Apuntes sobre Adapters

$\{F_{ij}\} \subseteq \mathbb{R}^{3\times 3}$ son los parámetros del modelo pre-entrenado, mientras que $\{G_{ij}\} \subseteq \mathbb{R}^{3\times 3}$ son los del modelo fine-tuneado. Asumimos que el modelo pre-entrenado es tan bueno que existen coeficientes $\{\alpha_{ij}\} \subseteq \mathbb{R}$ tales que:

$$G_{ij} = \sum_{k=1}^{C} \alpha_{ik} F_{kj}, \quad \forall i,j.$$

De esta forma, si definimos

$$f(x) := F * x = \left[\sum_{j=1}^{C} F_{ij} * x_j \right]_{i=1}^{C},$$

$$g(x) := G * x = \left[\sum_{j=1}^{C} G_{ij} * x_j \right]_{i=1}^{C},$$

para un set de features $x = \{x_j\}$ con $x_j \in \mathbb{R}^{H\times W}$, entonces tomando $y = f(x)$ vemos que se cumple:

$$g(x) = \left[\sum_{j=1}^{C} \left(\sum_{k=1}^{C} \alpha_{ik} F_{kj} \right) * x_j \right]_{i=1}^{C} = \left[\sum_{j=1}^{C} \sum_{k=1}^{C} \alpha_{ik} (F_{kj} * x_j) \right]_{i=1}^{C} = \left[\sum_{k=1}^{C} \alpha_{ik} \sum_{j=1}^{C} F_{kj} * x_j \right]_{i=1}^{C}$$

$$= \left[\sum_{k=1}^{C} \alpha_{ik} y_k \right]_{i=1}^{C} = \left[\sum_{j=1}^{C} \alpha_{ij} y_j \right]_{i=1}^{C},$$

de modo que si abusamos de la notación y vemos los α_{ij} como filtros $\alpha_{ij} \in \mathbb{R}^{1\times 1}$, la igualdad anterior se reescribe como:

$$g(x) = \left[\sum_{j=1}^{C} \alpha_{ij} * y_j \right]_{i=1}^{C} = \alpha * y = \alpha * f(x).$$

A los adapters del tipo anterior Rebuffi los denominó serial adapters. Un año después proposo los parallel adapters.

Con la misma notación asumamos ahora que el modelo pre-entrenado es tan bueno que existen coeficientes $\{\alpha_{ij}\} \subseteq \mathbb{R}$ tales que:

$$G_{ij} = F_{ij} + \alpha_{ij} D$$

en donde $D \in \mathbb{R}^{3\times 3}$ es la matriz

$$D = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix}.$$

De lo anterior se sigue que:

$$g(x) = \left[\sum_{j=1}^{c} G_{ij} * x_j \right]_{i=1}^{c} = \left[\sum_{j=1}^{c} (F_{ij} + \alpha_{ij} D) * x_j \right]_{i=1}^{c} = f(x) + \left[\sum_{j=1}^{c} (\alpha_{ij} D) * x_j \right]_{i=1}^{c},$$

pero ya que $\alpha_{ij} D \in \mathbb{R}^{3\times 3}$ tiene solo la entrada del centro diferente a cero, se sigue que

$$(\alpha_{ij} D) * x_j = \alpha_{ij} * x_j,$$

viendo α_{ij} en la expresión del lado derecho como un filtro $\alpha_{ij} \in \mathbb{R}^{1\times 1}$, así que

$$g(x) = f(x) + \alpha * X.$$