This file is the supplementary file of the work "Zhang Y., Guo X., Leung H, et al. A Hierarchical Meta-Transfer Framework with Attention Mechanism for SAR Target Recognition ". It tends to explain that why the values are scaled by dividing by $\sqrt{d}$ in the scaled dot-product attention to avoid the gradient vanishing problem.

Firstly, the input with a large variance tends to push the softmax function into regions where it has extremely small gradients. The proof can be given as follows.
As we all know, the softmax function can be shown as follows.

$$y_i = \text{Softmax}(x_i) = \frac{e^{x_i}}{\sum_j e^{x_j}} \tag{1}$$

Then the gradient can be computed by

$$\frac{\partial y_i}{\partial x_j} = \begin{cases} y_i - y_i^2, i = j \\ -y_i y_j, i \neq j \end{cases} \tag{2}$$

If the input value x_i is much larger than the other values $\{x_j\}_{j \neq i}$, y_i is close to 1 and the gradient is close to 0. This situation tends to occur when the variance of the inputs is a big value.

Secondly, there exists the statistical property of dot product as follows.
Lemma. Assume that the components of $\boldsymbol{u}$ *and* $\boldsymbol{v}$ *are independent random variables with a mean of 0 and a variance of 1. Then their dot product* $z = \sum_{i=1}^{d} u_i v_i$ *has mean 0 and variance* d.

Proof. We denote $z_i = u_i v_i$ and then the mean and variance of $z = \sum_{i=1}^{d} u_i v_i$ can be given by

$$E\left(\sum_{i=1}^{d} z_i\right) = \sum_{i=1}^{d} E(z_i) = \sum_{i=1}^{d} E(u_i v_i) = \sum_{i=1}^{d} E(u_i) E(v_i) = 0 \tag{3}$$

$$D\left(\sum_{i=1}^{d} z_i\right) = \sum_{i=1}^{d} D(z_i) = d \tag{4}$$

where

$$\begin{aligned} D(z_i) = D(u_i v_i) &= E\left(u_i v_i - E(u_i v_i)\right)^2 \\ &= E\left(u_i^2 v_i^2\right) \\ &= E\left(u_i^2\right) E\left(v_i^2\right) \\ &= D(u_i) D(v_i) \\ &= 1 \end{aligned} \tag{5}$$

Therefore, for large values of d, the dot products grow large in magnitude, pushing the softmax function into regions where it has extremely small gradients. Therefore, the variance can be scaled to 1 by dividing by $\sqrt{d}$, which can avoid the gradient vanishing problem.