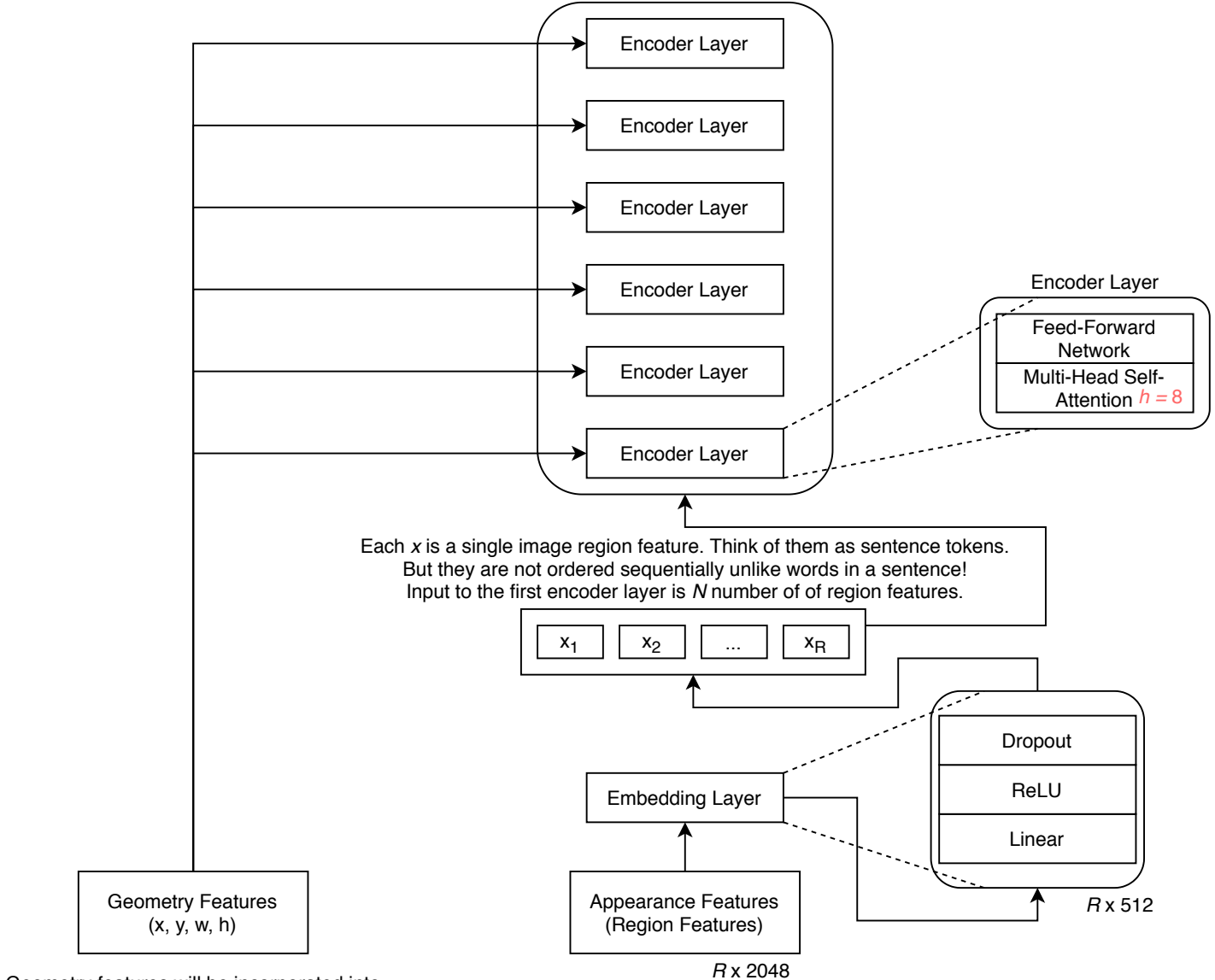

Geometry features will be incorporated into the attention weight matrix

d_{geom} = 4

Notations:

- ***R*** = number of detected regions per image
  - nowhere to found (not in the paper or in the code)
- d_{model} = 512, model inner dimension
- *h* = number of heads in self-attention

One Head

1. *X* contains all the input vectors (e.g. region features). Let R = 50.
Dimensions for all computations below: X: ***R* x 512,** W_q, W_k, W_v: **512 x 64**

$$Q = XW_Q, K = XW_K, V = XW_V$$

2. **Appearance attention weights**
are computed as follows:

$$\Omega_A = \frac{QK^T}{\sqrt{d_k}}$$

50 x 50

3. The next step is to compute **geometric attention weights.**

A displacement vector between
all variations of bounding boxes (*m, n*) among *R* is computed:

$$\lambda(m,n) = \left(\log\left(\frac{|x_m - x_n|}{w_m}\right), \log\left(\frac{|y_m - y_n|}{h_m}\right), \log\left(\frac{w_n}{w_m}\right), \log\left(\frac{h_n}{h_m}\right)\right) \quad *$$

Afterwards, this displacement vector is passed to compute weights:

$$\omega_G^{mn} = ReLU\left(\text{Emb}(\lambda)W_G\right) \quad **$$

4. **Combined attention weights** are computed as the next step:

$$\omega^{mn} = \frac{\omega_G^{mn}\exp(\omega_A^{mn})}{\sum_{l=1}^{N}\omega_G^{ml}\exp(\omega_A^{ml})} \quad ***$$

5. The output of the head is computed as the following:

$$\text{head}(X) = \text{self-attention}(Q, K, V) = \Omega V$$

Omega size: R x R

where every element in Omega is provided by combined attention weights

* For more information, check the slides.

** Check BoxRelationalEmbedding function in model implementation

*** Check box_attention function in model implementation