10주차(1/3)

다층 신경망 모델링

파이썬으로 배우는 기계학습

한동대학교
김영섭 교수

다층 신경망 모델링

- 학습 목표
  - 미분의 연쇄법칙을 학습한다.
  - 오차함수의 행렬 표기에서 미분하는 방법을 학습한다.
  - 다층 인공 신경망의 행렬 모델을 학습한다.

- 학습 내용
  - 미분의 연쇄법칙
  - 오차함수의 행렬 미분
  - 다층 인공 신경망 행렬 모델

1. 연쇄법칙: 연쇄법칙이란?

- 어떤 변수의 변화가 매개변수의 변화를 일으키고 그 후 최종 함수 값의 변화를 유발하는 것

1. 연쇄법칙: 연쇄법칙의 예

$$F(x) = f(g(x))$$

$$F'(x) = f'\big(g(x)\big)g'(x)$$

1. 연쇄법칙: 연쇄법칙의 예

$$F(x) = f(g(x))$$

$$F'(x) = f'(g(x))g'(x)$$

라이프니츠 표기법

$y = f(u), u = g(x)$ 일 때,

$$\frac{dy}{dx} = \frac{dy}{du} \cdot \frac{du}{dx}$$

(1)번 (2)번

2. $W^{[2]}$의 오차함수 미분 : 복습

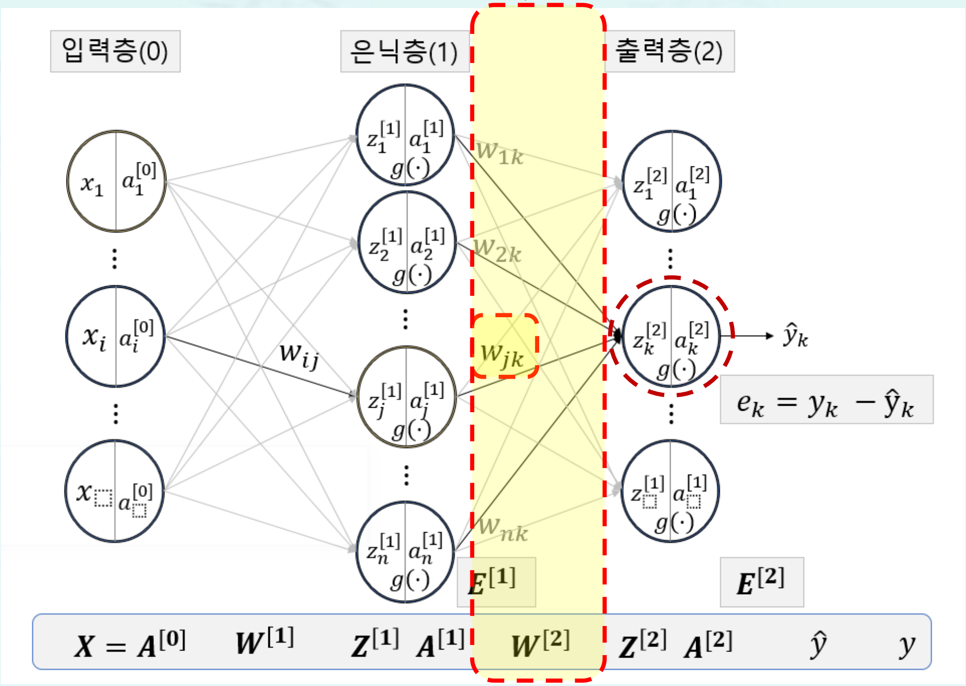
$$\frac{\partial E}{\boxed{\partial w_{jk}}} = -(y_k - \hat{y}_k) \cdot \frac{\partial}{\partial w_{jk}} g(z_k)$$

$$= -(y_k - \hat{y}_k) \cdot g'(z_k) \frac{\partial z_k}{\partial w_{jk}}$$

$$= -(y_k - \hat{y}_k) \cdot g'(z_k) \frac{\partial}{\partial w_{jk}} \left(\sum_j w_{jk} \cdot a_j\right)$$

$$= -(y_k - \hat{y}_k) \cdot g'(z_k) \cdot a_j$$

$$\frac{\partial E}{\partial W^{[2]}} = -E^{[2]} \cdot g'(Z^{[2]}) \cdot A^{[1]T}$$



2. $W^{[2]}$의 오차함수 미분 : 복습

- **1단계**

$$W^{[2]} := W^{[2]} - \alpha\, \Delta W^{[2]}$$

$$= W^{[2]} - \alpha \frac{\partial E}{\partial W^{[2]}}$$

- **2단계**

$$w_{jk}^{[2]} := w_{jk}^{[2]} - \alpha \Delta w_{jk}^{[2]}$$

$$= w_{jk}^{[2]} - \alpha \frac{\partial E}{\partial w_{jk}^{[2]}}$$

- **3단계**

$$\frac{\partial E}{\partial w_{jk}} = -(y_k - \hat{y}_k) \cdot \frac{\partial}{\partial w_{jk}} g(z_k)$$

- **4단계**

$$\Delta w_{jk}^{[2]} = \frac{\partial E}{\partial w_{jk}} = -(y_k - \hat{y}_k) \cdot g'(z_k) \cdot a_j$$

$$\Delta W^{[2]} = \frac{\partial E}{\partial W^{[2]}} = -E^{[2]} \cdot g'(Z^{[2]}) \cdot A^{[1]T}$$

- **결론**

$$W^{[2]} := W^{[2]} - \alpha \Delta W^{[2]}$$

$$= W^{[2]} - \alpha \frac{\partial E}{\partial W^{[2]}}$$

$$= \boxed{W^{[2]} + E^{[2]} \cdot g'(Z^{[2]}) \cdot A^{[1]T}}$$

3. 오차함수의 행렬 미분 : $W^{[2]}$의 오차함수 미분

$$W^{[2]} := W^{[2]} - \alpha \boxed{\Delta W^{[2]}}$$

$$= W^{[2]} - \alpha \frac{\partial E}{\partial W^{[2]}}$$

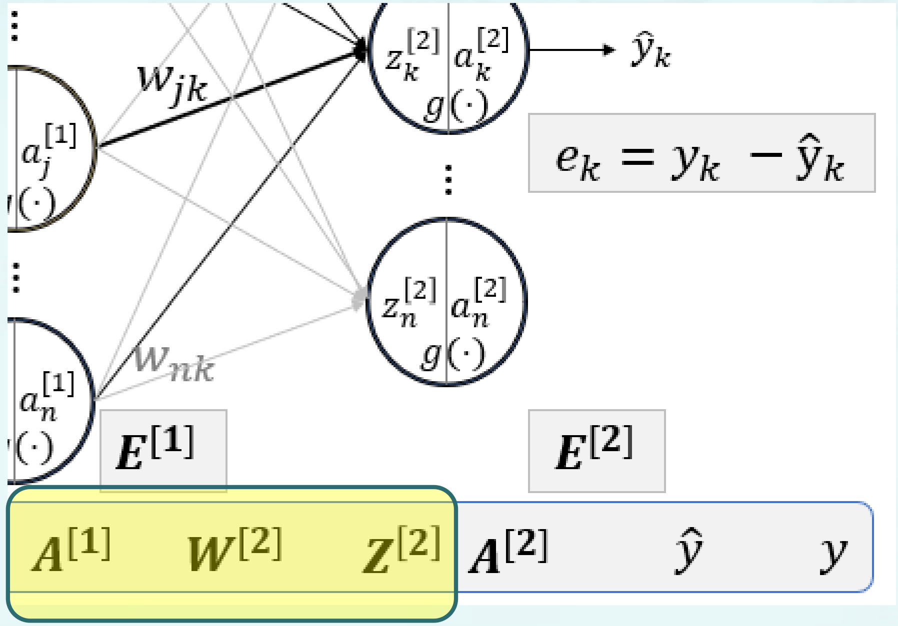
$$\boxed{\Delta W^{[2]}} = \frac{\partial E}{\partial W^{[2]}}$$

3. 오차함수의 행렬 미분 : $W^{[2]}$의 오차함수 미분

$$\Delta W^{[2]} = \frac{\partial E}{\partial W^{[2]}}$$

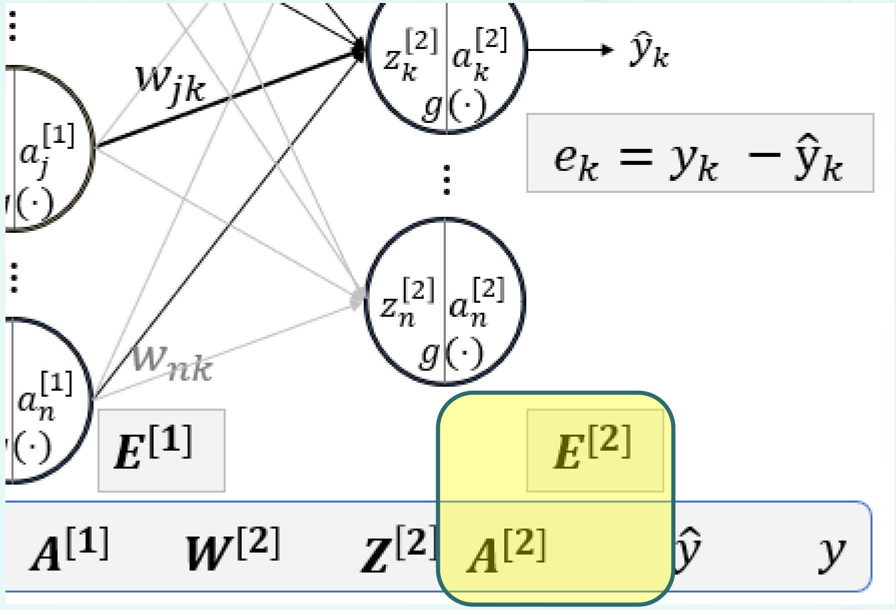
$$= \frac{\partial E}{\partial A^{[2]}} \cdot \frac{\partial A^{[2]}}{\partial W^{[2]}}$$

$$= \frac{\partial E}{\partial A^{[2]}} \cdot \frac{\partial A^{[2]}}{\partial Z^{[2]}} \cdot \frac{\partial Z^{[2]}}{\partial W^{[2]}}$$



3. 오차함수의 행렬 미분 : $W^{[2]}$의 오차함수 미분

$$\Delta W^{[2]} = \frac{\partial E}{\partial W^{[2]}}$$
$$= \frac{\partial E}{\partial A^{[2]}} \cdot \frac{\partial A^{[2]}}{\partial W^{[2]}}$$
$$= \boxed{\frac{\partial E}{\partial A^{[2]}}} \cdot \frac{\partial A^{[2]}}{\partial Z^{[2]}} \cdot \frac{\partial Z^{[2]}}{\partial W^{[2]}}$$



3. 오차함수의 행렬 미분 : $W^{[2]}$의 오차함수 미분

$$E = \frac{1}{2}\left(A^{[2]} - Y\right)^2$$

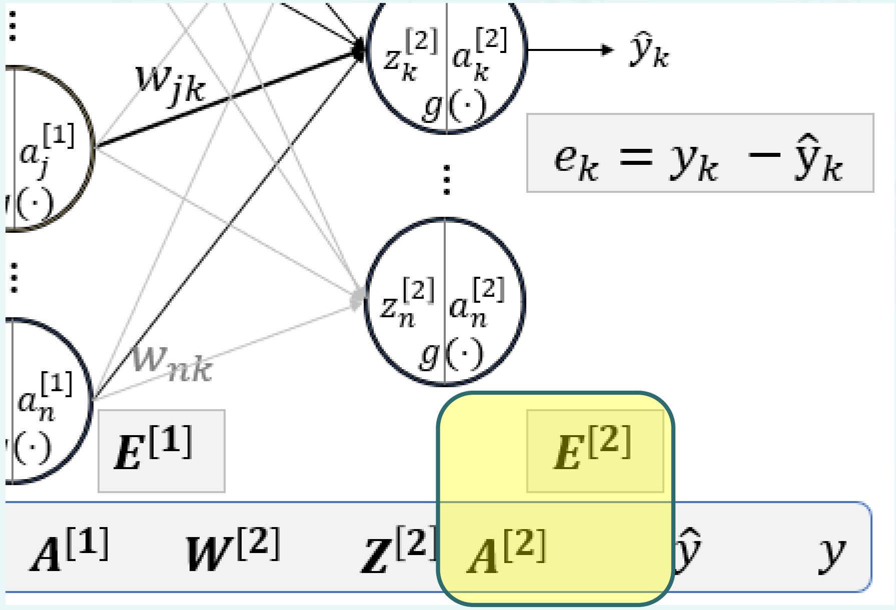
$$E^{[2]} := \left(A^{[2]} - Y\right)$$

$$\Delta W^{[2]} = \frac{\partial E}{\partial W^{[2]}}$$

$$= \frac{\partial E}{\partial A^{[2]}} \cdot \frac{\partial A^{[2]}}{\partial W^{[2]}}$$

$$= \boxed{\frac{\partial E}{\partial A^{[2]}}} \cdot \frac{\partial A^{[2]}}{\partial Z^{[2]}} \cdot \frac{\partial Z^{[2]}}{\partial W^{[2]}}$$

$$= E^{[2]} \cdot \frac{\partial A^{[2]}}{\partial Z^{[2]}} \cdot \frac{\partial Z^{[2]}}{\partial W^{[2]}}$$



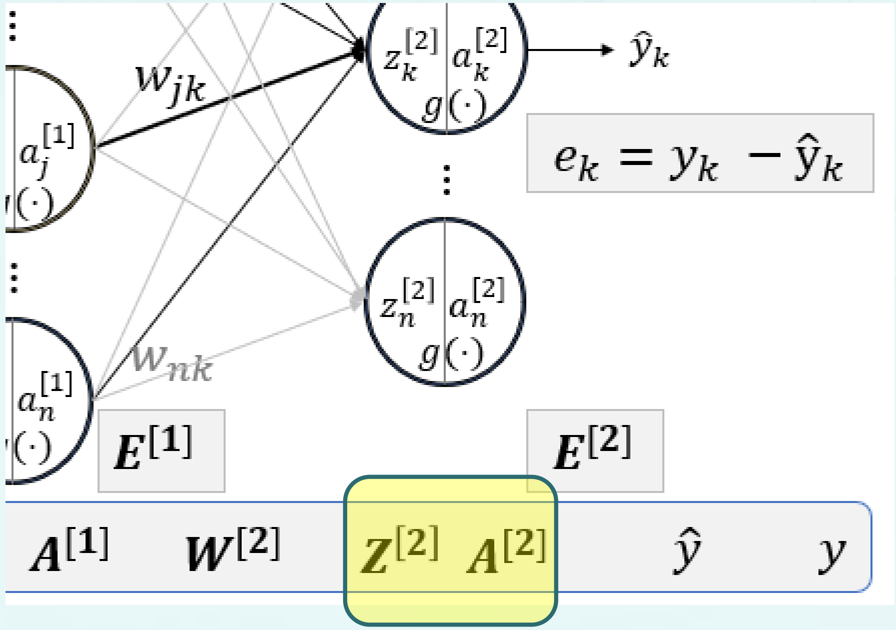
3. 오차함수의 행렬 미분 : $W^{[2]}$의 오차함수 미분

$$\Delta W^{[2]} = \frac{\partial E}{\partial W^{[2]}}$$

$$= \frac{\partial E}{\partial A^{[2]}} \cdot \frac{\partial A^{[2]}}{\partial W^{[2]}}$$

$$= \frac{\partial E}{\partial A^{[2]}} \cdot \frac{\partial A^{[2]}}{\partial Z^{[2]}} \cdot \frac{\partial Z^{[2]}}{\partial W^{[2]}}$$

$$= E^{[2]} \cdot \boxed{\frac{\partial A^{[2]}}{\partial Z^{[2]}}} \cdot \frac{\partial Z^{[2]}}{\partial W^{[2]}}$$



3. 오차함수의 행렬 미분 : $W^{[2]}$의 오차함수 미분

$$A^{[2]} = g(Z^{[2]})$$

$$\begin{aligned}
\Delta W^{[2]} &= \frac{\partial E}{\partial W^{[2]}} \\
&= \frac{\partial E}{\partial A^{[2]}} \cdot \frac{\partial A^{[2]}}{\partial W^{[2]}} \\
&= \frac{\partial E}{\partial A^{[2]}} \cdot \frac{\partial A^{[2]}}{\partial Z^{[2]}} \cdot \frac{\partial Z^{[2]}}{\partial W^{[2]}} \\
&= E^{[2]} \cdot \frac{\partial A^{[2]}}{\partial Z^{[2]}} \cdot \frac{\partial Z^{[2]}}{\partial W^{[2]}} \\
&= E^{[2]} \cdot \boxed{g'(Z^{[2]})} \cdot \frac{\partial Z^{[2]}}{\partial W^{[2]}}
\end{aligned}$$

3. 오차함수의 행렬 미분 : $W^{[2]}$의 오차함수 미분

$$E^{[2]} \cdot \boxed{\sigma(Z^{[2]})(1-\sigma(Z^{[2]}))} \cdot \frac{\partial Z^{[2]}}{\partial W^{[2]}}$$

$$\Delta W^{[2]} = \frac{\partial E}{\partial W^{[2]}}$$

$$= \frac{\partial E}{\partial A^{[2]}} \cdot \frac{\partial A^{[2]}}{\partial W^{[2]}}$$

$$= \frac{\partial E}{\partial A^{[2]}} \cdot \frac{\partial A^{[2]}}{\partial Z^{[2]}} \cdot \frac{\partial Z^{[2]}}{\partial W^{[2]}}$$

$$= E^{[2]} \cdot \frac{\partial A^{[2]}}{\partial Z^{[2]}} \cdot \frac{\partial Z^{[2]}}{\partial W^{[2]}}$$

$$= E^{[2]} \cdot \boxed{g'(Z^{[2]})} \cdot \frac{\partial Z^{[2]}}{\partial W^{[2]}}$$

3. 오차함수의 행렬 미분 : $W^{[2]}$의 오차함수 미분

$$\Delta W^{[2]} = \frac{\partial E}{\partial W^{[2]}}$$

$$= \frac{\partial E}{\partial A^{[2]}} \cdot \frac{\partial A^{[2]}}{\partial W^{[2]}}$$

$$= \frac{\partial E}{\partial A^{[2]}} \cdot \frac{\partial A^{[2]}}{\partial Z^{[2]}} \cdot \frac{\partial Z^{[2]}}{\partial W^{[2]}}$$

$$= E^{[2]} \cdot \frac{\partial A^{[2]}}{\partial Z^{[2]}} \cdot \frac{\partial Z^{[2]}}{\partial W^{[2]}}$$

$$= E^{[2]} \cdot \boxed{g'(Z^{[2]})} \cdot \frac{\partial Z^{[2]}}{\partial W^{[2]}}$$

$$E^{[2]} \cdot \boxed{\sigma(Z^{[2]})(1 - \sigma(Z^{[2]}))} \cdot \frac{\partial Z^{[2]}}{\partial W^{[2]}}$$

$$= E^{[2]} \cdot \boxed{A^{[2]}\,(1 - A^{[2]})} \cdot \frac{\partial Z^{[2]}}{\partial W^{[2]}}$$

3. 오차함수의 행렬 미분 : $W^{[2]}$의 오차함수 미분

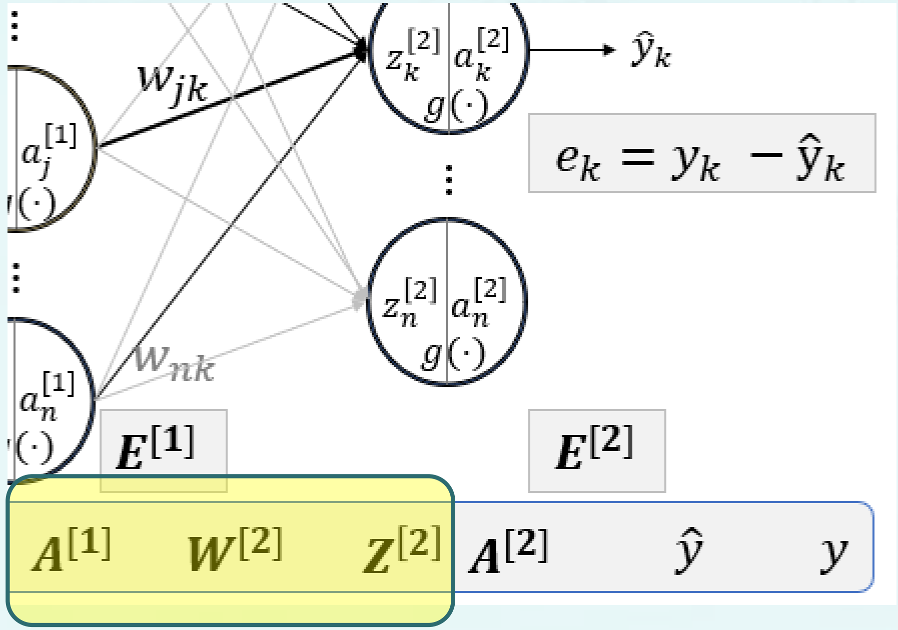
$$\Delta W^{[2]} = \frac{\partial E}{\partial W^{[2]}}$$

$$= \frac{\partial E}{\partial A^{[2]}} \cdot \frac{\partial A^{[2]}}{\partial W^{[2]}}$$

$$= \frac{\partial E}{\partial A^{[2]}} \cdot \frac{\partial A^{[2]}}{\partial Z^{[2]}} \cdot \frac{\partial Z^{[2]}}{\partial W^{[2]}}$$

$$= E^{[2]} \cdot \frac{\partial A^{[2]}}{\partial Z^{[2]}} \cdot \frac{\partial Z^{[2]}}{\partial W^{[2]}}$$

$$= E^{[2]} \cdot g'(Z^{[2]}) \cdot \boxed{\frac{\partial Z^{[2]}}{\partial W^{[2]}}}$$



3. 오차함수의 행렬 미분 : $W^{[2]}$의 오차함수 미분

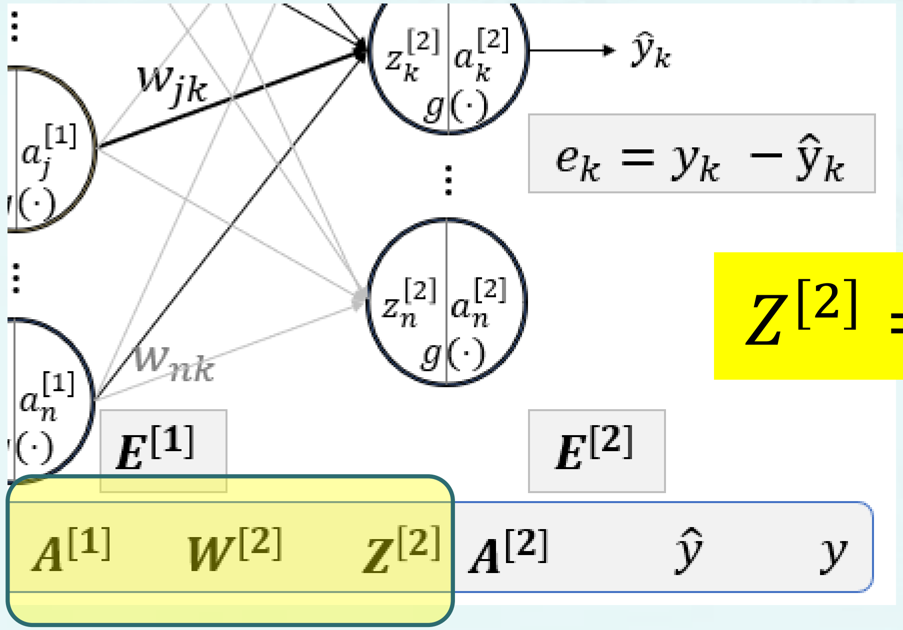
$$\Delta W^{[2]} = \frac{\partial E}{\partial W^{[2]}}$$

$$= \frac{\partial E}{\partial A^{[2]}} \cdot \frac{\partial A^{[2]}}{\partial W^{[2]}}$$

$$= \frac{\partial E}{\partial A^{[2]}} \cdot \frac{\partial A^{[2]}}{\partial Z^{[2]}} \cdot \frac{\partial Z^{[2]}}{\partial W^{[2]}}$$

$$= E^{[2]} \cdot \frac{\partial A^{[2]}}{\partial Z^{[2]}} \cdot \frac{\partial Z^{[2]}}{\partial W^{[2]}}$$

$$= E^{[2]} \cdot g'(Z^{[2]}) \cdot \boxed{\frac{\partial Z^{[2]}}{\partial W^{[2]}}}$$

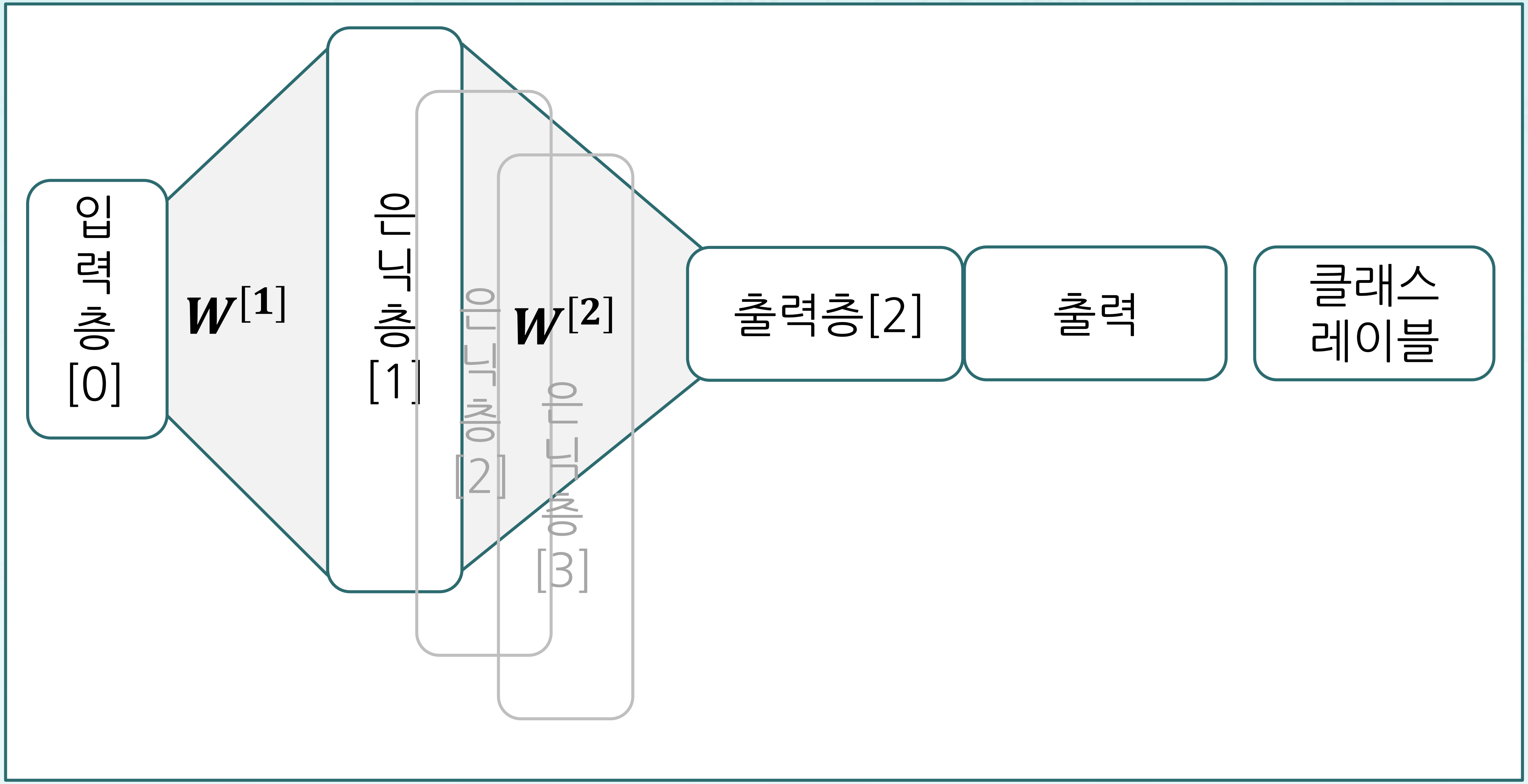


$$Z^{[2]} = W^{[2]} A^{[1]}$$

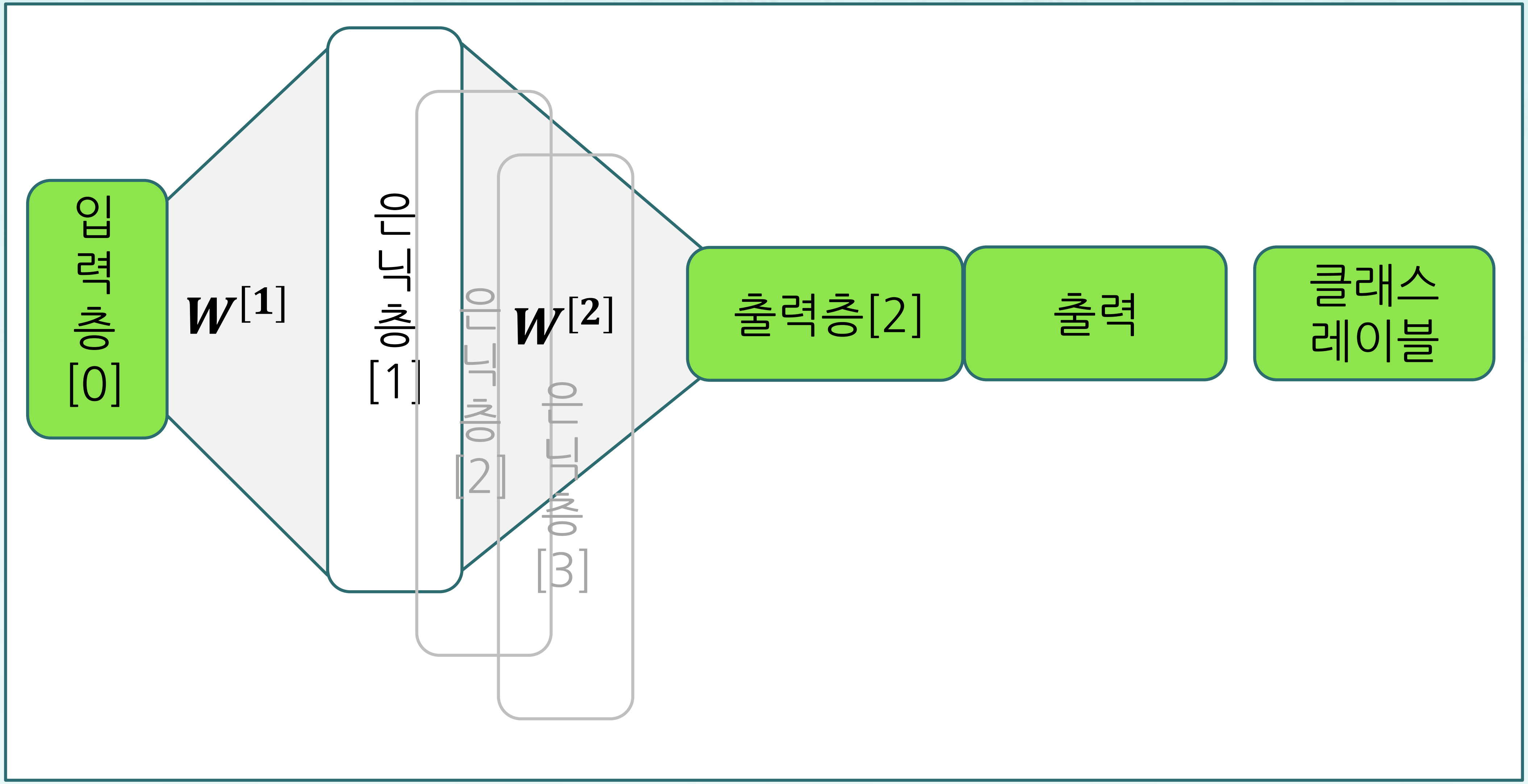
3. 오차함수의 행렬 미분 : $W^{[2]}$의 오차함수 미분

$$
\begin{aligned}
\Delta W^{[2]} &= \frac{\partial E}{\partial W^{[2]}} \\
&= \frac{\partial E}{\partial A^{[2]}} \cdot \frac{\partial A^{[2]}}{\partial W^{[2]}} \\
&= \frac{\partial E}{\partial A^{[2]}} \cdot \frac{\partial A^{[2]}}{\partial Z^{[2]}} \cdot \frac{\partial Z^{[2]}}{\partial W^{[2]}} \\
&= E^{[2]} \cdot \frac{\partial A^{[2]}}{\partial Z^{[2]}} \cdot \frac{\partial Z^{[2]}}{\partial W^{[2]}} \\
&= E^{[2]} \cdot g'(Z^{[2]}) \cdot \frac{\partial Z^{[2]}}{\partial W^{[2]}} \\
&= E^{[2]} \cdot g'(Z^{[2]}) \cdot \boxed{A^{[1]\cdot T}}
\end{aligned}
$$

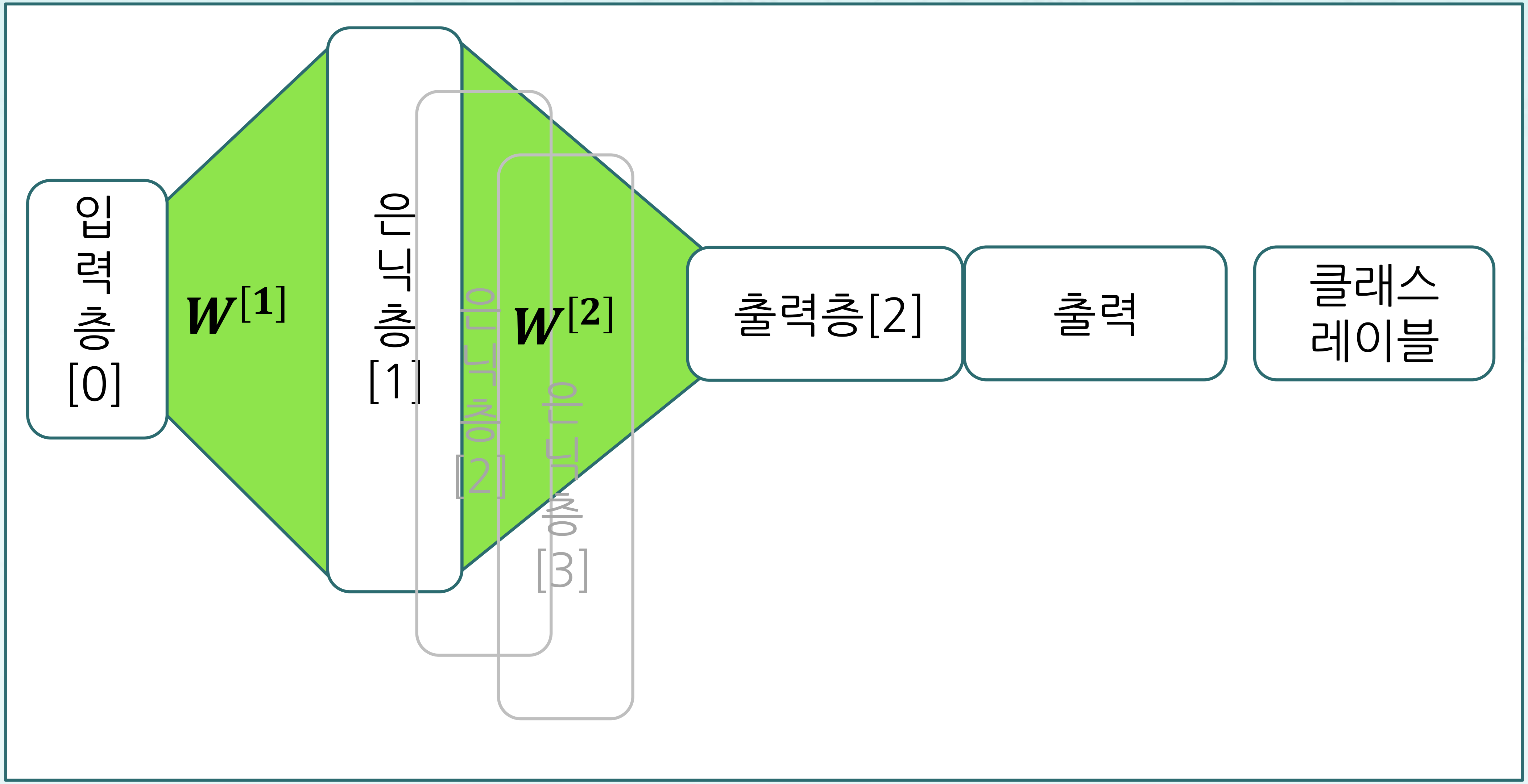
4. 다층 인공 신경망 행렬 모델: 다층 신경망의 구조



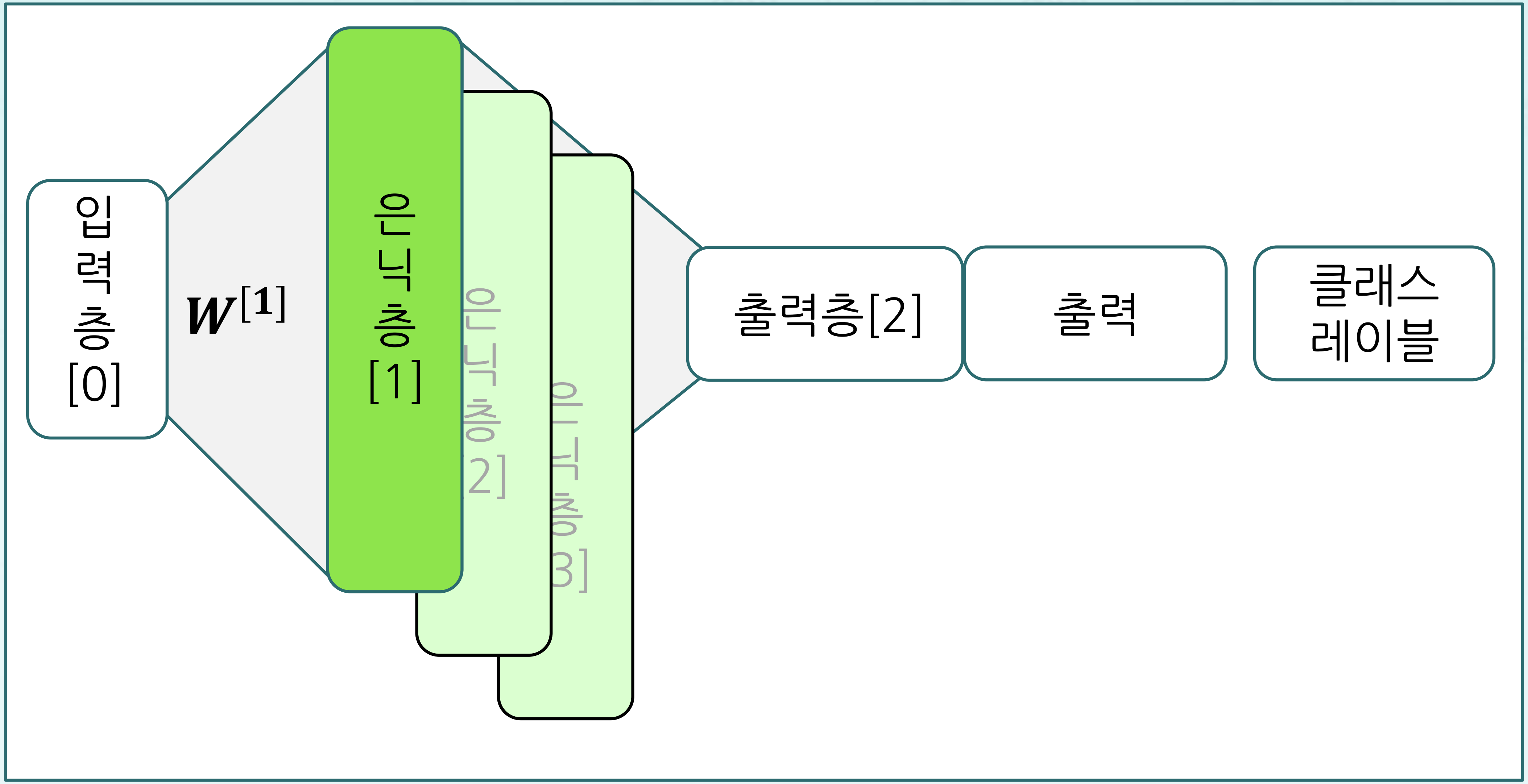
5. 다층 신경망의 구조: 입력층과 출력층



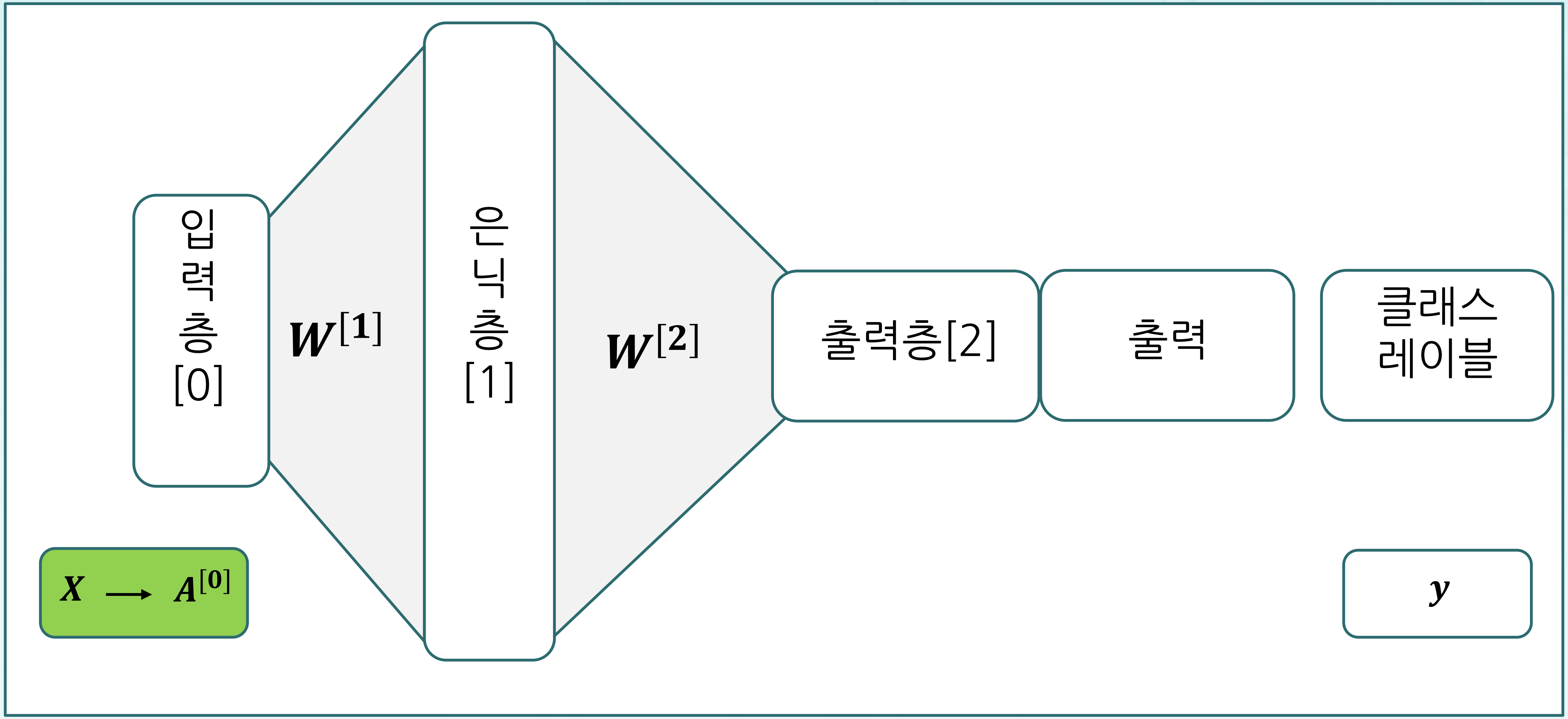
5. 다층 신경망의 구조: 가중치



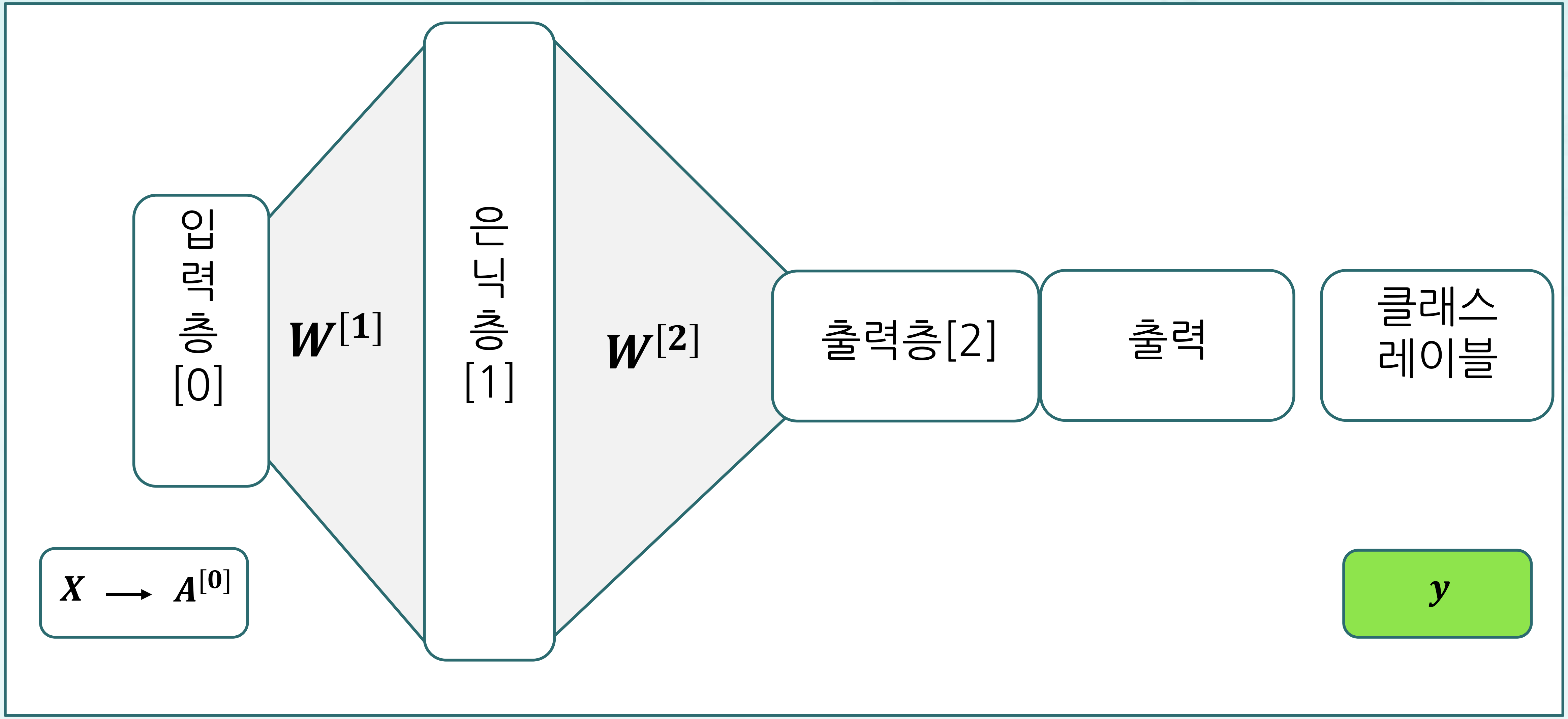
5. 다층 신경망의 구조: 은닉층



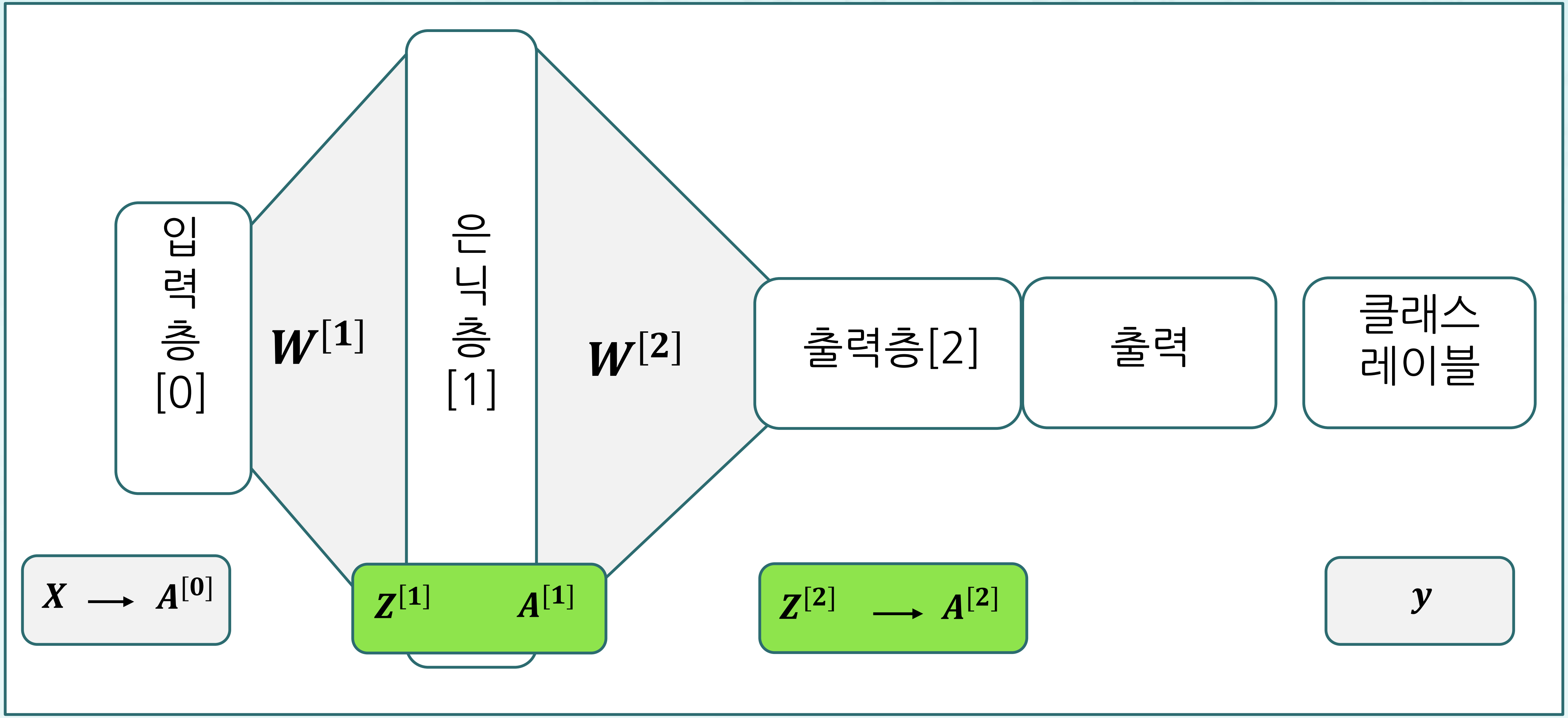
6. 다층 인공 신경망 행렬 모델: 입력



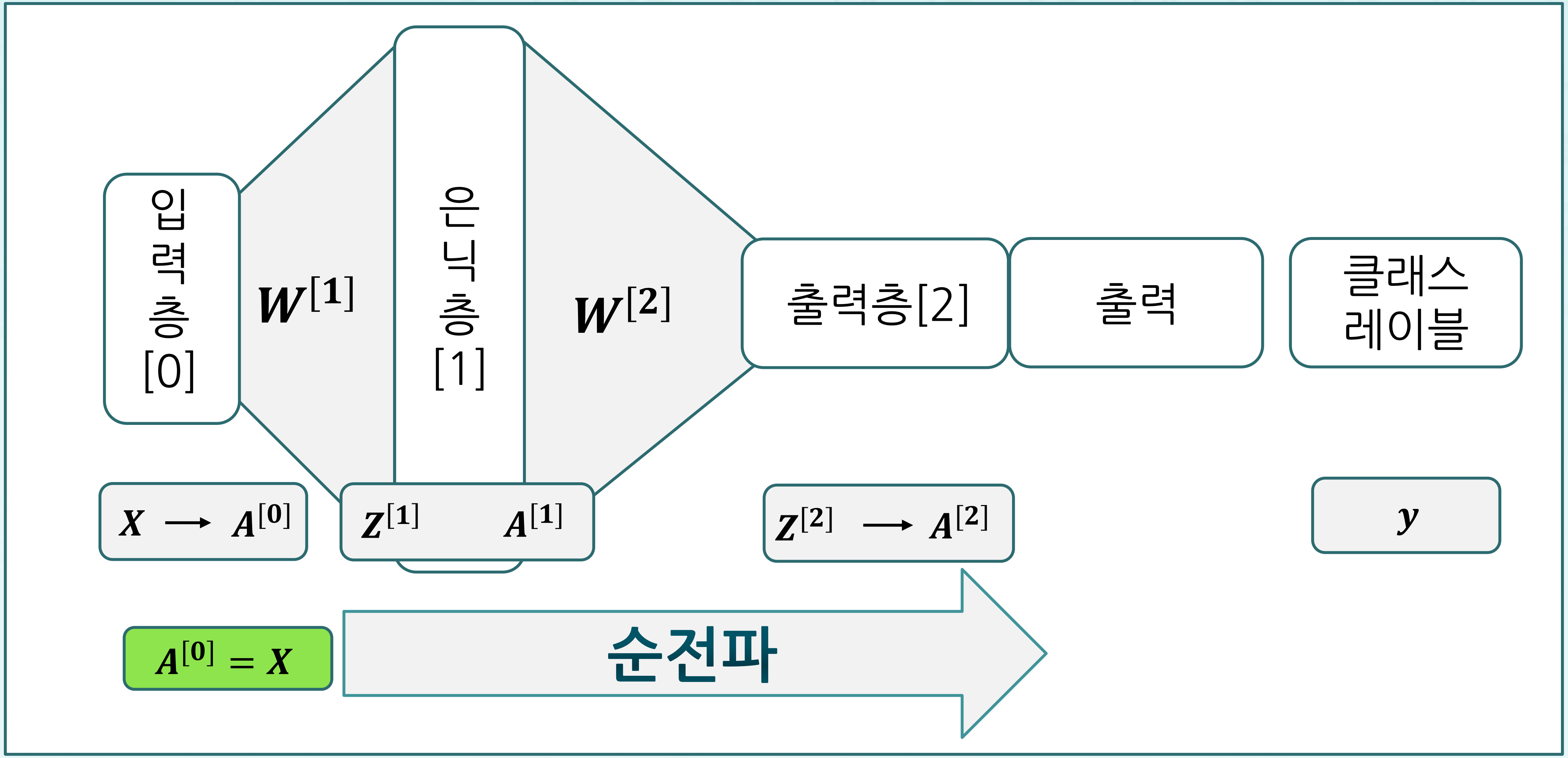
6. 다층 인공 신경망 행렬 모델: 레이블



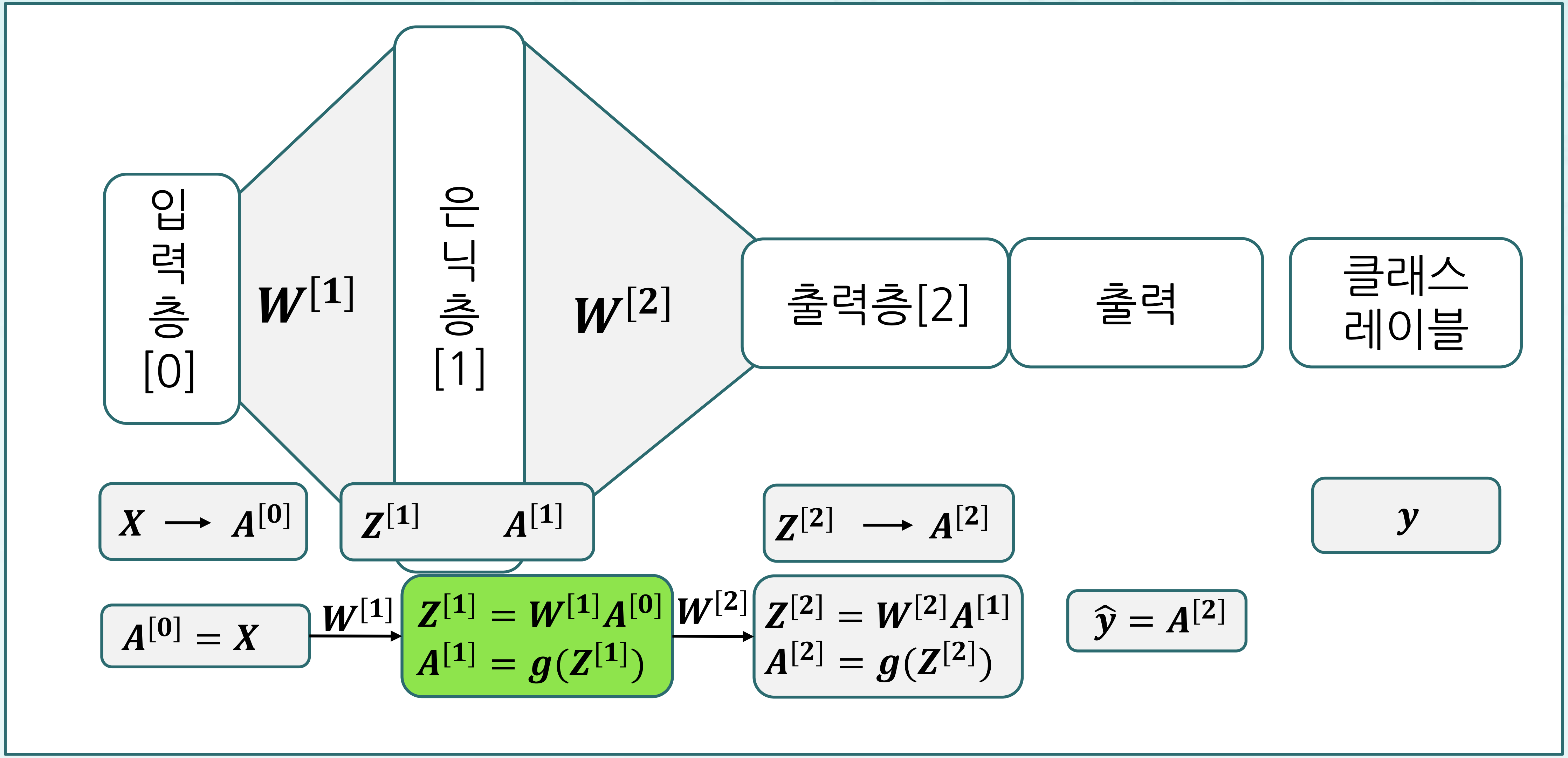
6. 다층 인공 신경망 행렬 모델: 입력과 출력



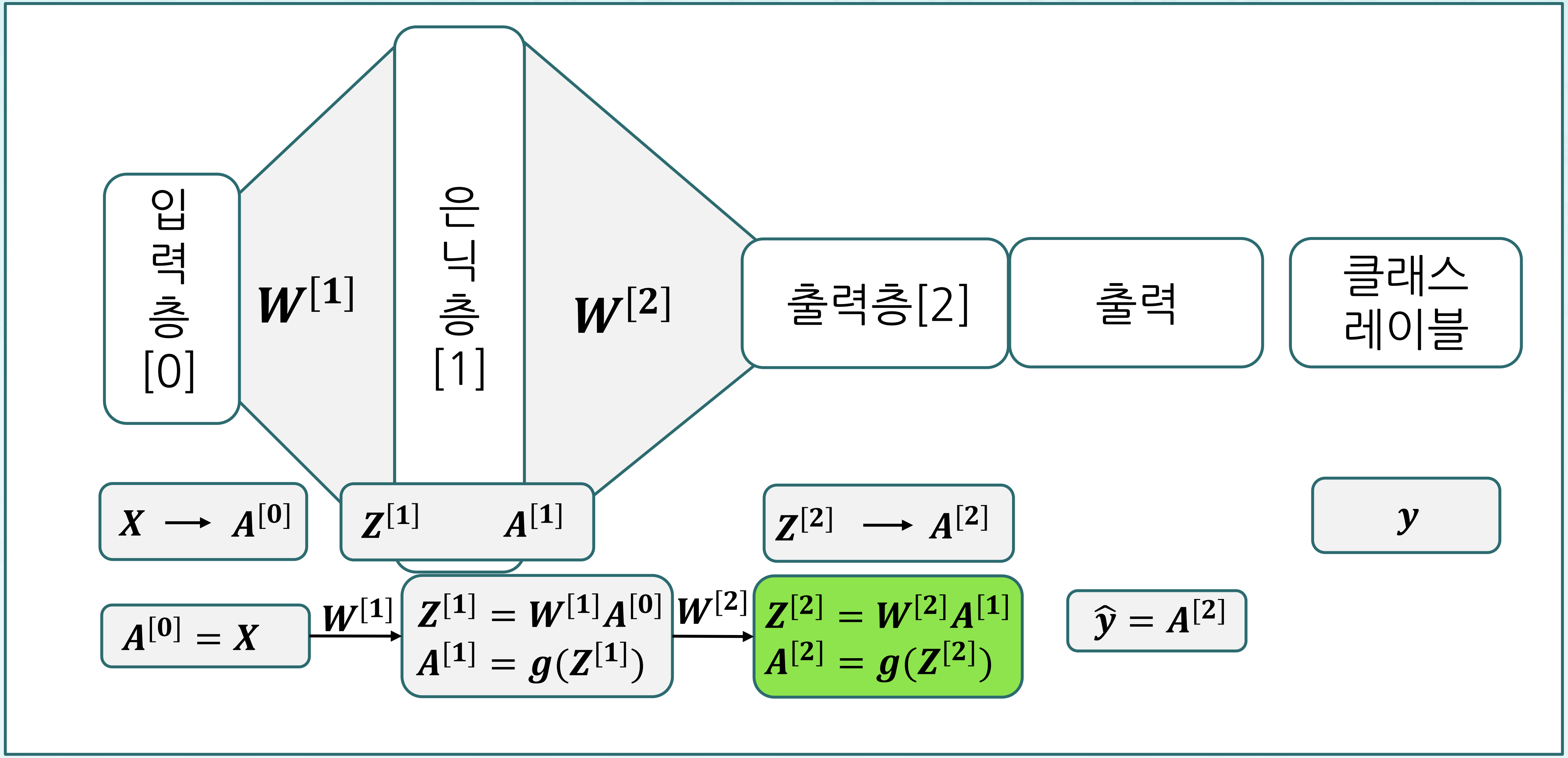
6. 다층 인공 신경망 행렬 모델: 순전파



7. 순전파: 은닉층 계산

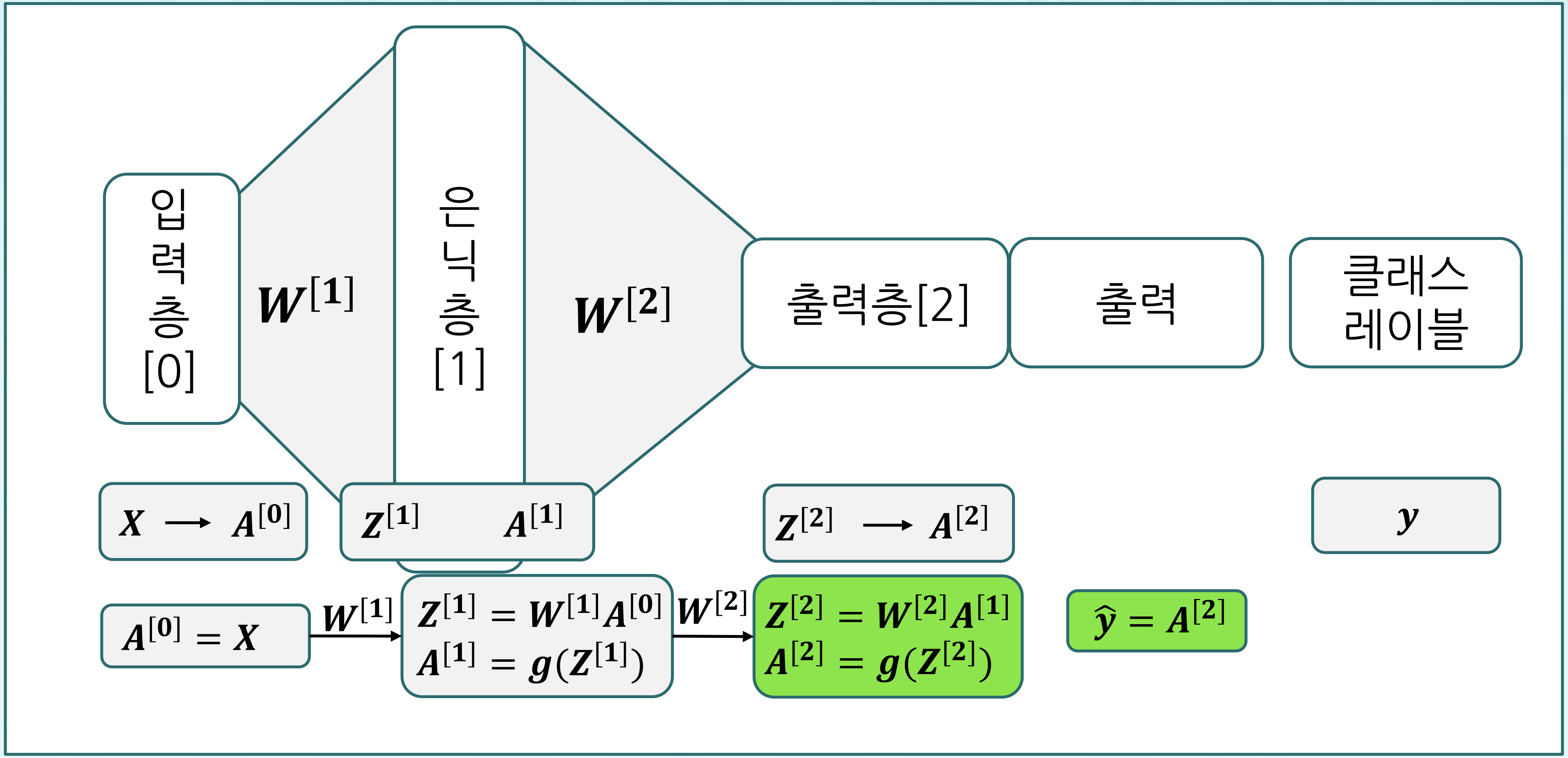

입력층[0]
$W^{[1]}$
은닉층[1]
$W^{[2]}$
출력층[2]
출력
클래스 레이블
$X \longrightarrow A^{[0]}$
$Z^{[1]}$ $A^{[1]}$
$Z^{[2]} \longrightarrow A^{[2]}$
y
$A^{[0]} = X$
$W^{[1]}$
$Z^{[1]} = W^{[1]}A^{[0]}$
$A^{[1]} = g(Z^{[1]})$
$W^{[2]}$
$Z^{[2]} = W^{[2]}A^{[1]}$
$A^{[2]} = g(Z^{[2]})$
$\hat{y} = A^{[2]}$

7. 순전파: 출력층 계산



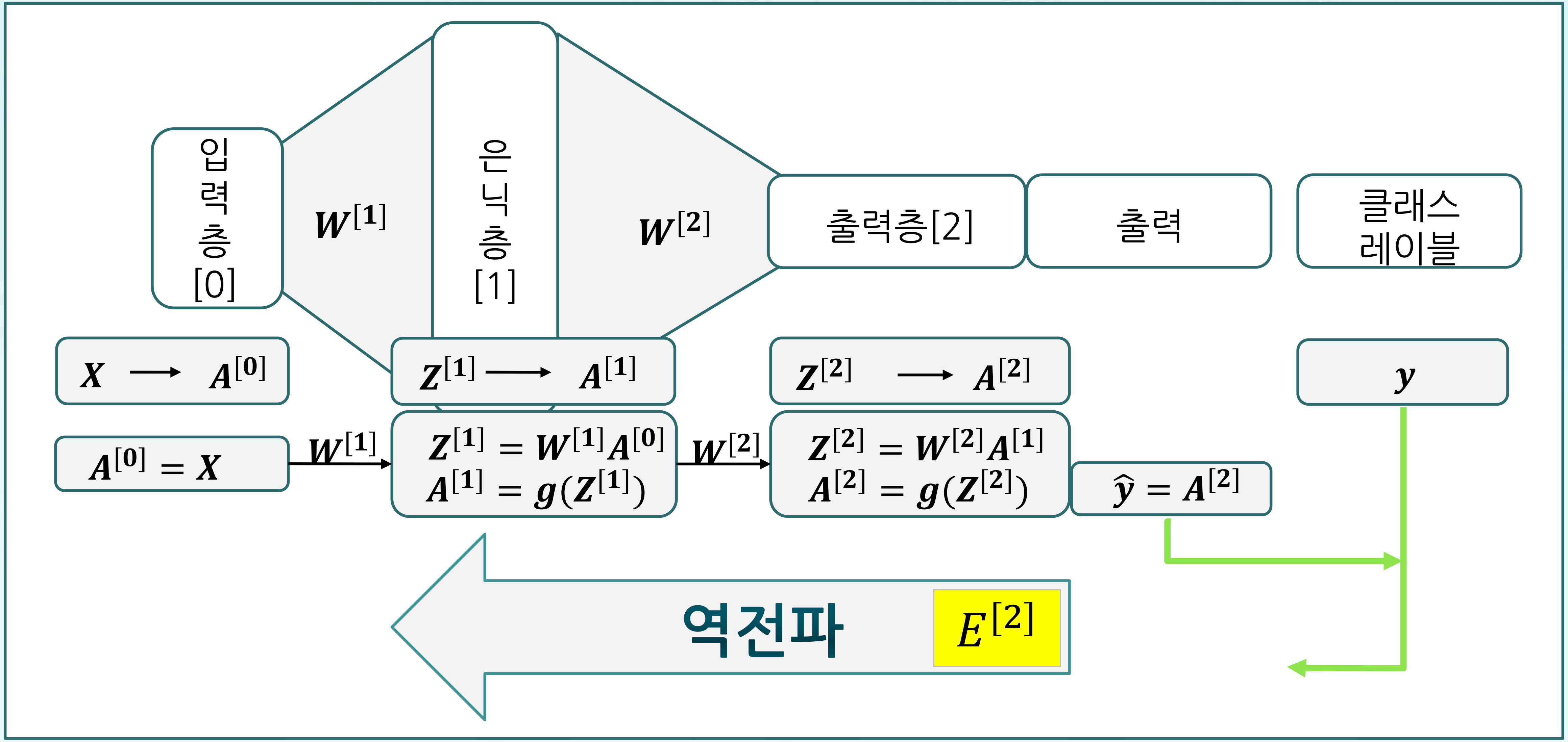
입력층[0]
$W^{[1]}$
은닉층[1]
$W^{[2]}$
출력층[2]
출력
클래스 레이블
$X \longrightarrow A^{[0]}$
$Z^{[1]} \quad A^{[1]}$
$Z^{[2]} \longrightarrow A^{[2]}$
y
$A^{[0]} = X$
$W^{[1]}$
$Z^{[1]} = W^{[1]}A^{[0]}$
$A^{[1]} = g(Z^{[1]})$
$W^{[2]}$
$Z^{[2]} = W^{[2]}A^{[1]}$
$A^{[2]} = g(Z^{[2]})$
$\hat{y} = A^{[2]}$

7. 순전파: 출력층 계산



입력층 [0]
$W^{[1]}$
은닉층 [1]
$W^{[2]}$
출력층[2]
출력
클래스 레이블
$X \longrightarrow A^{[0]}$
$Z^{[1]} \quad A^{[1]}$
$Z^{[2]} \longrightarrow A^{[2]}$
y
$A^{[0]} = X$
$W^{[1]}$
$Z^{[1]} = W^{[1]}A^{[0]}$
$A^{[1]} = g(Z^{[1]})$
$W^{[2]}$
$Z^{[2]} = W^{[2]}A^{[1]}$
$A^{[2]} = g(Z^{[2]})$
$\hat{y} = A^{[2]}$

8. 역전파: 출력층 계산


입력층[0]
$W^{[1]}$
은닉층[1]
$W^{[2]}$
출력층[2]
출력
클래스 레이블
$X \longrightarrow A^{[0]}$
$Z^{[1]} \longrightarrow A^{[1]}$
$Z^{[2]} \longrightarrow A^{[2]}$
y
$A^{[0]} = X$
$W^{[1]}$
$Z^{[1]} = W^{[1]}A^{[0]}$
$A^{[1]} = g(Z^{[1]})$
$W^{[2]}$
$Z^{[2]} = W^{[2]}A^{[1]}$
$A^{[2]} = g(Z^{[2]})$
$\hat{y} = A^{[2]}$
역전파
$E^{[2]}$

8. 역전파: 가중치 조정

$$W^{[2]} := W^{[2]} - \alpha \Delta W^{[2]}$$

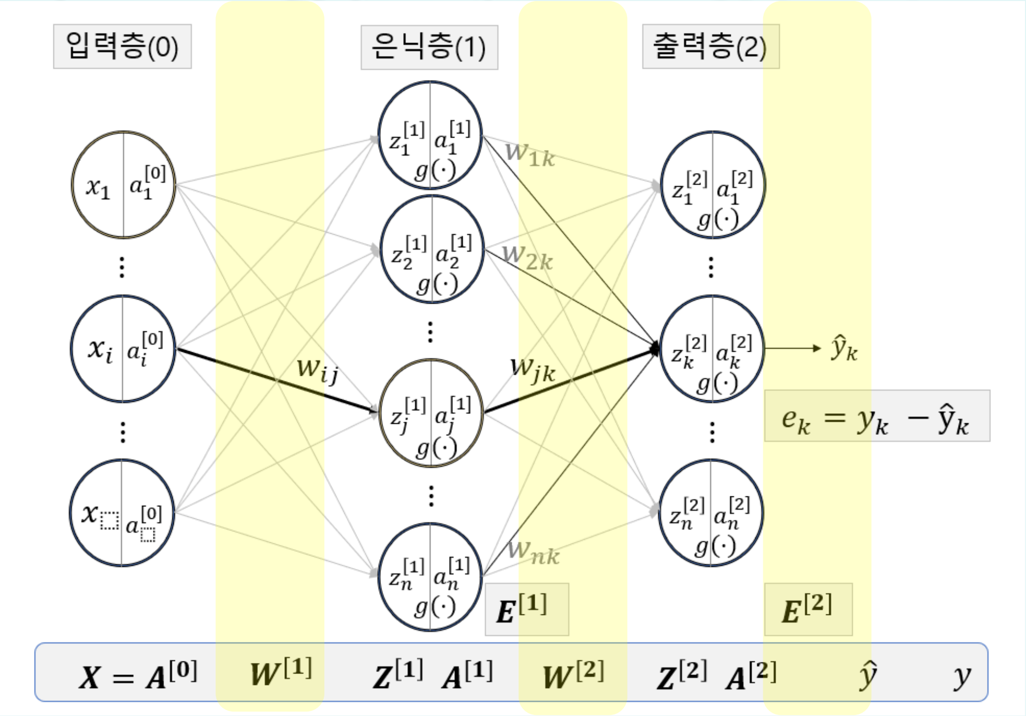
$$= W^{[2]} - \alpha \frac{\partial E}{\partial W^{[2]}}$$

$$= \mathrm{W}^{[2]} + \boxed{E^{[2]} \cdot g'(Z^{[2]}) \cdot A^{[1]T}}$$

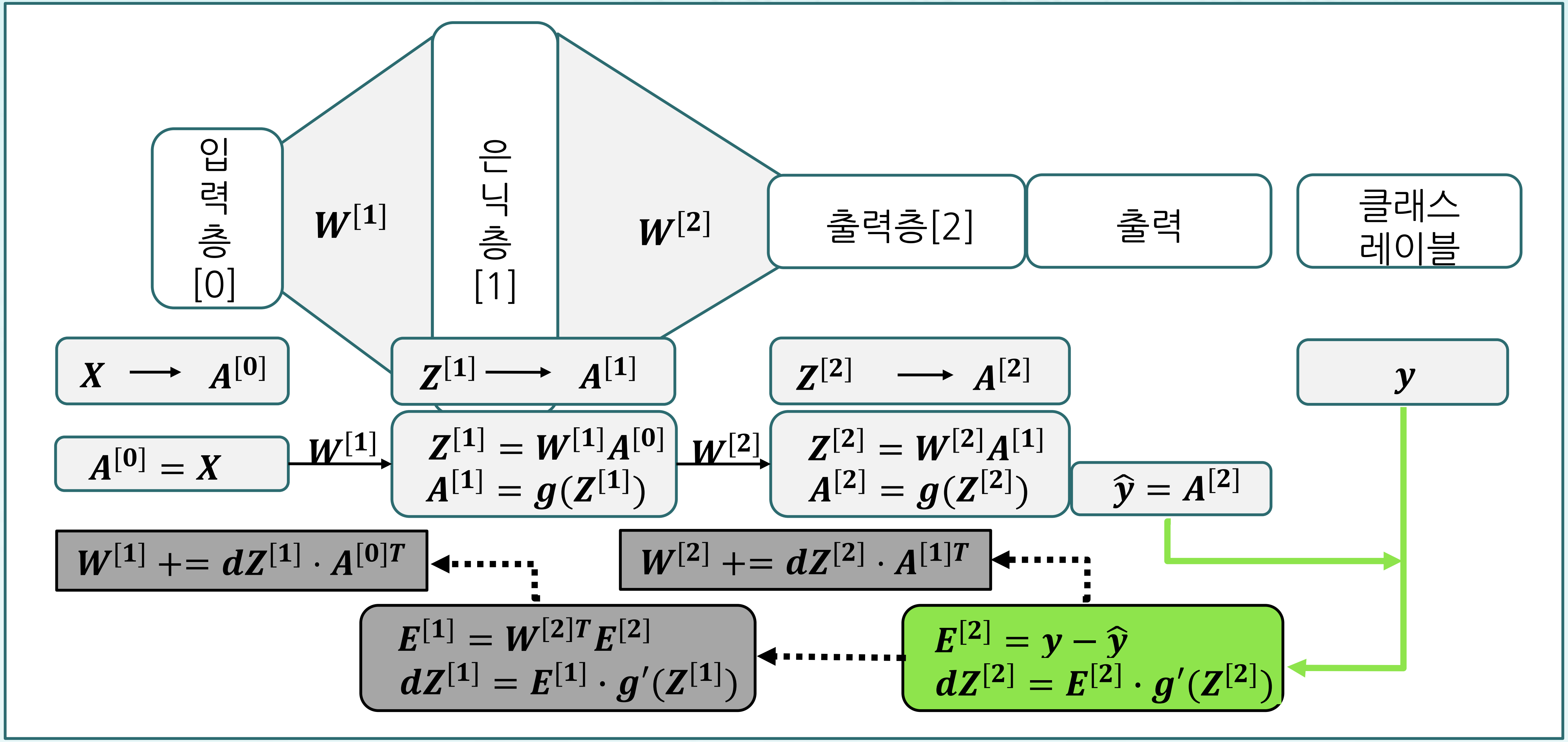
$$W^{[1]} := W^{[1]} - \alpha \Delta W^{[1]}$$

$$= W^{[1]} - \alpha \frac{\partial E}{\partial W^{[1]}}$$

$$= \mathrm{W}^{[1]} + \boxed{E^{[1]} \cdot g'(Z^{[1]}) \cdot A^{[0]T}}$$

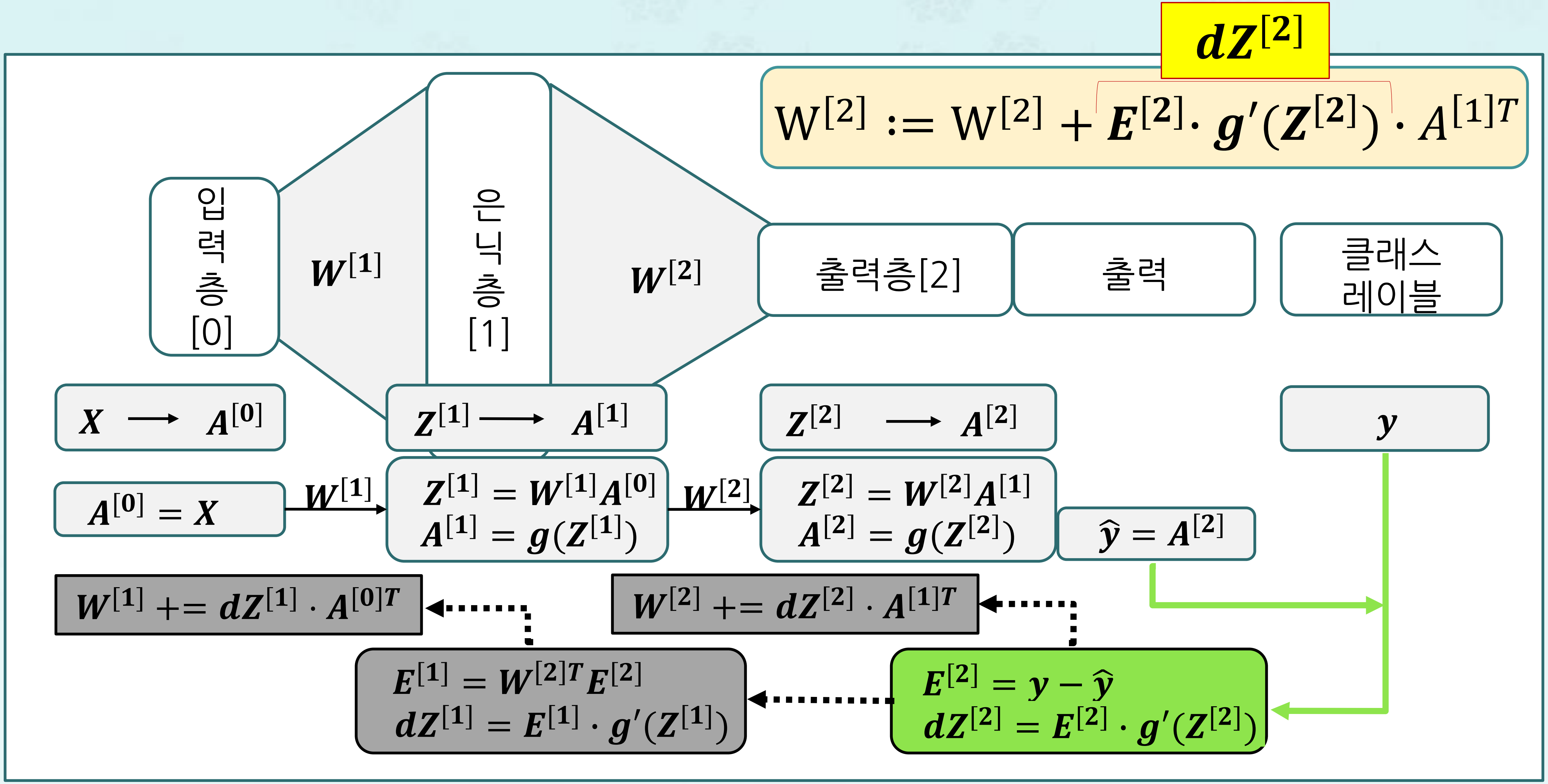


8. 역전파: E2 값 계산



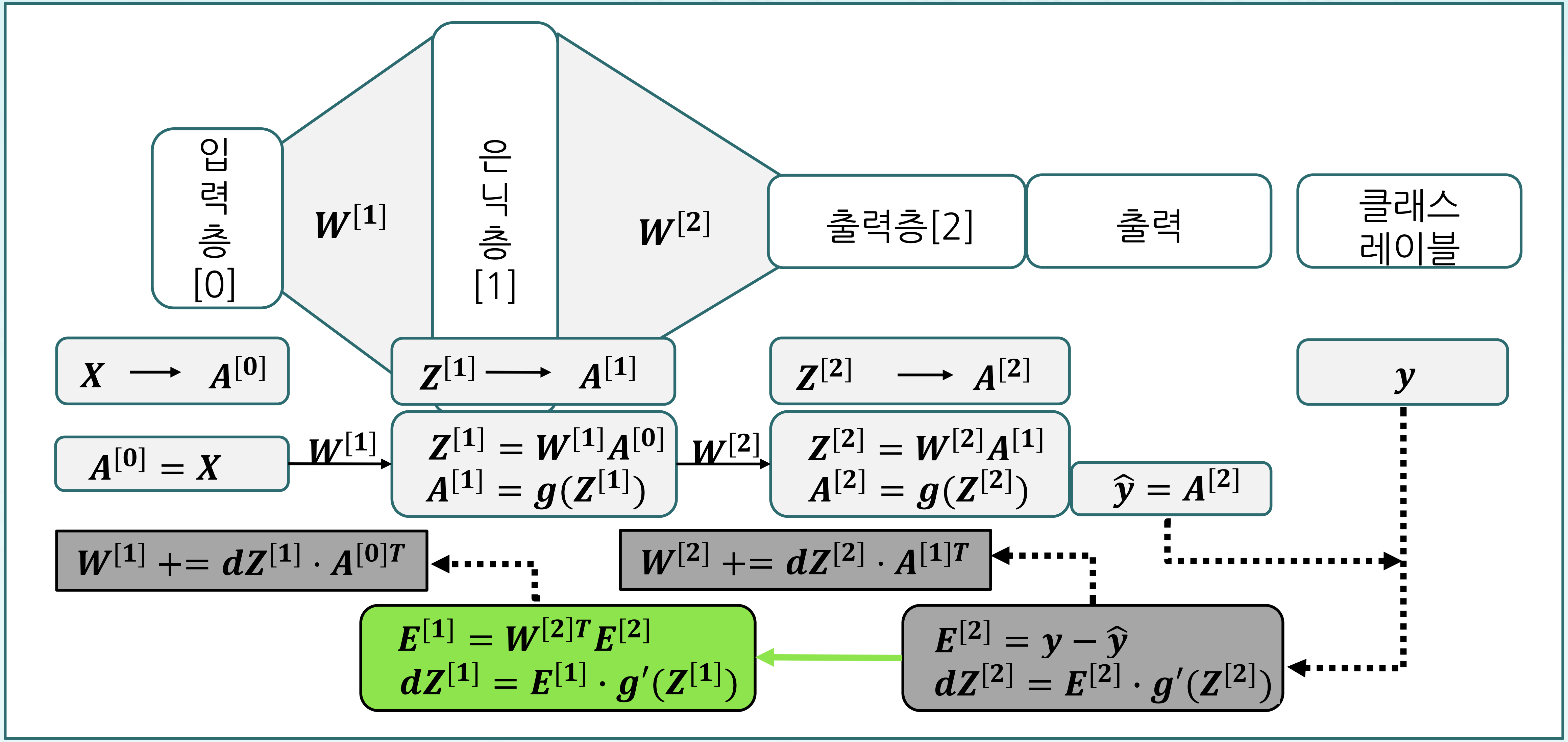
입력층[0]
$W^{[1]}$
은닉층[1]
$W^{[2]}$
출력층[2]
출력
클래스 레이블
$X \longrightarrow A^{[0]}$
$Z^{[1]} \longrightarrow A^{[1]}$
$Z^{[2]} \longrightarrow A^{[2]}$
y
$A^{[0]} = X$
$W^{[1]}$
$Z^{[1]} = W^{[1]}A^{[0]}$
$A^{[1]} = g(Z^{[1]})$
$W^{[2]}$
$Z^{[2]} = W^{[2]}A^{[1]}$
$A^{[2]} = g(Z^{[2]})$
$\hat{y} = A^{[2]}$
$W^{[1]} += dZ^{[1]} \cdot A^{[0]T}$
$W^{[2]} += dZ^{[2]} \cdot A^{[1]T}$
$E^{[1]} = W^{[2]T}E^{[2]}$
$dZ^{[1]} = E^{[1]} \cdot g'(Z^{[1]})$
$E^{[2]} = y - \hat{y}$
$dZ^{[2]} = E^{[2]} \cdot g'(Z^{[2]})$

8. 역전파: dZ2 값 계산



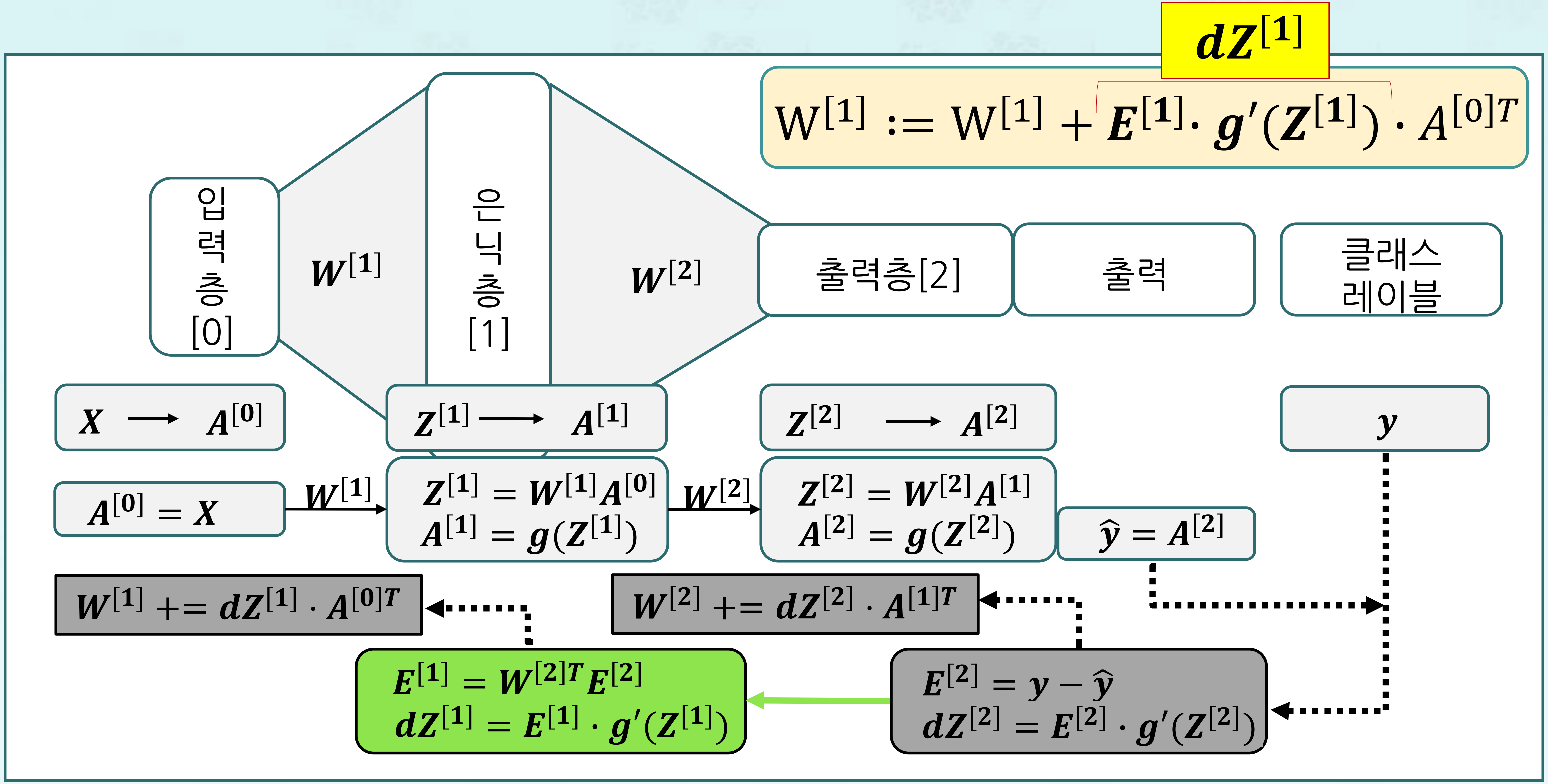
$dZ^{[2]}$
$W^{[2]} := W^{[2]} + E^{[2]} \cdot g'(Z^{[2]}) \cdot A^{[1]T}$
입력층[0]
$W^{[1]}$
은닉층[1]
$W^{[2]}$
출력층[2]
출력
클래스 레이블
$X \longrightarrow A^{[0]}$
$Z^{[1]} \longrightarrow A^{[1]}$
$Z^{[2]} \longrightarrow A^{[2]}$
y
$A^{[0]} = X$
$W^{[1]}$
$Z^{[1]} = W^{[1]}A^{[0]}$
$A^{[1]} = g(Z^{[1]})$
$W^{[2]}$
$Z^{[2]} = W^{[2]}A^{[1]}$
$A^{[2]} = g(Z^{[2]})$
$\hat{y} = A^{[2]}$
$W^{[1]} += dZ^{[1]} \cdot A^{[0]T}$
$W^{[2]} += dZ^{[2]} \cdot A^{[1]T}$
$E^{[1]} = W^{[2]T}E^{[2]}$
$dZ^{[1]} = E^{[1]} \cdot g'(Z^{[1]})$
$E^{[2]} = y - \hat{y}$
$dZ^{[2]} = E^{[2]} \cdot g'(Z^{[2]})$

8. 역전파: E1 값 계산



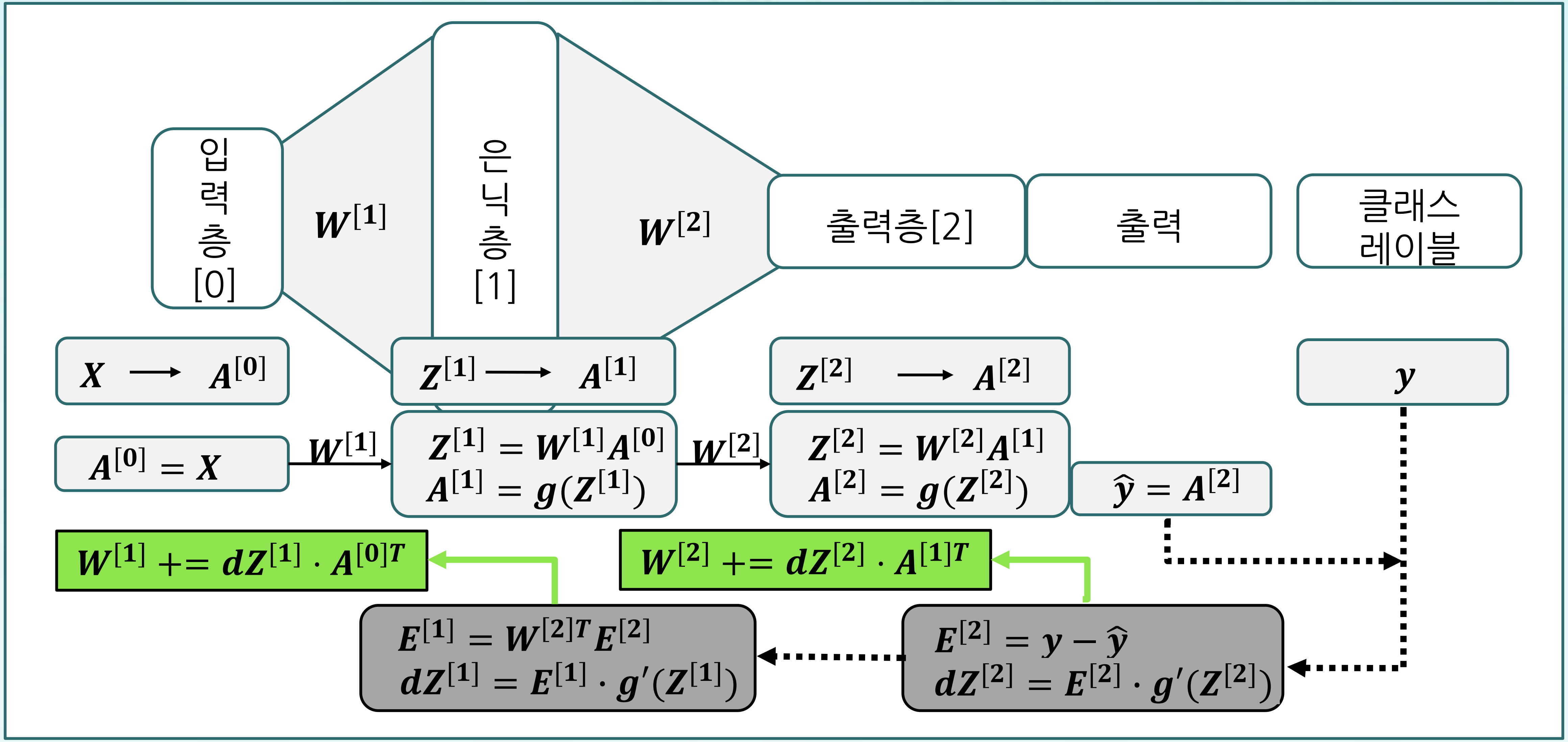
입력층[0]
$W^{[1]}$
은닉층[1]
$W^{[2]}$
출력층[2]
출력
클래스 레이블
$X \longrightarrow A^{[0]}$
$Z^{[1]} \longrightarrow A^{[1]}$
$Z^{[2]} \longrightarrow A^{[2]}$
y
$A^{[0]} = X$
$W^{[1]}$
$Z^{[1]} = W^{[1]}A^{[0]}$
$A^{[1]} = g(Z^{[1]})$
$W^{[2]}$
$Z^{[2]} = W^{[2]}A^{[1]}$
$A^{[2]} = g(Z^{[2]})$
$\hat{y} = A^{[2]}$
$W^{[1]} += dZ^{[1]} \cdot A^{[0]T}$
$W^{[2]} += dZ^{[2]} \cdot A^{[1]T}$
$E^{[1]} = W^{[2]T}E^{[2]}$
$dZ^{[1]} = E^{[1]} \cdot g'(Z^{[1]})$
$E^{[2]} = y - \hat{y}$
$dZ^{[2]} = E^{[2]} \cdot g'(Z^{[2]})$

8. 역전파: dZ1 값 계산



$dZ^{[1]}$
$W^{[1]} := W^{[1]} + E^{[1]} \cdot g'(Z^{[1]}) \cdot A^{[0]T}$
입력층[0]
$W^{[1]}$
은닉층[1]
$W^{[2]}$
출력층[2]
출력
클래스 레이블
$X \longrightarrow A^{[0]}$
$Z^{[1]} \longrightarrow A^{[1]}$
$Z^{[2]} \longrightarrow A^{[2]}$
y
$A^{[0]} = X$
$W^{[1]}$
$Z^{[1]} = W^{[1]} A^{[0]}$
$A^{[1]} = g(Z^{[1]})$
$W^{[2]}$
$Z^{[2]} = W^{[2]} A^{[1]}$
$A^{[2]} = g(Z^{[2]})$
$\hat{y} = A^{[2]}$
$W^{[1]} += dZ^{[1]} \cdot A^{[0]T}$
$W^{[2]} += dZ^{[2]} \cdot A^{[1]T}$
$E^{[1]} = W^{[2]T} E^{[2]}$
$dZ^{[1]} = E^{[1]} \cdot g'(Z^{[1]})$
$E^{[2]} = y - \hat{y}$
$dZ^{[2]} = E^{[2]} \cdot g'(Z^{[2]})$

8. 역전파: 가중치 조정


입력층[0]
$W^{[1]}$
은닉층[1]
$W^{[2]}$
출력층[2]
출력
클래스 레이블
$X \longrightarrow A^{[0]}$
$Z^{[1]} \longrightarrow A^{[1]}$
$Z^{[2]} \longrightarrow A^{[2]}$
y
$A^{[0]} = X$
$Z^{[1]} = W^{[1]}A^{[0]}$
$A^{[1]} = g(Z^{[1]})$
$Z^{[2]} = W^{[2]}A^{[1]}$
$A^{[2]} = g(Z^{[2]})$
$\hat{y} = A^{[2]}$
$W^{[1]} += dZ^{[1]} \cdot A^{[0]T}$
$W^{[2]} += dZ^{[2]} \cdot A^{[1]T}$
$E^{[1]} = W^{[2]T}E^{[2]}$
$dZ^{[1]} = E^{[1]} \cdot g'(Z^{[1]})$
$E^{[2]} = y - \hat{y}$
$dZ^{[2]} = E^{[2]} \cdot g'(Z^{[2]})$

다층 신경망 모델링

- 학습 정리
  - 미분에서의 연쇄법칙
  - 오차함수의 행렬 미분
  - 다층 인공 신경망의 행렬 모델링