/newsasset-006-68323f.jpg, (参照 2023-01-14).

[3] NEC. ”短縮動画＋説明文で要約　映像認識 AI × LLM”. NEC の最先端技術. 2023. https://jpn.nec.com/rd/technologies/202314/index.html, (参照 2023-01-14).

[4] 公立はこだて未来大学 HP. ”プロジェクト学習”. 公立はこだて未来大学. 2022. https://www.fun.ac.jp/project-learning, (参照 2023-07-14).

[5] Peiyuan Jiang, Daji Ergu, Fangyao Liu, Ying Cai, Bo Ma. (2022) “A Review of Yolo Algorithm Developments” Procedia Computer Science, 199:1066-1073.

[6] rinna. ”rinna、日本語に特化した 36 億パラメータの GPT 言語モデルを公開”. rinna 株式会社. 2023. https://rinna.co.jp/news/2023/05/20230507.html, (参照 2024-01-04).