**김형원\_스터디 및 계획**

**1.공부방법**

1) 유튜브에서 모두의 딥러닝 기초 강의를 수강

<https://hunkim.github.io/ml/>



2) 위 강의를 바탕으로 개인 블로그에 강의 노트들 작성 및 복습

<https://blog.naver.com/rlagudndjs95>

스크린샷이(가) 표시된 사진

자동 생성된 설명

스크린샷이(가) 표시된 사진

자동 생성된 설명

3) 현재 모두의 딥러닝 교재 정독중

-> 텐서플로와 Keras사용법 위주로 공부.

-> 헷갈리는 개념들이 많아서 지속적인 복습 요구.

**2. Tensorflow를 이용한 신경망 실습**

1) CNN(일부) -> 기존에 있던 텐서플로 소스들을 바탕으로 이해에 중점

스크린샷, 텍스트이(가) 표시된 사진

자동 생성된 설명

2) RNN(일부) -> 기존에 있던 텐서플로 소스들을 바탕으로 이해에 중점

스크린샷이(가) 표시된 사진

자동 생성된 설명

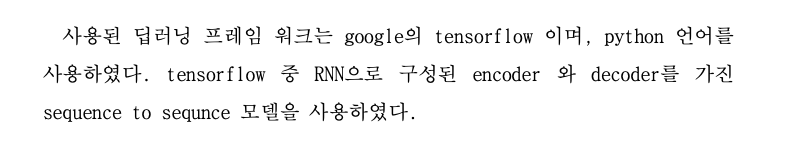
**3. 찾아본 선행 연구**

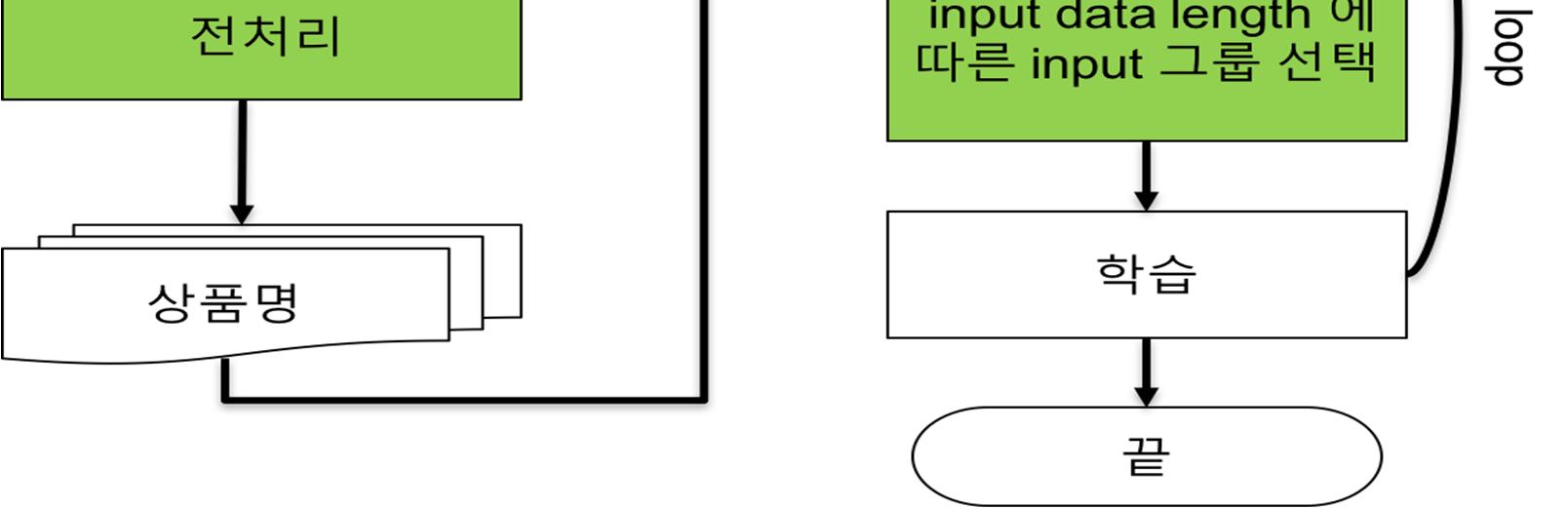
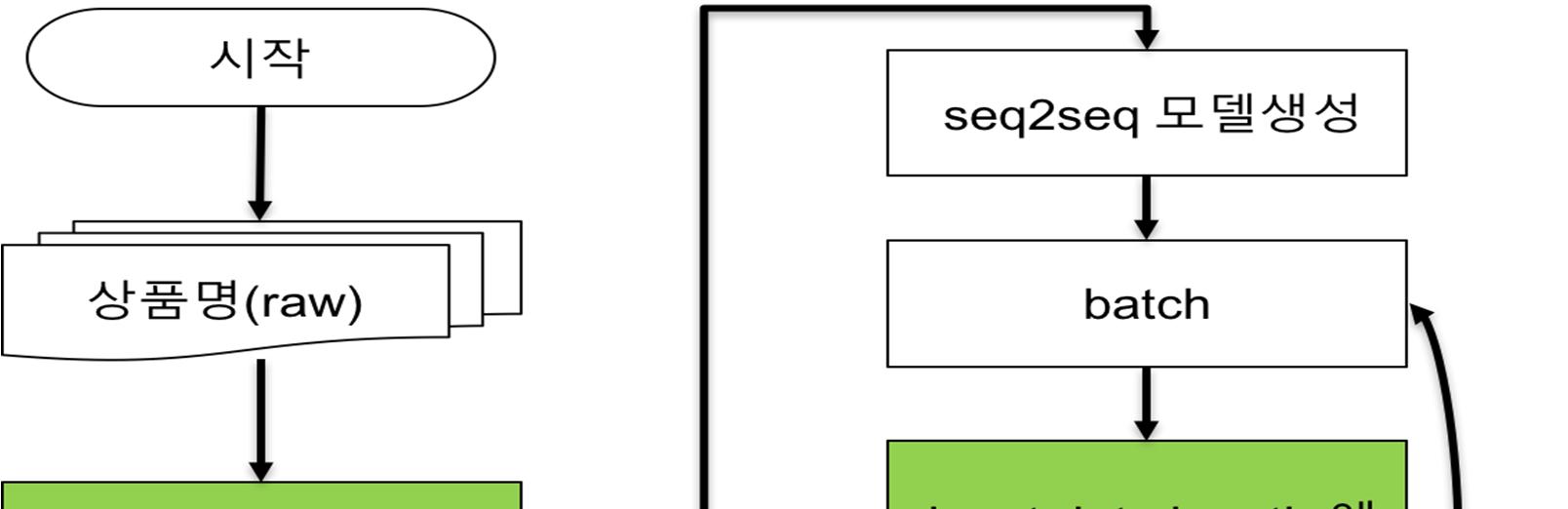
1)찾아본 논문에 나온 기술 요약

**<출처>**

**딥러닝을 이용한 형태소 분석 기반의 상품 카테고리 분류 기법**

**Product Category Classification based on Morphological Analysis using Deep Learning**

**(김진삼, 조인휘)**



1. **전처리 단계**

상품명과 브랜드명이 가장 필요한 정보, 작지만 영향을 끼치는 정보에는 용량, 수량

전혀 상관이 없는 정보는 기호(대괄호, 소괄호 등), 프로모션 여부, 배송 정보등이 있다.

전처리 단계에서 구두점과 함께 분류와 상관없는 문자들을 제거 하였다.

또한 일반적 전처리 외에 모든 텍스트에 적용 가능한 규칙이 필요 하였으므로 트위터 형태소 분석기를 사용하여 정규화, 토큰화, 어근화, 어구 추출의 단계를 거쳤다.

1. **모델 생성**

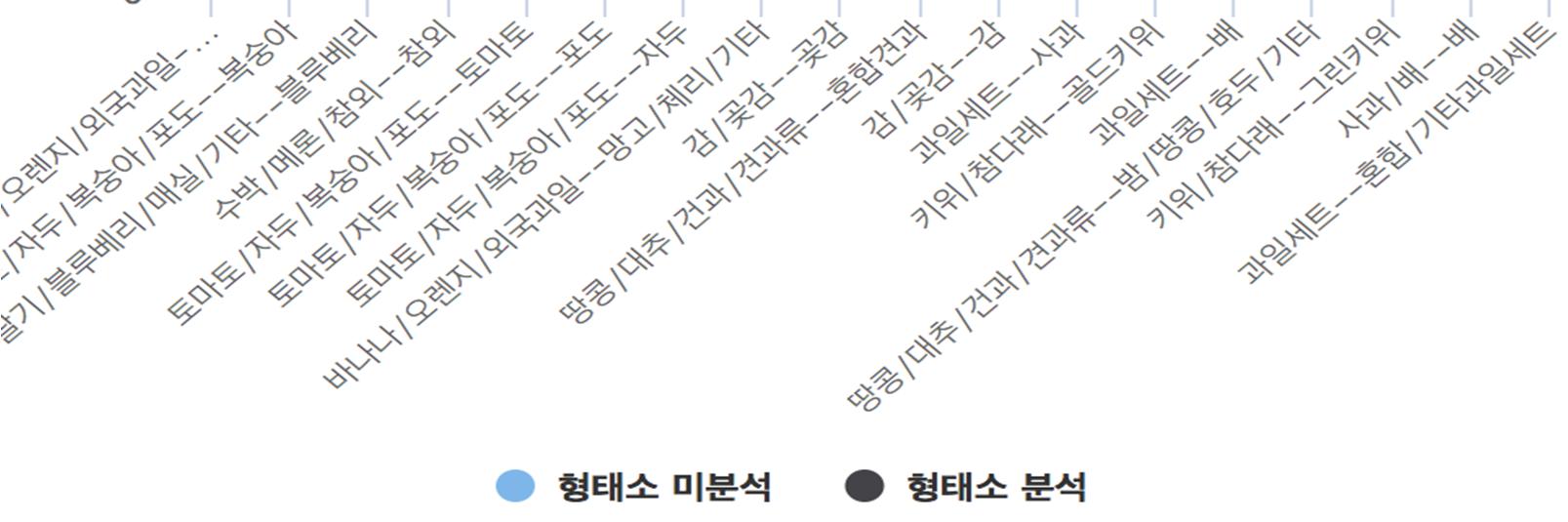
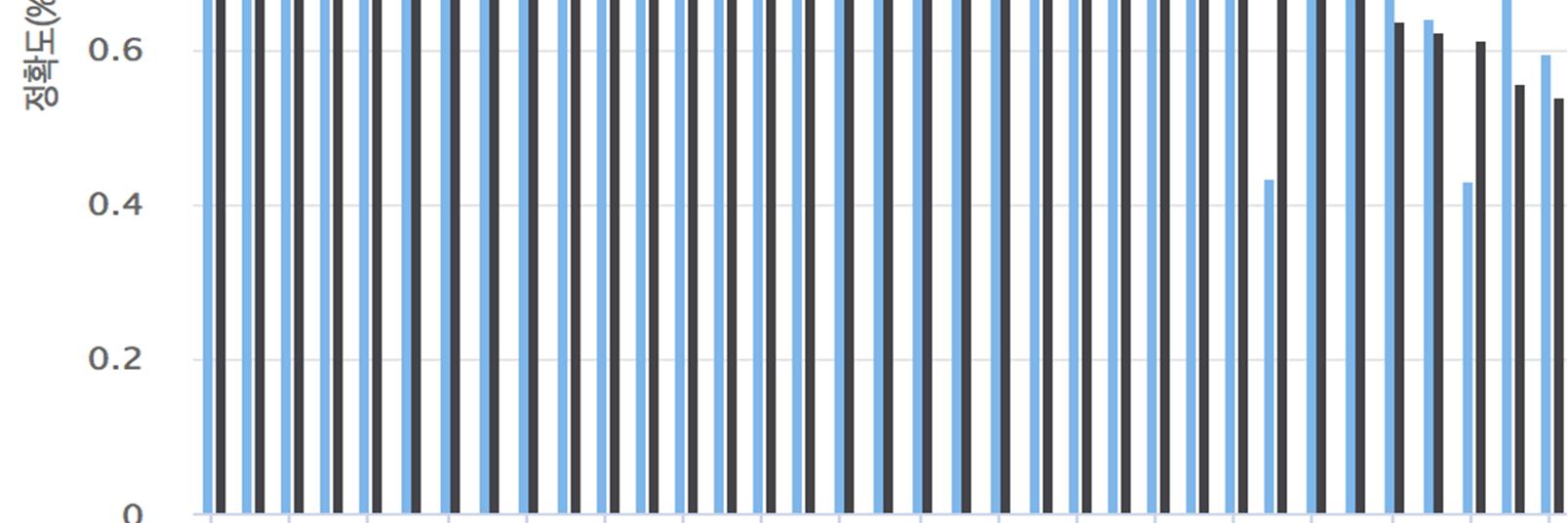
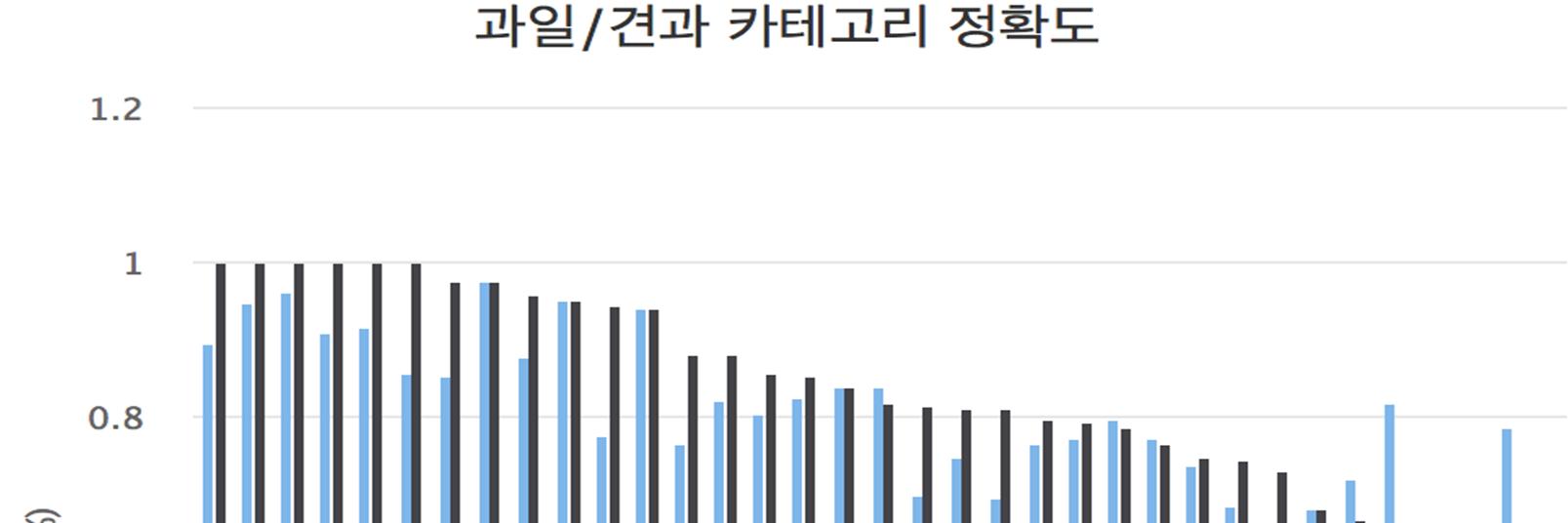
Encoder RNN과 Decoder RNN으로 연결된 sequence to sequence(seq2seq) 모델 사용(텐서플로에서 라이브러리 제공) 또한 텍스트를 계산 하기 위해서는 embedding 시켜서 2차원 벡터 값으로 변경을 하여 입력 값으로 사용. Word embedding은 가공되지 않은 텍스트 데이터를 2차원 수치형 데이터로 만드는 것.

1. **결과**

실험 1 - 과일/견과 카테고리

총 35개의 카테고리를 가지고 있고, 15,966개의 트레이닝 데이터와 1,774 개의 테스트 데이터로 실험 하였다.

그림 11. 과일/견과 카테고리 정확도

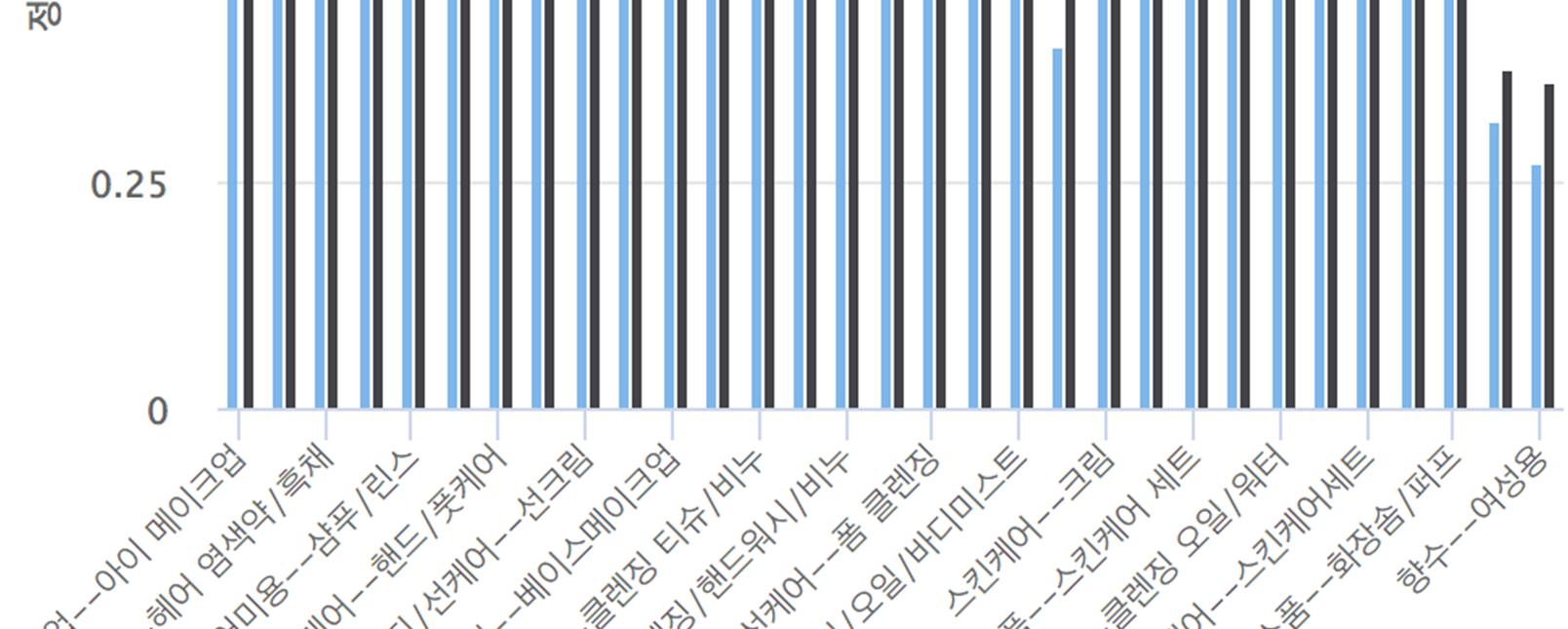
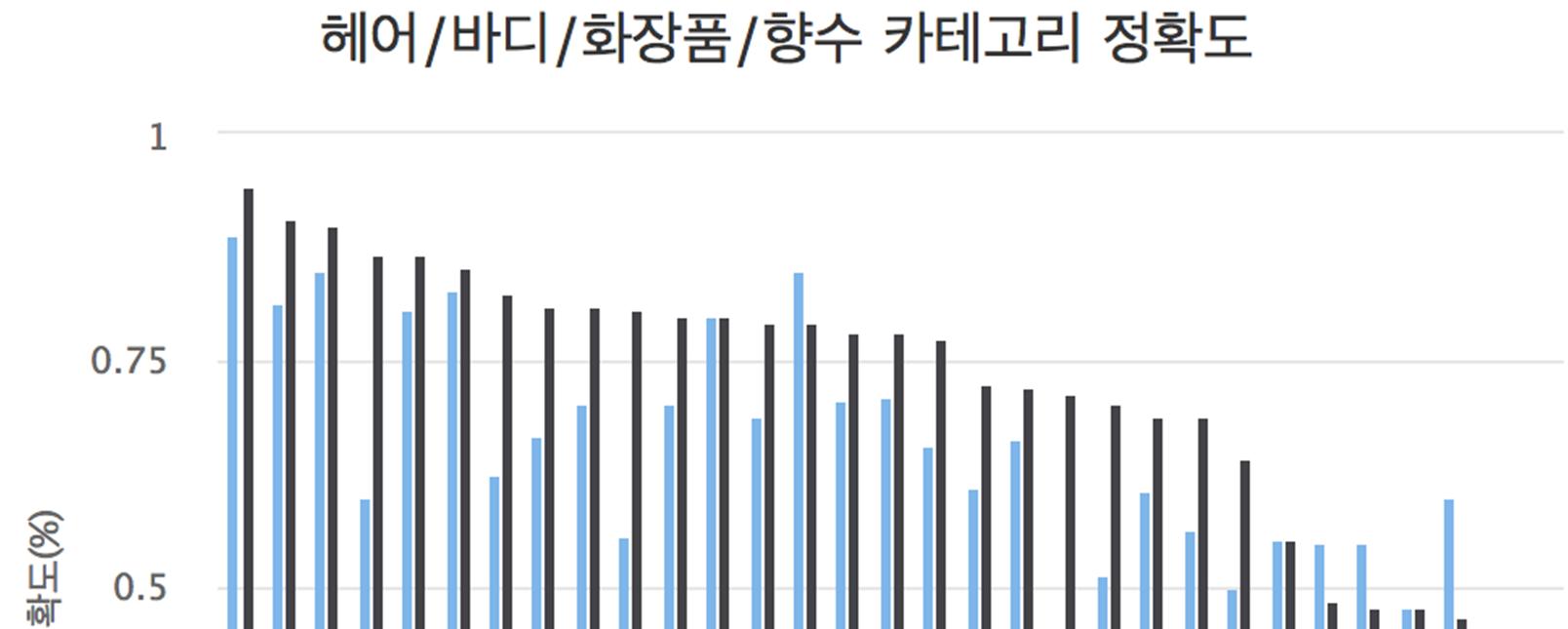


\*일반적인 전처리 : 77.29%의 정확도

\*전처리와 형태소 분석을 이용한 입력값 고도 정제시 80.89%의 정확도

실험 2 - 헤어/바디/화장품/향수 카테고리

총 31개의 카테고리를 가지고 있고, 15,789개의 트레이닝 데이터와 1,754 개의 테스트 데이터로 실험 하였다.

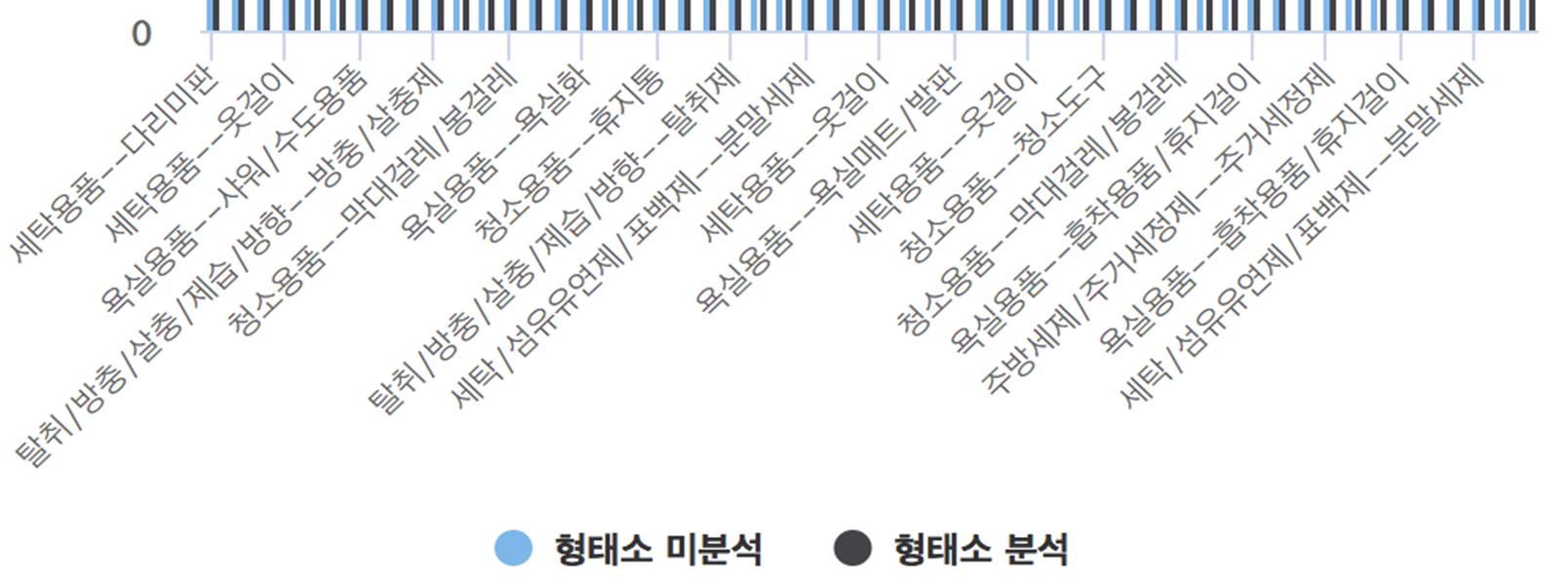
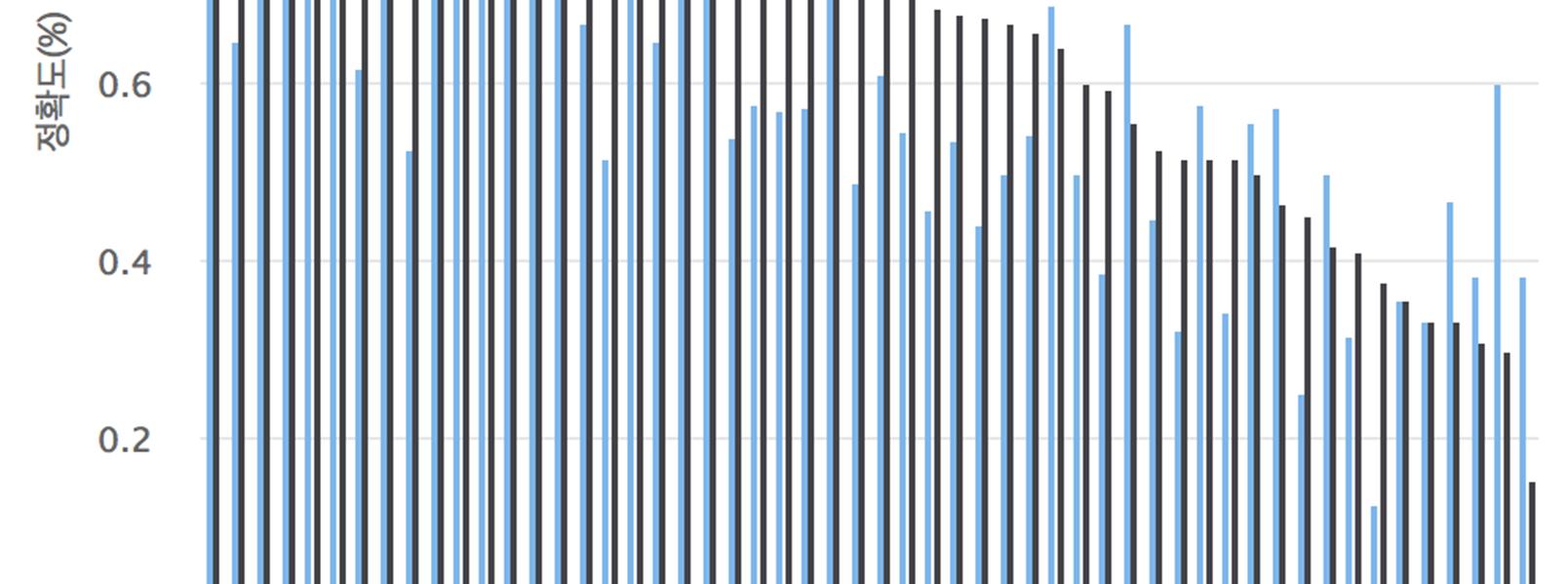
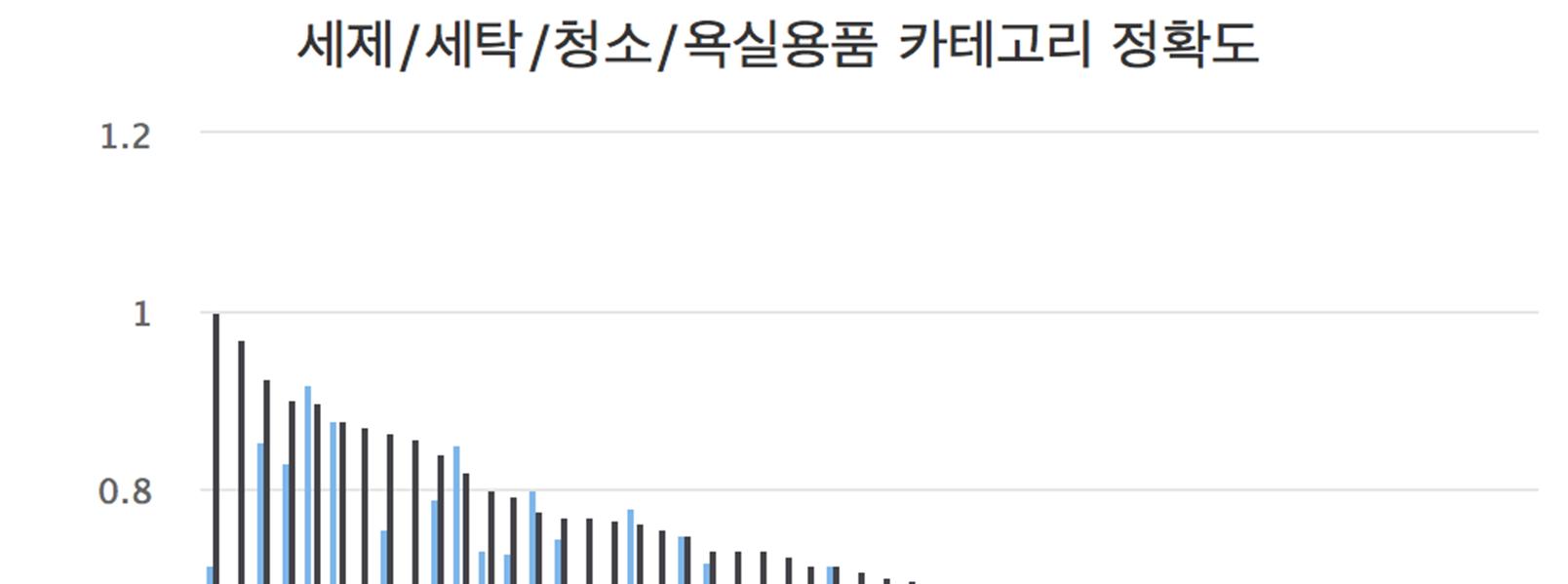


\*일반 적인 전처리 : 66.65%의 정확도

\*전처리와 형태소 분석을 이용한 입력값 고도 정제시 75.09%의 정확도

실험 3 - 세제/세탁/청소/욕실용품 카테고리 정확도

총 54개의 카테고리를 가지고 있고, 21,173개의 트레이닝 데이터와 2,356 개의 테스트 데이터로 실험 하였다.



\*일반 적인 전처리 : 66.33%의 정확도

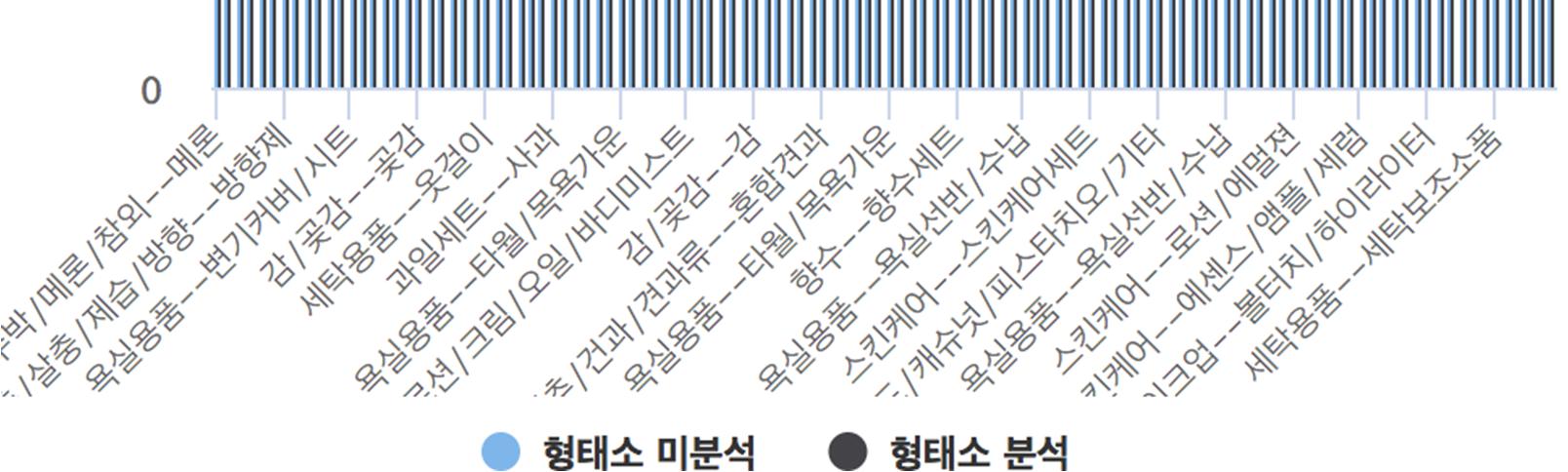
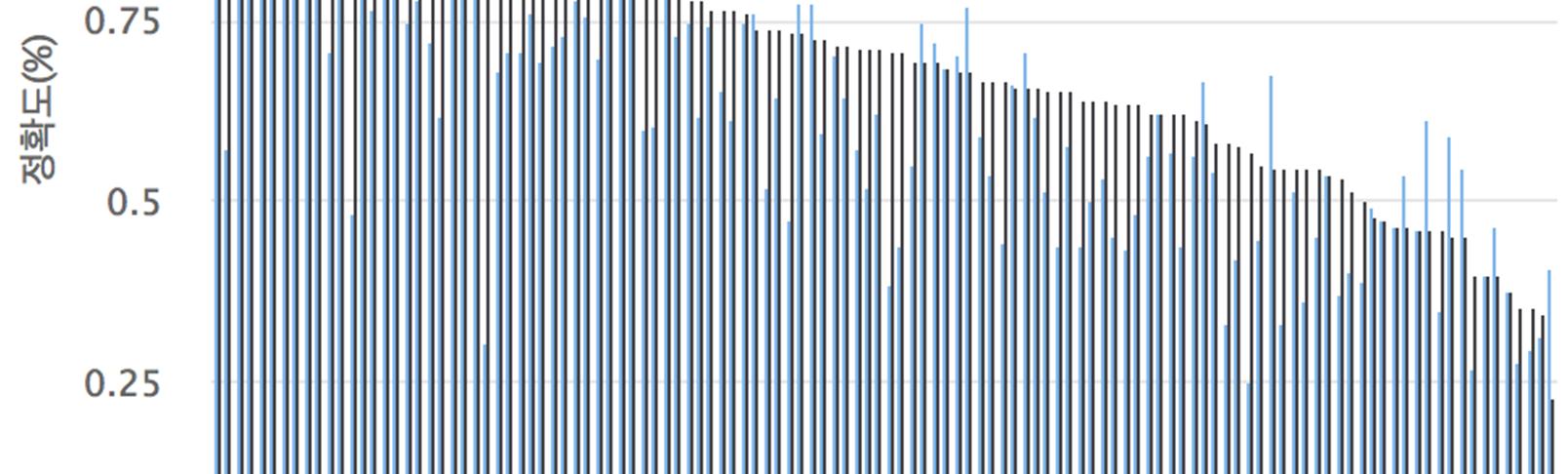
\*전처리와 형태소 분석을 이용한 입력값 고도 정제시 74.53%의 정확도

3개의카테고리전체테스트

이번 실험에서는 위에서 진행한 3 개의 상품-카테고리 데이터 전체를 테스트 하였다.

위의 실험들과 동일한 조건으로 실험이 되었다. Cell 의 개수와 Hidden layer 의 개수, 트레이닝 회수도 동일하다.

실험 데이터는 52,934 개의 트레이닝 데이터와 5,878 개의 테스트 데이터로 이루어진다.



\*일반 적인 전처리 : 65.93%의 정확도

\*전처리와 형태소 분석을 이용한 입력값 고도 정제시 74.52%의 정확도

…….(이하 중략)……

2)Git hub README에 저장

쇼핑몰 카테고리 분류하기 참고용 블로그 [https://medium.com/mighty-data-science-bootcamp/번역-머신러닝을-활용한-제품-카테고리-분류하기-cd75f11d588a](https://medium.com/mighty-data-science-bootcamp/%EB%B2%88%EC%97%AD-%EB%A8%B8%EC%8B%A0%EB%9F%AC%EB%8B%9D%EC%9D%84-%ED%99%9C%EC%9A%A9%ED%95%9C-%EC%A0%9C%ED%92%88-%EC%B9%B4%ED%85%8C%EA%B3%A0%EB%A6%AC-%EB%B6%84%EB%A5%98%ED%95%98%EA%B8%B0-cd75f11d588a)

카카오 쇼핑몰 카테고리 분류하기 깃허브(1위, 파이토치) <https://github.com/kakao-arena/shopping-classification>

카카오 쇼핑몰 카테고리 분류하기 깃허브(3위,텐서플로) <https://github.com/SeoHyeonDeok/S2Bstar-kakao-arena-shopping-classification>

카카오 쇼핑몰 카테고리 분류하기 설명해놓은 블로그 [https://medium.com/@tantara/카카오-아레나-쇼핑몰-상품-카테고리-분류-대회-리뷰-part-1-전처리-8e8778fb5db4](https://medium.com/@tantara/%EC%B9%B4%EC%B9%B4%EC%98%A4-%EC%95%84%EB%A0%88%EB%82%98-%EC%87%BC%ED%95%91%EB%AA%B0-%EC%83%81%ED%92%88-%EC%B9%B4%ED%85%8C%EA%B3%A0%EB%A6%AC-%EB%B6%84%EB%A5%98-%EB%8C%80%ED%9A%8C-%EB%A6%AC%EB%B7%B0-part-1-%EC%A0%84%EC%B2%98%EB%A6%AC-8e8778fb5db4)

네이버 쇼핑몰 카테고리 분류 과정 상세 설명 [https://deview.kr/data/deview/2019/presentation/[125]이미지와%20Text정보들을%20이용한%20쇼핑카테고리%20분류AI.pdf](https://deview.kr/data/deview/2019/presentation/%5B125%5D%EC%9D%B4%EB%AF%B8%EC%A7%80%EC%99%80%20Text%EC%A0%95%EB%B3%B4%EB%93%A4%EC%9D%84%20%EC%9D%B4%EC%9A%A9%ED%95%9C%20%EC%87%BC%ED%95%91%EC%B9%B4%ED%85%8C%EA%B3%A0%EB%A6%AC%20%EB%B6%84%EB%A5%98AI.pdf)

sklearn과 keras를 이용한 딥러닝 기반 상품 카테고리 자동분류 서버 과제 <https://grd0001.tistory.com/19>

**4. 앞으로 계획**

**\*Default = 복습 또 복습**

1) 관련 연구 논문 찾아보고 모르는 부분 공부, 기회가 된다면 딥러닝 코세라 강의 들어보기

2) 데이터 추출, 역할 분담

3) 환경 구축(colab? Linux? AWS? PC?)

4) 전체적인 카테고라이징 시스템 설계

5) 구현