|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 학과 | 컴퓨터 전자시스템공학 | 학번 | 201904458 | 이름 | 이준용 |
| 구분 | 내용 | | | | |
| 학습 범위 | 기계학습 5장 딥러닝 최적화  5.1목적함수: 교차 엔트로피와 로그우도  5.2 성능 향상을 위한 요령 | | | | |
| 학습 내용 | **딥 러닝의 최적화**  과학과 공확에서 최적화  기계학습에 비해서 딥 러닝의 최적화가 훨씬 간단함  일반화 능력이 뛰어나야 함을 전제로 최적화  훈련집합과 테스트집합 간의 관계(대리자 관계)가 최적화를 어렵게 하고 있음  목적함수의 비블록적인 성질, 고차원 특징 공간, 데이터의 희소성 등  긴 시간이 소요  **목적함수: 교차 엔트로피와 로그 우도**  MSE 목적함수(평균 제곱 오차)  -> 오차가 클수록 e의 값이 커지니 벌점으로 표기하기는 알맞음  한계: 벌점의 학습이 더디기 때문에 벌점의 상대적인 수치가 적절하지 않을 수 있음  교차 엔트로피 목적함수  딥러닝은 적어도 분류 문제에서는 로스함수로 MSE대신 CE(Cross Entropy)를 주로 사용한다. (물론, 이를 확률분포로 생각한다면, 연속적인 값에 대한 손실함수는 가우시안 분포를 얻는다고 가정하는 MSE, 이산적인 값에 대해서는 멀티놀리 확률분포를 따를 것이라 가정하는 Cross entropy를 쓰는 것이 맞다.)  Softmax 활성함수  Max를 모방하여 만들어진 함수로, 최댓값은 더욱 활성화되고 작은 값은 억제되는 특징을 가진다  **데이터 전처리**  데이터는 각 종류별로 데이터의 규모를 같은 선상에서 가늠하기 힘들다. 키와 몸무게를 기준으로 하였을 때 m와 kg를 비  교할 경우 kg가 100배 정도 더 큰 규모이기 때문에 첫 번째 특징의 가중치는 100배 정도 느리게 학습된다.  1. 가중치 초기화  처음에 0 이 되면 문제발생. 다 균일한 값을 주면 backprogation 시 두 노드가 같은 일을 하는 중복성 발생한다. 따라서 난수로 가중치를 초기화한다. 가우시언 분포 또는 균일 분포에서 난수를 추출한다.  2. 모멘텀  Gradient의 잡음 현상을 의미하며, 기계 학습은 훈련 집합을 이용하여 Gradient를 추정하므로 잡음이 생길 가능성이 높다. 모멘텀은 그레디언트 스무딩을 가하여 잡음을 줄인다. -> 수렴 속도 빨라진다.  모멘텀의 효과는 오버슈팅 현상을 누그러뜨린다.  3. 적응적 학습률  학습률 p는 너무 크게 주어질 경우 Overshooting 현상으로 인한 진자 현상이 일어날 가능성이 높고 너무 작으면 수렴하는 속도가 매우 느려진다.  이럴 때 Gradient에 학습률 p를 곱하여 모든 매개변수가 같은 크기의 학습률을 사용하게 유도하여 step이 진행됨에 따라 값의 진행을 조절한다.   * AdaGrad – Parameter 중 Element의 움직임이 많아질수록 학습률이 낮아지게 하여 element마다 학습률을 다르게 감소시키는 방법. * RMSProp – 기울기를 단순 누적하는 것이 아니라 지수 가중 이동 평균을 사용하여 최신 기울기들이 AdaGrad에 비해 더 크게 반영하도록 하는 것. * Adam – AdaGrad와 RMSProp을 융합한 형태로, 두 기법에서 AdaGrad와 RMSProp의 계수가 초기에 0으로 biased 되는 문제를 해결하기 위하여 고안된 방법   4. 활성함수  활성값 z를 계산하여 나온 결과 r을 적용하는 과정.  초기에는 선형의 활성함수를 사용하였으나 비선형 문제를 해결하기 위해 퍼셉트론에서 계단식 함수를 사용하였다. 그 이후 tanh 형태에서 현재 쓰이고 있는 ReLU 형태로 진화하였다.  Tanh는 활성값이 커지면 포화 상태가 되고 그레디언트는 0 에 가까워진다. 매개변수 갱신이 매우 느린 요인이다. Relu 함수 살짝 negative 하게 하려면 Leaky Relu 사용한다.  5. 배치 정규화  공변량 시프트 현상  학습이 진행되면서 층 1의 매개변수가 바뀜에 따라 결과값이 따라 바뀜 -> 층 2에 입력되는 데이터의 분포가 변경됨 -> 층3....으로 갈수록 더 심각해진다  이는 학습을 방해하는 요인으로 작용  Gradient Vanishing / Exploding Problem이 발생하지 않도록 하면서 Learning Rate 값을 크게 설정해야 함. 이를 누그러뜨리기 위해 이전의 데이터 정규화를 모든 층에 적용하는 기법. 입력 데이터가 아닌 중간 결과 데이터에 적용하는 것이 더 유리하다  ->다른 곳에 해도 성능 향상에는 도움이 됨.  훈련 집합보다는 미니 배치에 적용하는 것이 훨씬 유리하다.  -> 어느 방향으로 업데이트 할 것인지 계산하는 쪽에 정규화를 하는 것이 더 성능에 도움이 된다.  효과: Traning 결과와 Test 결과 분포가 다를 경우 Traning Data Set에 대한 결과가 Test Data Set의 결과에 맞춰지게됨.   * CNN에서는 노드 단위가 아니라 특징 맵 단위로 학습, 테스트를 적용 * 가중치 초기화에 덜 민감함 * 학습률을 크게 하여 수렴 속도가 향상 * Sigmoid를 활성함수로 사용하는 깊은 신경망도 학습이 이루어짐 * Drop Out을 따로 적용하지 않아도 높은 성능이 이루어짐. | | | | |