**[CSE3301-04] Video Summarization Project:** **Data Analysis Report**

20200118 조창희

목차

1. 딥러닝 공부 상황 정리3

모두를 위한 딥러닝 강의 수강 3

코세라 강의 수강3

강의 수강 소감5

1. Video Summarization Project 6

데이터 셋 분석6

1. Video title, category 분석
2. Video data set 내부 세부 분석

적절하지 않은 클래스가 나온 이유10

분석결과 및 한계 지점10

분석결과 및 개선방안 제시11

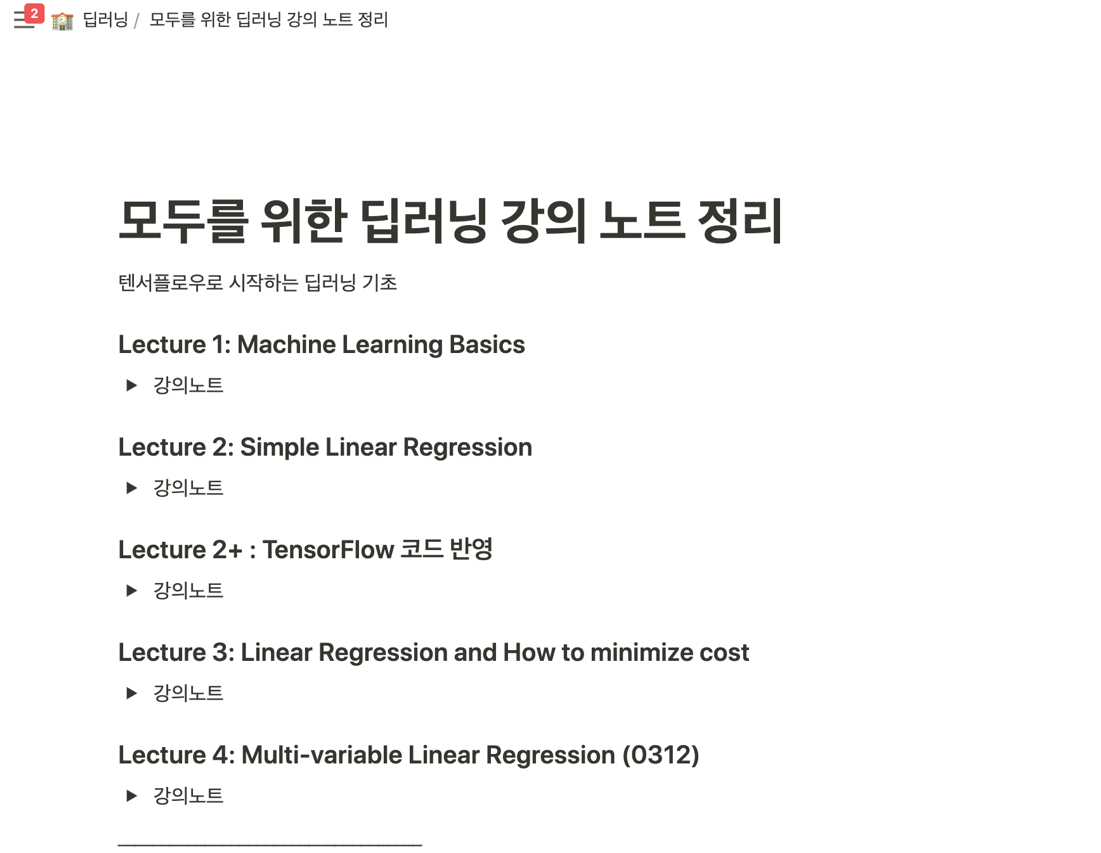
새롭게 제시한 라벨 예시11

소감11

**딥러닝 공부 상황 정리**

**모두를 위한 딥러닝 강의 수강 (23.03.05 ~ 23.03.16)**

Coursera 강의를 수강하기 전, 한국어로 되어 있는 모두를 위한 딥러닝 강의 1,2를 선별적으로 수강하여 개념을 익혔다.



[Notion 툴](https://www.notion.so/7a15ab2685e64d54b1f33197ca4559ae?pvs=4)을 통해 강의 내용을 정리하였다.

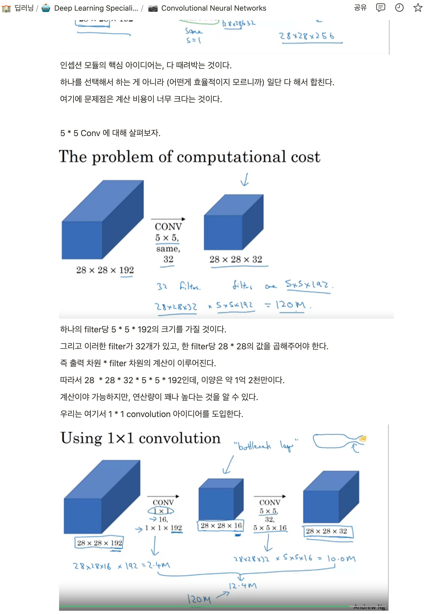
**Coursera 강의 수강 (23.3.16 ~ 23.5.30)**

연구프로젝트 초 중반 기간에는 courser에서 제공하는 Andrew ng의 Deep Learning Specialization 강의를 수강하며 공부하는 기간을 가졌다. 강의는 총 5개의 코스로 구성되어 있으며, 한 강의당 3-4주 분량의 수업이 있다. 본인은 3번째 코스를 제외하고 다음 4가지의 코스를 수강하였다.

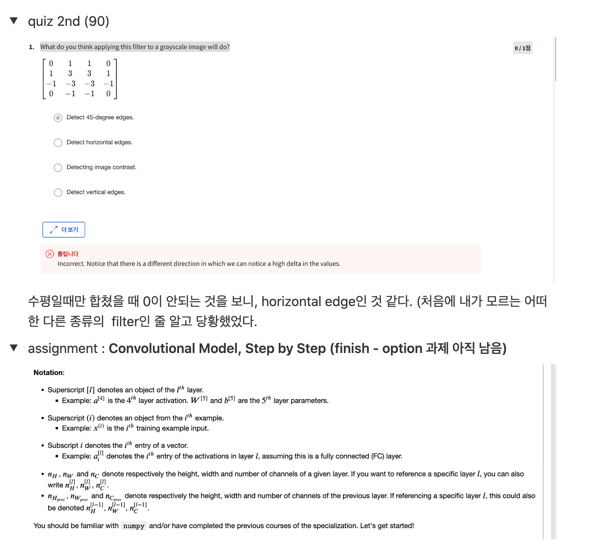
1. **Neural Networks and Deep Learning**
2. **Improving Deep Neural Networks: Hyperparameter Tuning, Regularization and Optimization**
3. **Convolutional Neural Networks**
4. **Sequence Models**

딥러닝에 대한 개념이 많이 부족했기 때문에, 처음 course 1을 완주하는 데에 시간이 오래 걸렸었다. 단순히 강의 수강에서 그치지 않고, 확실하게 개념을 이해하고 익히기 위해 과제와 퀴즈까지 진행했었다.

텍스트, 스크린샷, 소프트웨어, 웹사이트이(가) 표시된 사진

자동 생성된 설명

<강의 정리 예시>



<퀴즈 오답정리 예시>

강의, 퀴즈, 과제 등의 공부 내용은 위와 같이 [notion 툴](https://www.notion.so/Deep-Learning-Specialization-cd39ad82eaf843168f2aa0314a282292?pvs=4)을 사용하여 정리하였다.

**수료증**

텍스트, 스크린샷, 원, 폰트이(가) 표시된 사진

자동 생성된 설명텍스트, 스크린샷, 폰트, 원이(가) 표시된 사진

자동 생성된 설명

\* Convolutional Neural Networks, Sequence Models는 부분적으로 수강하여 아직 수료증이 나오지 않았지만, 거의 다 들었기 때문에 추후 완강할 계획이다.

**강의를 수강한 소감**

원래 딥러닝에 흥미는 있었으나, 막대한 수식에 대한 두려움이 있었던 상태였다. 그러나 Andrew ng의 강의는 기본적인 Intuition부터 차근차근 설명해주었기 때문에 수식과 모델의 목적을 생각하며 공부할 수 있었다. 이번 연구 프로젝트 수업을 통해 coursera 강의를 듣게 되면서 딥러닝의 기본적인 개념을 잡을 수 있었다. 또한 강의만 있는 것이 아니라 매주 과제, 퀴즈도 있어서 다른 3학점짜리 수업을 듣는 느낌이었다. 과제가 많이 친절한 편이긴 했지만, 직접 코드를 짜면서 개념적으로만 이해했던 것을 코드로 적용할 수 있었다. 아쉬운 점이 있다면, 아무래도 강의가 열린 시점이 5-6년 전이다보니, 아주 최신 기술까지는 익힐 수 없었다는 점인데, 기본적인 개념을 익히기에는 충분히 훌륭한 강의라고 생각한다.

강의를 들으며 느꼈던 딥러닝이라는 학문의 독특한 점을 다음 세 가지로 정리할 수 있겠다. 첫째로, 딥러닝 관련 task 로 흔히 얘기하는 컴퓨터 비전(CV), 자연어처리(NLP), 추천시스템 등은 독립적인 분야인 것 같지만 서로 연계되어 있는 점이 많다는 것이다. 비교적 최근에 발전하기 시작한 자연어 처리의 경우에도 논문의 상당 수는 컴퓨터비전에 있었던 기존 아이디어를 적용한 경우가 많았다. 음성 인식의 경우에도 RNN으로 풀 수 있지만, 스펙트로그램화 해서 CNN으로 접근할 수도 있다. 두 번째로, 컴퓨터 공학이라는 학문은 어느정도 역사가 있는 편이라 명확한 교재와 강의가 많이 존재한다. 하지만 인공지능의 경우 급진적으로 발전하기 시작한 시점이 비교적 최근이라, 정립된 교재가 많지 않으며 파장을 일으킨 논문 자체가 교재가 된다는 것이다. 따라서 해당 수업에서도 발표 논문을 토대로 설명하는 경우가 많았다. 마지막으로, 높은 성능을 보였던 모델을 살펴보면 복잡한 수식을 바탕으로 최적화하는 경우도 많지만, 성능이 높은 모델은 단순한 직관으로 시작된 경우도 많이 있다. 그리고 이러한 직관은 보통 다른 분야에서 사용한 개념을 바탕으로 생겨나는 경우가 많다. 따라서 딥러닝 개발자는 딥러닝이라는 한 분야에 매몰되기 보다는, 다양한 분야에 대한 지식이 있는 상태에서 적용할 수 있다면 재미있는 결과를 많이 만들 수 있을 것 같다.

**Video summarization Project (23.6.1 ~ 23.6.27)**

**Base model 소개**

Base mode은 현재 video summarization에서 state-of-art인 [Video MAE v2](https://github.com/OpenGVLab/VideoMAEv2)이다. 해당 모델은 self-supervised learning(자기지도학습)을 하여 Masked Auto encode (MAE)를 비디오에 적용했다고 한다.

이와 관련한 기반 논문은 대표적으로 다음과 같다.

[Masked Autoencoders Are Scalable Vision Learners](https://arxiv.org/pdf/2111.06377.pdf)

[VideoMAE: Masked Autoencoders are Data-Efficient Learners for Self-Supervised Video Pre-Training](https://arxiv.org/abs/2203.12602)

**데이터 셋 분석**

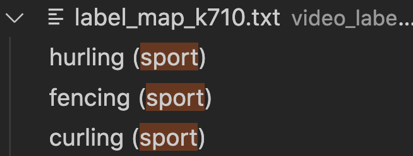
해당 연구프로젝트에서는 모델에 앞서 기존 데이터 셋에 대한 분석을 하였다. 주어진 비디오 영상들 중에서 비디오 제목, 카테고리에 관해 관련 라벨이 존재하는지 분석하고, 각 영상당 2초씩 잘라 라벨링이 적절히 이루어져 있는지 분석하는 과제였다.

**1) Video title, category 분석**

|  |  |  |  |
| --- | --- | --- | --- |
| 실제 비디오 이름 | 실제 카테고리 | 이름과 연관된 class | 카테고리와 연관된 class |
| Mexican Fried Chicken Sandwich Recipe | Making Sandwich | cooking chicken, making sandwich | making sandwich |
| Reuben Sandwich with Corned Beef & Sauerkraut | making sandwich |
| Poor Man's Meals: Spicy Sausage Sandwich | making sandwich, cooking sausages (not on barbeque) |
| Saigon Sandwich - Vietnamese Sandwiches in San Francisco CA | making sandwich |
| GoogaMooga Sneak Peek Joseph Leonard's Fried Chicken Sandwich cooking video | making sandwich, cooking chicken |
| David Belle | Fondateur du parkour | Reportage de TF1 | \*ParKour | X | X  (sport 관련은 있으나 ex. hurling (sport) Parkour을 지칭하는 것은 없다. 기술에 대한 클래스도 없다.) |
| Charlotte Parkour | Charlotte Video Project | X |
| Singapore Parkour Free Running | JC Boy Late for School | running on treadmill |
| Parkour Camp Leipzig | X |
| Vivencias: Jam Parkour Vi횄짹a del Mar 2012 | X |
| Chinese New Year Parade 2012 New York City Chinatown | PaRade | Celebrating | Celebrating  (완전히 정확한 매칭은 X) |
| 2013 Alumnae Parade & Laurel Chain Ceremony | Celebrating |
| LA Kings Stanley Cup South Bay Parade 2014 | Celebrating |
| Chinatown Parade | Celebrating |
| Southend-on-Sea and Kortrijk Bike Friendly Cities Parades | riding a bike, Celebrating |
| Oliver's Show - Dog's tale | Dog Show | training dog | training dog |
| TODAY- Obie the obese dog works toward weight loss | training dog |

\*Parkour : 야마카시의 멤버 다비드 벨에 의해 창시된 이동 기술

제목, 영상 초반부 소개를 기반으로 적절한 클래스를 찾아보았다. ParKour의 경우, 개념 자체가 독특하다보니 연관된 라벨을 찾기 어려웠다. Sport를 예시로 들었을 때 다음과 같은 클래스가 있었는데, 전체적인 스포크를 포괄하는 것이 아닌, 지엽적인 데이터 셋에 맞춰져있는 라벨로 보였다.



**2) Video data set 내부 세부 분석**

분석한 결과물은 양이 너무 많아 따로 엑셀 파일로 첨부할 예정이다. 전반적인 분석 방식과 분석 과정에서 발견하였던 지점에 대해 레포트에 적도록 하겠다.

**분석 방식**

2초마다 라벨이 제대로 이루어져 있는지 확인하는 작업을 거쳤다. 처음에는 2초별로 스크린 샷과 함께 상세한 내용을 정리하였었다.



<정리한 노션 참고>

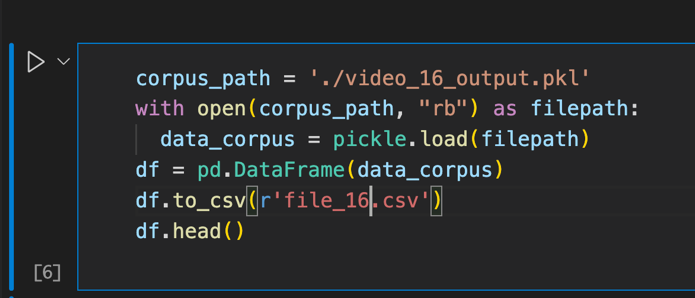
그러나 시간이 과도하게 오래걸려 다음과 같은 분석 기준을 정의하고, 엑셀에서 일괄적으로 정리하였다.

확인하는 기준은 다음과 같이 나누었다.

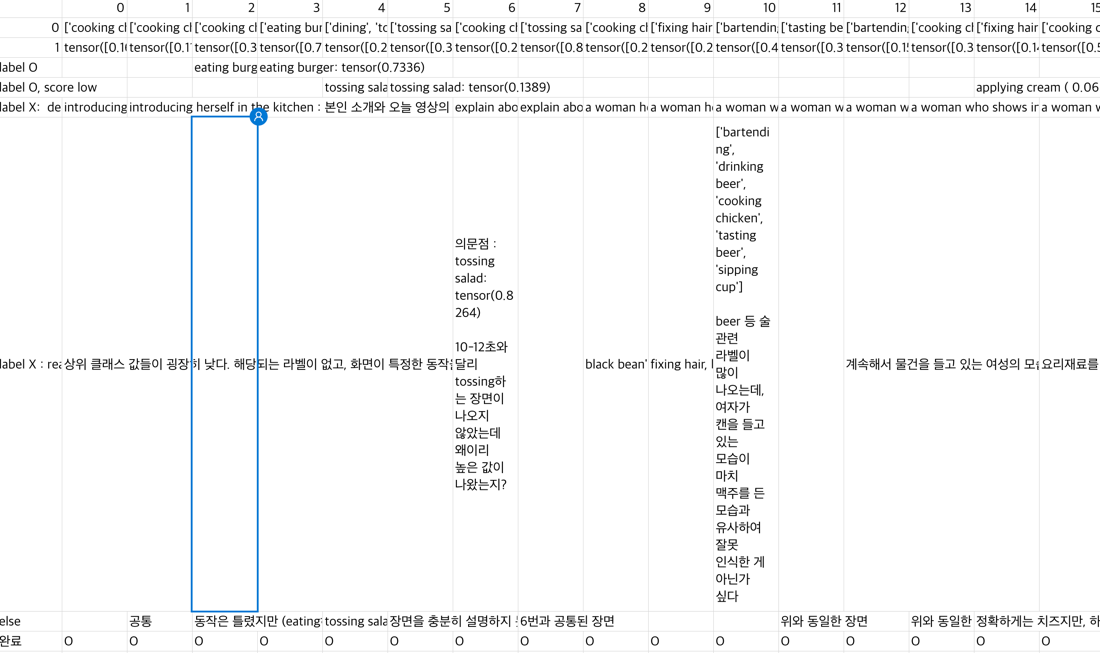
1. 모델이 출력한 상위 5개의 클래스에 적절한 클래스가 있고, 해당 클래스의 수치가 0.5(올림 기준)를 넘어서는 경우
2. 모델이 출력한 상위 5개의 클래스에 적절한 클래스가 있으나, 해당 클래스의 수치가 0.5 미만인 경우
3. 모델이 출력한 상위 5개의 클래스에 적절한 클래스가 없는 경우, 새로운 클래스 제시 (기존 다른 클래스에 이미 있는 경우, 비고에 기존 클래스임을 명시)
4. 3번에 해당하는 경우, 적절한 클래스를 출력하지 않은 이유를 적기
5. 비고

분석한 방식은 다음과 같다.

1. 먼저 Video mae v2를 이용하여 영상 별로 pkl 파일을 생성한다.



1. Pkl 파일을 열어서 csv 파일로 전환한다.

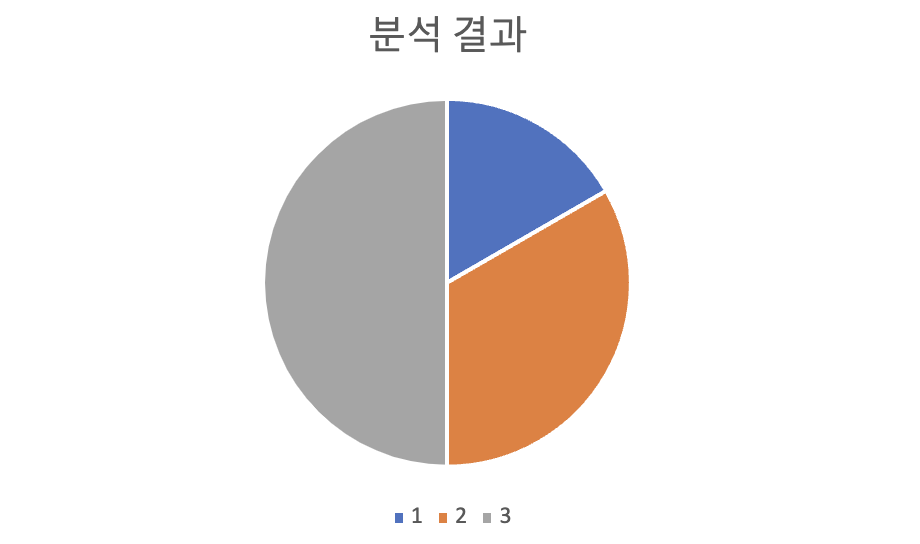


1. 엑셀에서 각 2초 구간마다 위에서 정의한 기준에 맞게 클래스를 구분한다.

Video-16 기준, 총 198개의 결과 클래스가 존재했다.

**분석 내용**

**비율**



분석 결과, 총 188개의 클래스에서 적절한 클래스이며, 그 비율도 높은 것은 44개(1번 기준), 적절한 클래스이지만 비율이 낮은 것은 58개(2번), 적절한 클래스가 아니어서 새롭게 클래스를 지정해주어야 했던 것은 97개(3번)로 가장 높았다. 전반적으로 옳은 경우가 약 23%로 낮은 성능을 보인다.

**적절하지 않은 클래스가 나온 이유**

첨부한 엑셀 파일에 각 상황의 이유를 적어 놓았다. 대부분의 경우, 영상에 나오는 동작에 해당되는 라벨이 없는 경우에 발생하였다. 또한 적절한 라벨이 없는 경우 나온 결과값들이 모두 낮은 비율을 보였다. 또한 동작이라고 판명되기 어려운 상황도 많았는데, 그러한 경우에는 cooking chicken으로 라벨링하는 편향성이 보였다.

**분석결과 및 한계 지점**

1. **라벨링 프레임의 한계** : 사람의 눈으로 판단할 때 2초가 정확히 어느 프레임까지인지 맞추기 어렵다. 만약 경계선에 있는 경우 값이 달라질 수 있다.
2. 적절한 라벨이 이미 존재하며 프레임 안에 확실히 동작이 정의되는 데에도 불구하고 인식되지 않는 클래스가 있다. ex. 32, 33번에서 tasting food
3. 동작이라고 정의 내리기 애매한 화면이 나오는 경우 무조건 cooking chicken 이 상위 클래스로 출력되는 편향이 있다.
4. 라벨링 자체가 주관적인 영역이기 때문에, 어디까지가 옳은 라벨이고 어디까지가 적절하지 않은 라벨인지 판단하는 기준이 애매하다.
5. 특별한 동작 없이 설명하는 장면들이 있다. 이러한 경우와 같이 동작을 라벨링하는 것 자체가 애매한 상황이 있다.
6. **중복된 클래스 존재**: tasting food와 eating burger은 중복된 클래스이다. tasting food가 더 상위개념이다.
7. 라벨은 (동작 행위) + (동작 대상) 의 형태로 주로 구성되어 있다. 이 둘 중 하나만 맞는 경우에는 어떻게 판단해야 할지 애매한 측면이 있었다.
8. **동작 인식에 있어 한계** : 만약 사람이 샌드위치를 만들고 있고, 그 옆에 커다란 치킨이 있는 경우, 모델은 치킨 관련 동작을 가장 높은 점수로 예측한다.
9. **같은 동작도 크기에 따라 다른 결과** : 줌아웃이 되어 이미지가 적게 차지될 때는 적은 값이, 줌인되어 이미지가 클 때는 큰 값이 출력되는 경향이 있다.

위는 분석에서 느꼈던 한계 지점들이다.

**분석 결과 및 개선방안 제시**

1. **필요하지 않은 장면 무시** : 2초 내에서 사람의 설명 장면, 장면 전환, 다음 영상 예고 등 비디오 요약에 있어 필요하지 않은 구간이 많이 발생하였다. 해당 장면을 적절히 무시할 수 있다면 좋을 것 같다. 본인은 데이터를 분석하면서 이러한 장면을 **상위 클래스의 점수가 높지 않은 장면**으로 정의하였다.
2. **중복 장면 처리** : 같은 장면이 10초 이상 지날 경우 복된 정보가 발생한다. 따라서 중복된 장면으로 판단되면 (이전 장면과 어느정도 이상 비슷한 수치를 보이면) 무시하는 과정이 필요할 것 같다.
3. **정적인 이미지와 동적인 이미지의 분류** : 8번 한계 사항을 해결하기 위해 지속적으로 가만히 있는 이미지와 계속해서 움직이는 이미지를 구분할 수 있다면, 더 정확한 동작 인식을 할 수 있을 것이다.
4. **멀티모달** : 비디오가 어떠한 방식으로 전개될 것인지에 대해 초반 10초 정도에 대해 압축적으로 설명하고 있다. 모든 종류의 비디오가 그렇지는 않지만, 16-20와 같은 요리 관련 영상의 경우 이러한 구조를 따르고 있다. 따라서 두괄식 형태의 비디오라면 앞 정보의 음성 데이터를 수집하여 압축하는 데에 도움이 될만한 정보를 얻을 수 있을 것이다.
5. **데이터셋 재구조화** : 사용한 라벨 목록인 label\_map\_k710.txt의 경우, 포괄적인 행동에 대한 라벨은 거의 존재하지 않고, 매우 지엽적인 라벨들로 구성되어 있다. 따라서 아예 존재하지 않는 라벨에 대한 동작이 있는 경우에는 엉뚱한 결과를 갖게 된다. 이러한 상황을 방지하기 위해 라벨의 형태를 단순 동작으로 크게 우선 정의하고, 동작을 하고 있는 대상으로 세분화하는 식으로 정의한다면 더 성능을 높일 수 있을 것이라고 생각한다. 가장 좋은 방법은 데이터 셋을 늘리는 것인데, 적은 라벨링으로 높은 성능을 내려면 재구조화하는 과정이 필요할 것 같다.
6. 설명 혹은 서 있는 상황에 대한 데이터 셋 정의가 필요하다. 이러한 상황일 때 엉뚱한 라벨이 들어가게 되는데, 전반적인 요약에 있어 성능 저해가 있을 것 같다.

**새롭게 제시한 라벨 예시**

다음은 적절한 라벨이 존재하지 않는 경우, 임의로 해당 동작에 맞게 새롭게 만든 라벨들 몇가지를 제시한다. 더욱 많은 라벨은 엑셀 파일에 첨부되어 있다.

|  |  |  |  |
| --- | --- | --- | --- |
| introducing herself in the kitchen | explain about cooking | holding a can of black beans | Showing ingredients for cooking |
| cutting tomatoes | put food on the table | making cheese | putting food on the table |
| cutting avocado | holding a bowl | explain something | showing chicken |

**소감**

한학기동안 연구프로젝트를 진행하면서 딥러닝에 대한 기본적인 내용을 익힐 수 있어 유익한 시간이었다. 우선, 코세라 강의 학습을 통해 이전보다 더 확실히 개념 정립이 된 것 같아 좋았다. 그리고 데이터 셋 분석도 의미 있는 시간이었다. 과제나 프로젝트 등을 통해 다양한 하이퍼 파라미터로 모델 실험을 하면서, 모델의 성능을 좌우하는 것은 모델의 구조 자체보다 데이터 셋이 중요하다는 생각을 많이 했었다. 하지만 정작 데이터 셋을 깊이 있게 분석할 기회는 거의 없었는데, 이번 기회에 세부적으로 분석할 수 있어 의미 있는 시간이었다. 또한 코드에 들어가기 전 모델의 구현 결과를 통한 분석만 해보더라도 이후 어떤 방향으로 개선되어야 할지 많은 아이디어를 얻을 수 있었다. 다만 시간 관계 상 많은 영상 데이터에 대한 분석을 하지 못했고, 분석 결과로 제시된 지점들을 직접 시도하고 재학습 시키지 못한 점은 아쉬움이 남는다.