Gradient Gating

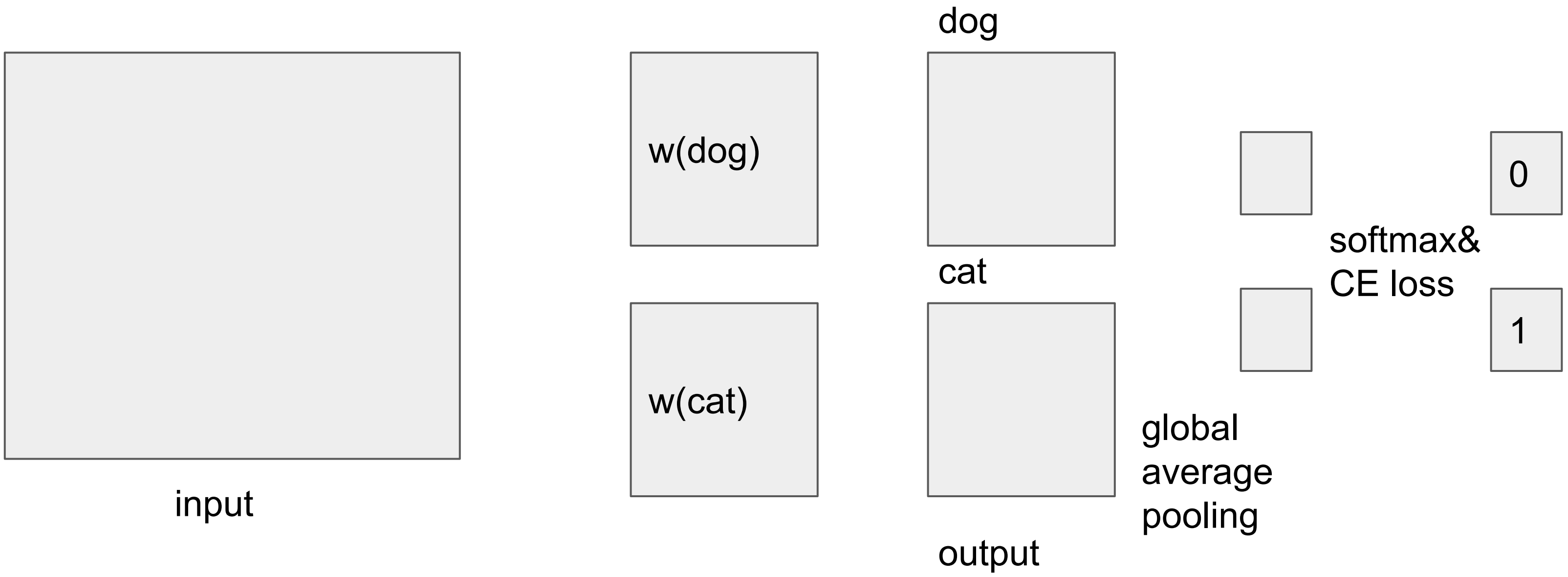
By BeomGon Yu (2021-06-28)

in backpropagation,(Global average pooling or convolution),
how about use its position info also for gradient update(label info + position info)

use gradient gating by feature map of each layer
feature map include the how much each map include the feature

Forward pass

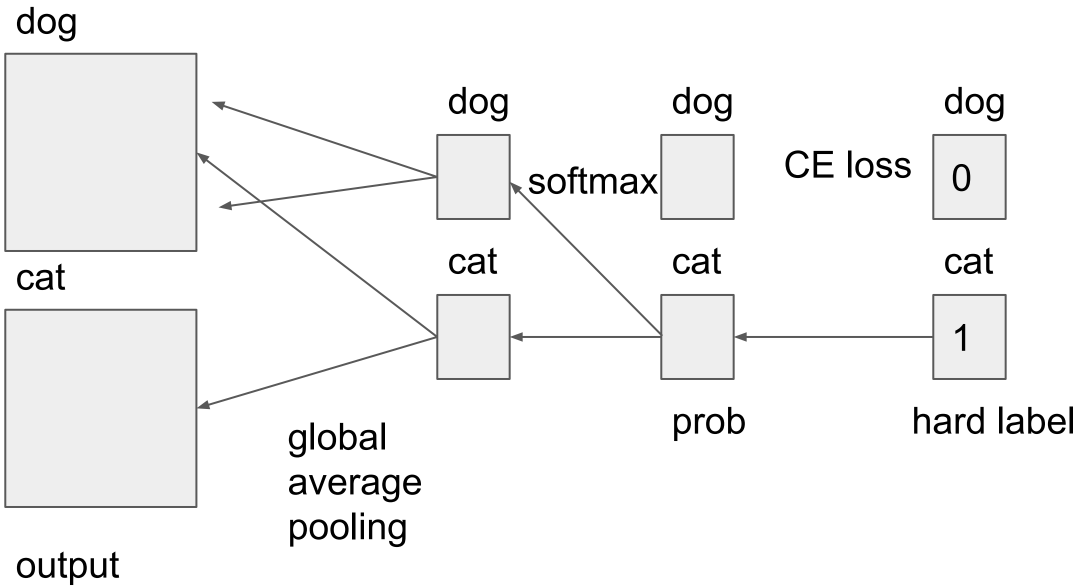
cat is on right/top



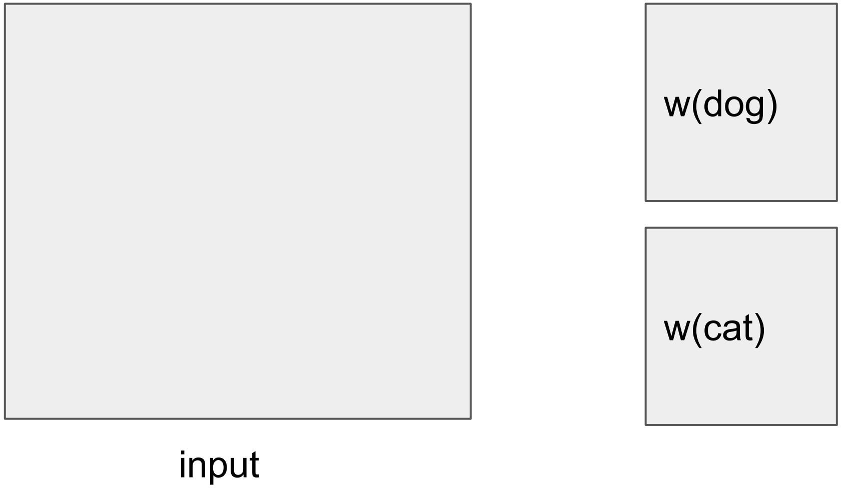
input image is cat and suppose its on the top/left, therefore, cat's output feature map's top left would be high, and other part will be noise.

Backward

gradient flow backwardly,
in global average pooling,
gradient is devided by N,
equally,
But I think more gradient
should flow to top left



when w(cat) filter will do convolution from left to right, top to down, only left, top is useful, other part is noise. however in backpropagation in convolution, other pard is also used. when update the w(cat), how can we add more value on top/left featuremap??



x is input y is output feature map, y11 is high, near value is middle, other value is low. w is trained for capturing cap

use output feature map for more gradient to flow through y11

$$x = \begin{bmatrix} x_{11} & x_{12} & x_{13} & x_{14} \\ x_{21} & x_{22} & x_{23} & x_{24} \\ x_{31} & x_{32} & x_{33} & x_{34} \\ x_{41} & x_{42} & x_{43} & x_{44} \end{bmatrix}$$

$$w = \begin{bmatrix} w_{11} & w_{12} \\ w_{21} & w_{22} \end{bmatrix}$$

$$y = \begin{bmatrix} y_{11} & y_{12} & y_{13} \\ y_{21} & y_{22} & y_{23} \\ y_{31} & y_{32} & y_{33} \end{bmatrix}$$

$$dw = \begin{bmatrix} x_{11} & x_{12} & x_{13} & x_{14} \\ x_{21} & x_{22} & x_{23} & x_{24} \\ x_{31} & x_{32} & x_{33} & x_{34} \\ x_{41} & x_{42} & x_{43} & x_{44} \end{bmatrix} * \begin{bmatrix} dy_{11} & dy_{12} & dy_{13} \\ dy_{21} & dy_{22} & dy_{23} \\ dy_{31} & dy_{32} & dy_{33} \end{bmatrix} = x * dy$$

$$dx = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & dy_{11} & dy_{12} & dy_{13} & 0 \\ 0 & dy_{21} & dy_{22} & dy_{23} & 0 \\ 0 & dy_{31} & dy_{32} & dy_{33} & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix} * \begin{bmatrix} w_{22} & w_{21} \\ w_{12} & w_{11} \end{bmatrix} = dy\_0 * w'$$

의의

기존 gradient update 방법에 새로운 방법을 제안하는 의미가 있다.

gradient gating 관련 최적화 또는 다른 방법과 조화시 성능 향상 기대

back propagation에 대한 이해 증진

adam과 비교시 adam은 항상 일정 수준으로 안정적으로 학습이 되나,
gated의 경우 종종 성능이 떨어지는 경우 존재.
대신 adam과 달리 추가적인 parameter나 연산이 거의 없어 메모리 saving이 가능하다.

Y = WX

dy/dx = W -> W*sigmoid(Y)

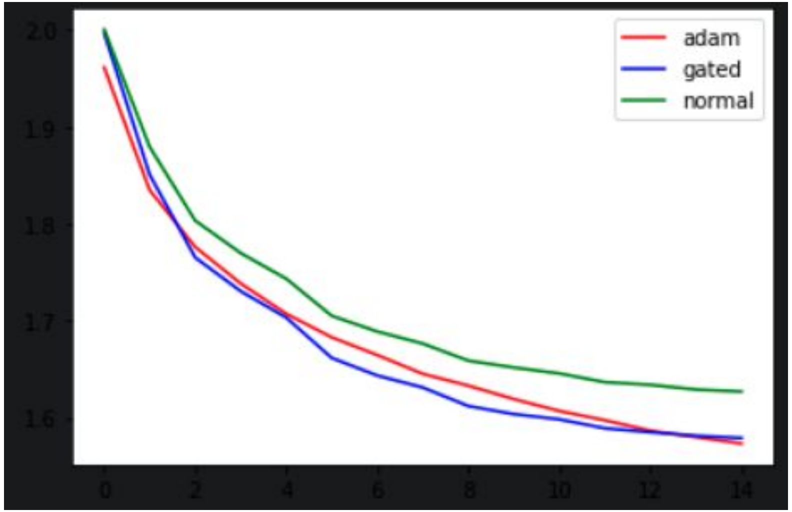
dy/dw = X -> X*sigmoid(Y)

Y is higher, more grad flow to that point(each map in layer)

직관적으로, if Loss is high, more grad is ok.

cifar 10 결과 비교

Training loss



cifar 10 결과 비교

validation accuracy