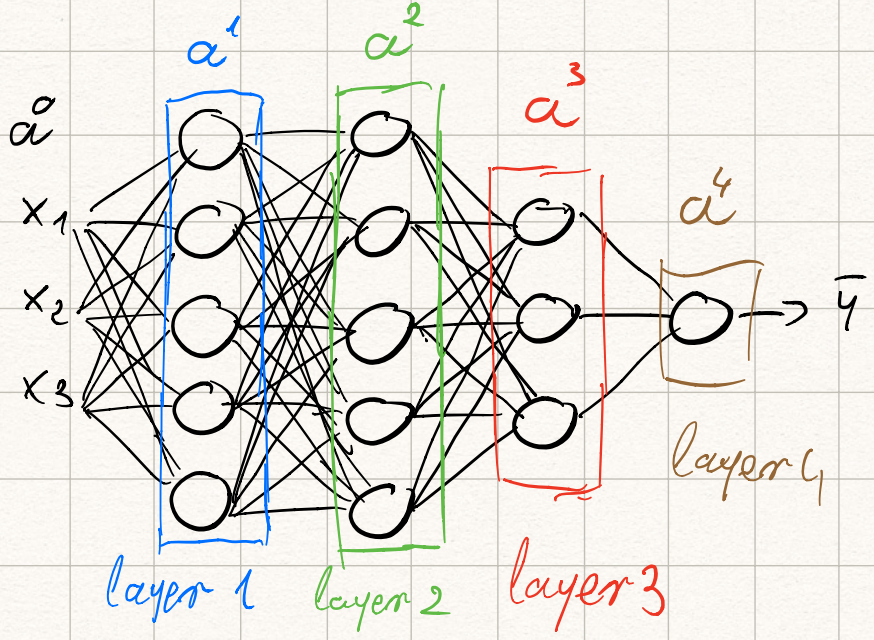
Forward propagation in deep networks



1) $z^1 = w^1 a^0 + b^1$
$a^1 = g^1(z^1)$

2) $z^2 = w^2 a^1 + b^2$
$a^2 = g^2(z^2)$

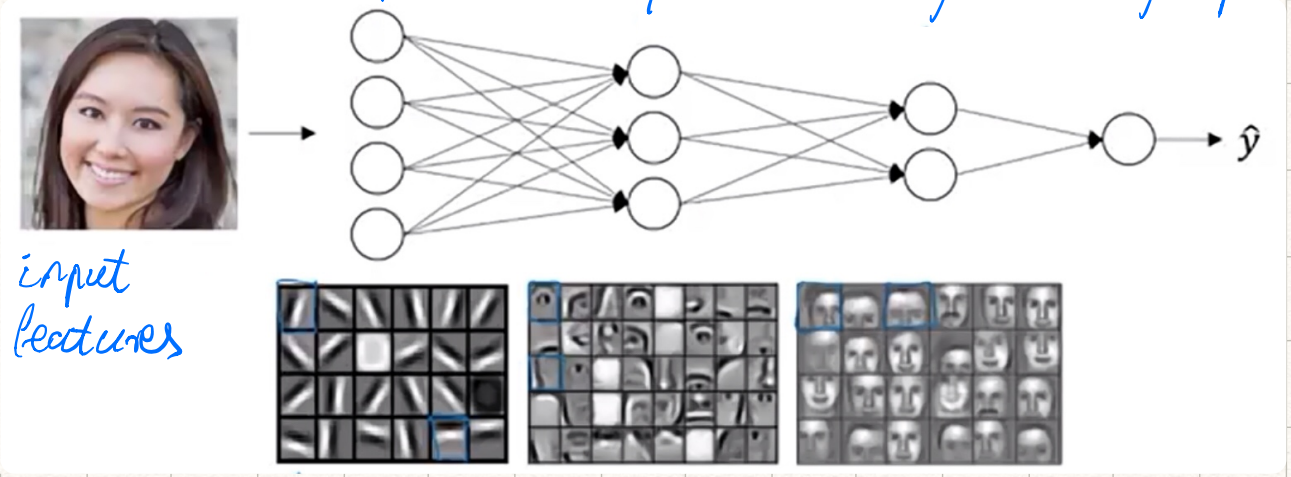
3) $z^3 = w^3 a^2 + b^3$
$a^3 = g^3(z^3)$

4) $z^4 = w^4 a^3 + b^4$
$a^4 = g^4(z^4)$

$$\Rightarrow \quad z^{[l]} = w^{[l]} \cdot a^{[l-1]} + b^{[l]}$$
$$a^{[l]} = g^{[l]}(z^{[l]})$$

Why deep representation

1)



2) Shallower networks require exponentially more hidden units

Building blocks of deep networks

Forward: Input $\Rightarrow a^{l-1}$, output $\Rightarrow a^l$, cache $\Rightarrow z^l$

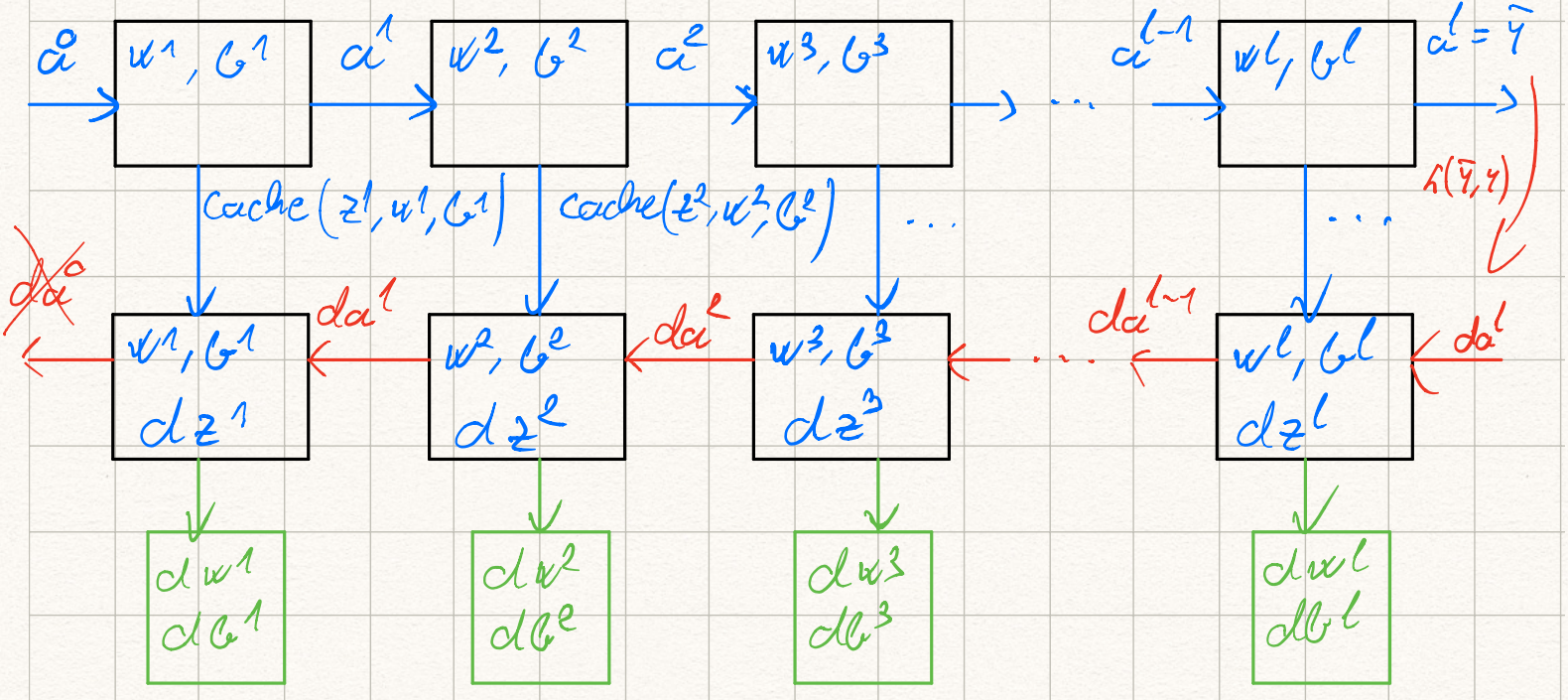
$$z^l = w^l a^{l-1} + b^l$$
$$a^l = f^l(z^l)$$

for layer l

Backward: Input $\Rightarrow da^l$, output $\Rightarrow da^{l-1}$, dw^l, db^l
cache $\Rightarrow z^l$

$$dz^l = w^{l+1T} dz^{l+1} * f^{l'}(z^l)$$
$$dw^l = dz^l \cdot a^{l-1}$$
$$db^l = dz^l$$
$$da^{l-1} = w^{lT} \cdot dz^l$$

for layer l



$$w^l := w^l - \alpha\, dw^l$$
$$b^l := b^l - \alpha\, db^l$$

Hyperparameters

Parameters: $W^{[l]}$, $b^{[l]}$

Hyperparameters:
- learning rate α
- # iterations
- # hidden layers
- # hidden units
- choice of activation function
- momentum
- mini-batch size
- regularization
- ⋮

Hyperparameters control $W^{[l]}$ and $b^{[l]}$.

Fine-tuning hyperparameters is an empirical process.