10/22 발표

안녕하십니까 10/22 발표를 시작하겠습니다.

우선 목차를 살펴보면 현재까지 확보한 말뭉치 데이터의 양, CIFAR-10 실습에 관한 내용, 수행계획서대로의 진행 내용 순으로 발표하겠습니다.

현재까지 확보한 말뭉치 데이터는 엑소브레인 말뭉치, 네이버 NLP 챌린지 말뭉치, 국립국어원 말뭉치, AI HUB 말뭉치, U-Corpus 말뭉치가 있습니다.

이 중 엑소브레인과 네이버, 국립국어원은 전년도 팀이 확보한 데이터입니다.

즉 추가적으로 확보한 데이터는 AI-HUB 4975문장, U-Corpus 17727문장으로 총 22719문장입니다.

이와 같은 측정 방식은 AI-HUB, U-Corpus 내의 문장을 추출할 때 한 문장 당 한 줄을 차지하게 출력하여 전체 라인 수를 파악하여 문장 수를 카운트 했습니다.

또한 각자 태그의 기준이 다르기 때문에 통일하기 위한 개체명 사전도 꾸준히 구축 중에 있습니다.

다음으로 CIFAR-10 실습에 대해 알아보겠습니다.

시파-10의 기본 코드를 진행한 결과 정확도는 약 70%로 예상한 75% 수준에는 조금 미치지 못했습니다.

그리고 정확도를 올리기 위해 조사하여 몇가지 방법을 적용해 봤습니다.

첫 번째로 진행한 방법은 배치 노멀라이제이션입니다.

오버피팅을 피하기 위한 근본적 방법은 아니지만 배니싱 그래디언트 문제를 해결할 수 있으며 러닝 레이트를 높여도 비교적 빠르게 학습을 할 수 있습니다.

이와 관련된 코드는 직접 작성이 아닌 조사 시 알게 된 코드로 테스트를 진행하였습니다. 진행한 결과, 정확도는 기본 코드에 조금 못 미치는 결과를 얻었습니다.

다음으로 진행한 방법은 드랍 아웃입니다.

오버피팅을 피하기 위해 주로 사용되는 방법으로 학습시에 신경망의 일부 유닛, 노드들을 제외하는 방법입니다. 물론 제외된 유닛들은 테스트 시에는 정상적으로 사용됩니다.

드랍 아웃의 비율은 0.2로 진행하였고, 결과는 70% 중후반대로 기본 코드에 비해 조금 상승했습니다.

다음으로 진행한 방법은 레귤러라이저입니다. 레귤러라이저는 모델의 복잡도를 제한시키는 방법이며, 여러 논문에서 가중치 감쇠, 웨이트 디케이 라고 표현하기도 합니다.

레귤러라이저의 진행 결과, 70% 초반대로 매우 적은 증가율을 보였습니다.

이러한 3가지 방법 외에도 에포크, 배치 사이즈, 필터 사이즈, 패딩, 스트라이드, 러닝레이트 등의 비율을 수정해봤지만, 드랍아웃만큼의 성능을 확인하진 못했습니다.

또한, 실습을 통해 클래시피케이션 레이어의 동작이나 원리 등을 파악하고자 하였지만 해당 부분에 대해서는 확실하게 알 수 없었습니다.

마지막으로 수행계획서의 계획에 따른 진행 사항입니다.

말뭉치 데이터 확보는 꾸준히 진행 중이며, 전년도 팀의 코드 분석 및 관련 기술 학습 또한 진행중에 있습니다.