3주차 과제

- Momentum (모멘텀)
  - 핵심 아이디어:

    '관성(Momentum)' 또는 '운동량'이라는 물리 법칙을 차용합니다. 현재의 기울기(gradient)뿐만 아니라, 과거의 이동 방향과 속도를 기억하는 '속도'를 도입합니다. 이 속도를 지수 이동 평균(exponential moving average)으로 관리하여, 올바른 방향으로의 이동은 가속화시킵니다.
  - 개선한 문제점:
    지그재그 완화 (Ravine 탐색 개선): 손실 함수 표면이 좁고 긴 골짜기(ravine) 형태일 때, 상하 진동은 상쇄되고 골짜기를 따라 내려가는 속도는 빨라집니다.
    Local Minima 탈출: 얕은 Local Minima에 빠지더라도, 이전에 쌓아온 관성의 힘으로 빠져나올 수 있습니다.
    결과적으로 수렴 속도가 향상됩니다.

- Adagrad (Adaptive Gradient)
  - 핵심 아이디어:

    'Adaptive'라는 이름처럼, 파라미터(가중치)별로 맞춤형 학습률을 적용합니다. 학습 과정 동안 '지금까지 해당 파라미터가 얼마나 변했는지'를 모든 과거 기울기의 제곱 합으로 저장합니다.
  - 적용 방식:

    이 기울기 제곱 합 값이 큰 파라미터(즉, 자주 업데이트되거나 변화량이 컸던 파라미터)는 학습률을 작게 조절하고, 이 값이 작은 파라미터(변화가 적었던 파라미터)는 학습률을 크게 유지합니다.
  - 개선한 문제점:
    SGD가 모든 파라미터에 동일한 학습률을 적용하던 문제를 해결합니다.
    데이터의 희소성(sparsity) 문제를 잘 처리합니다. (예: 자연어 처리에서 드물게 등장하는 단어의 가중치는 더 빠르게 학습시킬 수 있음)
  - 한계점: 기울기 제곱 합은 계속 커지기만 하므로, 학습이 오래 진행되면 분모가 무한히 커져 학습률이 0에 수렴하여 결국 학습이 멈추는(vanishing learning rate) 문제가 발생합니다.

- RMSProp (Root Mean Square Propagation)
  - 핵심 아이디어: Adagrad의 학습률 고갈 문제를 해결하기 위해 제안되었습니다. Adagrad처럼 과거의 모든 기울기 제곱을 더하는 대신, 최근 기울기 제곱의 '지수 이동 평균'을 사용합니다.
  - 적용 방식: '지수 이동 평균'을 사용함으로써, 오래된 기울기 정보의 영향력은 지수적으로 감소시키고 최신 기울기 정보 위주로 학습률을 조절합니다.
  - 개선한 문제점: Adagrad의 기울기 제곱 합이 무한정 커지는 것을 방지하여, 학습률이 0이 되어 학습이 멈추는 현상을 막아줍니다.

- Adam (Adaptive Moment Estimation)
  - 핵심 아이디어: 'Moment Estimation'이라는 이름처럼, 1차 모멘트(Momentum)와 2차 모멘트(RMSProp)의 아이디어를 결합합니다.
  - 적용 방식:
    1. 방향성: Momentum처럼 기울기의 지수 이동 평균을 계산하여 속도를 조절합니다.
    2. 학습률 크기: RMSProp처럼 기울기 제곱의 지수 이동 평균을 계산하여 파라미터별 학습률을 조절합니다.
  - 개선한 문제점: Momentum의 빠른 수렴(방향성)과 RMSProp의 적응적 학습률(안정성)의 장점을 모두 취하여, 빠르고 안정적인 수렴을 보여줍니다. 현재 가장 보편적으로 널리 사용되는 옵티마이저 중 하나입니다.

- AdamW (Adam with Weight Decay)
  - 핵심 아이디어: Adam에서 가중치 감쇠(Weight Decay, L2 정규화)를 적용하는 방식을 수정한 것입니다.
  - 개선한 문제점 (Adam의 문제):
    기존 Adam에서 L2 정규화(가중치 감쇠)를 적용하면, 이 정규화 항이 2차 모멘트에도 영향을 주어 의도한 대로 정규화가 작동하지 않았습니다. (즉, 학습률이 큰 파라미터일수록 정규화 효과가 약해짐)
  - 적용 방식 (개선):
    AdamW는 가중치 감쇠 항을 그래디언트 계산에서 분리(decouple)했습니다.
    Adam의 계산(방향, 크기)은 그대로 하되, 최종 가중치를 업데이트할 때 별도로 가중치 감쇠를 적용합니다.

- 결과: 정규화가 모든 가중치에 일관되게 적용되어, 더 나은 일반화 성능(과적합 방지)을 보여줍니다. (특히 Transformer 기반 모델에서 표준으로 사용됨)