밑바닥부터 시작하는 딥러닝 교재 관련 질문

* **4단원**

1. 활성화 함수의 사용 목적과 활성화 함수가 꼭 비선형 함수여야 하는 이유에 대해 설명하시오.

활성화 함수가 없는 상황을 가정하고 신경망의 입력과 출력간의 관계를 생각해볼 수 있습니다. 이 경우, 출력은 (입력의 선형결합 + 상수) 형태가 될 것 입니다. 즉, 다층 신경망의 의미가 없어지는 것 입니다.

위의 상황과 마찬가지의 상황이 선형활성화 함수를 사용하는 상황입니다. 선형활성화 함수는 입력에 대한 선형결합을 출력합니다. 따라서 (신경망의 layer + 선형활성화 함수) 쌍이 무수히 쌓여 있어도 그 결과는 입력에 대한 선형결합과 다른 것이 없게 되는 것 입니다.

위와 같은 이유로 비선형 활성화 함수는 신경망의 표현력을 증가시키는데 필수적인 요소입니다.

1. 손실함수 2가지를 아는 만큼 설명하고 경사 하강법에 대해 기술하시오.

MSE는 정답에 대한 출력의 오차를 제곱하고 2로 나눈값 입니다. 항등함수를 출력함수로 사용할 경우 그 이전 layer에 역전파시 정답 과 출력의 오차를 전달하게 됩니다. 회귀 문제를 해결할 때 주로 사용합니다.

CEE는 Softmax함수와 같이 분류문제 사용됩니다. Softmax 함수는 입력을 정규화 한 값을 출력하므로 분류문제에 유용하며 CEE와 함께 사용할 경우 이전 layer에 역전파값으로 출력과 정답의 오차를 반환합니다.

경사하강법은 손실함수의 기울기를 기반으로 신경망의 매개변수를 갱신하는 기법입니다. 단번에 손실함수를 최소로 만드는 매개변수값들을 찾는 것은 불가능하기 때문에 지역적 극소점을 찾아가는 방법입니다.

매개변수값을 좌표로한 지점의 손실함수의 기울기를 편미분을 이용해 구하며, 이 기울기는 손실함수의 값을 증가시키는 방향을 가리킵니다.

이 기울기의 반대 방향을 학습률 만큼 확대하여 매개변수에 더하는 방식으로 ‘지역적 극소점’으로 매개변수를 이동시키는 방식이 경사하강법입니다.

* **5단원**

1. 학습이 진행되는 전반적인 과정을 순전파와 역전파와 함께 설명하시오.

순전파는 앞 layer에서 뒤layer로 데이터가 이동하는 과정입니다. 이때 앞 layer의 데이터를 X, 가중치를 W, 편향을 B, 뒤 layer의 값을 Y, 활성화 함수를 h라고 할 때,

Y = h(XW + B) 로 표현할 수 있습니다.

순전파를 통해 입력 데이터로부터 각 layer들의 값이 얻어지며 마지막 출력layer 까지 데이터가 도달하여 값을 출력합니다.

순전파가 완료되면, 출력값과 정답인 one-hot-vector의 비교를 통해 loss function L이 얻어집니다. L에 대한 경사하강법을 적용하기 위해 각 매개변수로 L을 편미분한 값을 사용하는데 이때 역전파법을 사용하게됩니다.

역전파법은 연쇄법칙을 이용해서 뒤 layer의 L에 대한 미분값을 재사용해 앞 layer로 전달하는 방식입니다. 간단한 예시로 위의 layer간 관계에서 아래의 수식으로 역전파 값을 구할 수 있습니다.

dL/dX = (dL/dY)(dY/dX) = (dL/dY)transpose(W)

위의 연쇄법칙을 통한 역전파법은 기존의 미분 값을 재사용하거나 계산량이 적은 값들을 이용하므로 컴퓨터의 자원을 효율적으로 이용하는 방법이라 할 수 있습니다.

이 기울기 값들은 손실함수가 증가하는 방향을 나타내는것이므로 각 매개변수를 기울기\*학습률 만큼 감소시키는 방식으로 학습을 진행합니다.

위의 순전파와 역전파를 학습데이터에 대해 지정한 횟수만큼 반복시킵니다.

1. 분류 모델에서 마지막 계층에 SoftMax 계층을 추가하는 이유는 무엇인가?

먼저, Softmax 계층은 입력에 대한 정규화를 수행합니다. 출력값이 확률을 나타내기 때문에, 입력을 분류하는 문제에 적합합니다.

또한 정답 데이터는 one-hot-vector방식으로 표현되므로 Softmax 계층으로 정답과의 오차를 표현하기 적절하며 CEE를 사용할 경우 역전파시 미분값이 깔끔하게 도출됩니다.

* + **6단원**

1. 최적화 기법인 SGD, Momentum, Adagrad, Adam기법이 어떻게 작동하는지 설명하고 단점을 설명하시오.

SGD, 확률적 경사하강법으로 현재 손실함수의 기울기를 기반으로 매개변수를 갱신합니다. 기울기와 학습률만을 기반으로 하므로 미분값이 급격하게 변화하는 상황이라면 최적의 경로를 찾기 어려운 단점이 있습니다.

또한 최적이 아닌 극소점에서 더 이상 최적화가 작동하지 않습니다.

Momemtum, 속도 개념을 적용한 최적화 기법입니다. 매개변수를 갱신할 때 (속도\*상수) 를 더하고 속도는 (속도\*상수 + 학습률\*기울기)로 갱신됩니다. 이때 (속도\*상수)는 속도의 영향을 줄이는 역할을 합니다.

SGD의 단점인 ‘미분값 변화로 인한 급격한 매개변수 변화’를 방지할 수 있습니다. 지금까지의 매개변수 변화방향과 크기가 ‘속도’로 누적되어 매개변수 변경에 반영되기 때문입니다.

또한 속도를 통해 관성이 작동하므로 최적이 아닌 극소점에서 탈출할 가능성이 있습니다.

하지만 관성의 특징 때문에, 가파른 기울기가 누적되어 속도가 커졌을경우 최적의 point를 지나칠 가능성이 있습니다.

Adagrad, 각 매개변수의 변화율을 서서히 줄여가는 기법입니다. 기울기를 누적해서 저장하며 이것을 이용해 각 매개변수에 대해 학습률을 줄여갑니다. 학습률을 줄여가고, 각 매개변수에 개별적으로 적용하기 때문에 최적의 경로에 정밀하게 접근할 수 있습니다.

하지만 기울기가 누적되며 학습률을 줄이기 때문에 기울기의 크기가 클 경우, 학습률이 급속도로 낮아져 학습에 문제가 될 수 있다는 단점이 있습니다.

Adam, Momentum 기법과 Adagrad기법의 아이디어를 적용한것으로 이해할 수 있습니다. 관성을 적용해서 기울기의 급격한 변화에 따른 급격한 매개변수 변화를 제한하고 있으며, 학습률 또한 손실함수 곡면의 변화에 따라 변화시키는 기법입니다.

1. 가중치 초기값을 모두 0으로 설정하지 않는 이유를 기울기 소실과 함께 설명하시오.

가중치 초기값을 모두 0으로 설정할 경우, 앞 layer에 역전파시 기울기 값으로 0을 전달하게 됩니다. 이는 앞 layer 이전의 layer들이 더 이상 기울기 값을 통한 경사하강법을 실행하지 못하도록 만들어 학습을 중단시킵니다.

또한 가중치가 모두 0이므로 ReLU함수를 활성화 함수로 사용하는경우 0에 가까운 값을 입력하기 때문에 기울기 소실을 야기할 수 있습니다.

1. 배치정규화가 무엇을 수행하는지 설명하고 이에 대한 효과에 대해 설명하시오.

배치정규화는 활성화 layer의 출력값을 고르게 하도록 강제하는 작업입니다. Affine layer와 활성화 layer사이의 배치정규화 layer에서 작업이 이루어집니다.

해당 layer는 Affine layer로부터 입력을 받고 미니배치 단위로 데이터의 표준편차가 1, 평균이 0이 되도록 데이터를 변경합니다.

또한 활성화 함수의 특징에 따라 기울기 소실을 일으키는 입력값들이 있으므로 이를 피하기위해 데이터를 확대, 이동시킵니다.

이렇게 변경된 데이터를 활성화 layer에 출력합니다.

이것을 통해 초기 가중치 값들의 분산에 대해 크게 고려하지 않아도 활성화 함수의 출력의 표현력을 높이고, 기울시 소실을 막을 수 있습니다.