dropout과 maxnorm

동국대학교 컴퓨터공학과 김관우, 서혜민

- Max-norm

Max-norm은 모델의 매개변수 값인 신경만의 가중치를 사용자 정의 매개변수보다 작은 계수를 갖도록 제한하는 방법으로 주로 과적합을 피해 모델의 매개변수를 제한하는 방법이다.

weight에 upper bound를 정해 weight가 너무 크게 설정되지 못하게 막아 미분값도 제한된다.

Max-norm Regularization은 가중치 벡터의 길이가 미리 정해 높은 상한 값을 넘지 못하도록 제한하면서 gradient descent 연산도 제한된 조건 하에서만 계산되도록 하는 projected gradient descent를 사용합니다.

$$\|\mathbf{w}\|_2 \leq r$$

모든 뉴런에 대해 입력의 연결 가중치 w의 L2 norm인 $\|w\|_2$를 r(하이퍼파라미터)이하로 제한하는 regularization 기법.

Norm : 벡터의 크기를 측정하는 함수.

$$L_p = (\sum_i^n |x_i|^p)^{\frac{1}{p}}$$

L2 Norm : P가 2인 Norm 함수. n차원 공간에서의 가중치 벡터의 크기를 계산.

$$L_2 = \sqrt{\sum_i^n x_i^2}$$

```
tf.keras.constraints.MaxNorm(
    max_value=2, axis=0
)
```

max_value를 설정해 가중치가 max_value 이하의 표준을 갖도록 제한한다. 앞서 설명한 r의 값이 max_value에 해당한다.

axis는 norm을 계산할 축이다. 만약 Dense 레이어의 가중치 행렬이 (input_dim, output_dim)의 형태를 취할 때 axis를 0으로 설정하면 (input_dim,)의 길이를 갖는 가중치 벡터를 제약할 수 있다.

```
from keras.constraints import max_norm
model.add(Dense(64, kernel_constraint=max_norm(2.)))
```

r를 2로 설정한 실제 코드.

일반적으로 r은 2~4를 사용.

$$\mathbf{w} \leftarrow \mathbf{w} \frac{r}{\|w\|_2}$$

매 훈련 스텝이 끝날 때 마다 $\| w \|_2$를 계산 및 스케일을 조정.

r을 줄이면 w에 더 작은 값이 곱해지니 가중치가 감소하여 overfitting을 감소시킬 수 있다.

dropout은 엄청 효과적이고 간단하면서 maxnorm과 상호보완적이다.

Method	Test Classification error %
L2	1.62
L2 + L1 applied towards the end of training	1.60
L2 + KL-sparsity	1.55
Max-norm	1.35
Dropout + L2	1.25
Dropout + Max-norm	**1.05**

위와 같이 다양한 정규화 기법들을 사용한 테스트 결과가 있는데 error를 살펴보면 dropout과 max-norm을 같이 사용했을 때의 효과가 매우 좋은 것을 확인할 수 있다.

(위의 예시에서는 MNIST 데이터를 사용해 테스트된 결과이다. - MNIST: 손 글씨 데이터)

dropout은 음성, 영상 텍스트 등 응용분야에 관계없이 신경망의 성능을 개선시킬 수 있으며 BERT에 Max-norm과 같이 적용했을 경우 성능을 크게 개선할 수 있을 것으로 예상된다.

- Bach size

=> 메모리의 한계와 속도 저하 때문에 전체 데이터 셋을 여러 개의 그룹으로 나눠서 학습을 시키는 데, 이때 하나의 그룹에 속하는 데이터 수를 의미. 즉, 연산 한번에 들어가는 데이터 수.

크게 하는 경우) 학습 횟수가 줄어들고, gpu를 최대한 활용하여, 모델의 학습 속도가 향상. training set 분포를 좀 더 근사화하여 추정 가능. 그러나 overfitting이 일어날 수 있다.

작게 하는 경우) step이 많아지므로, 학습 횟수가 많아져서 local minima가 발생 가능. 너무 작게 하면 작은 데이터를 대상으로 가중치가 자주 업데이트 될 수 있기 때문에 학습이 불안정할 수 있음.