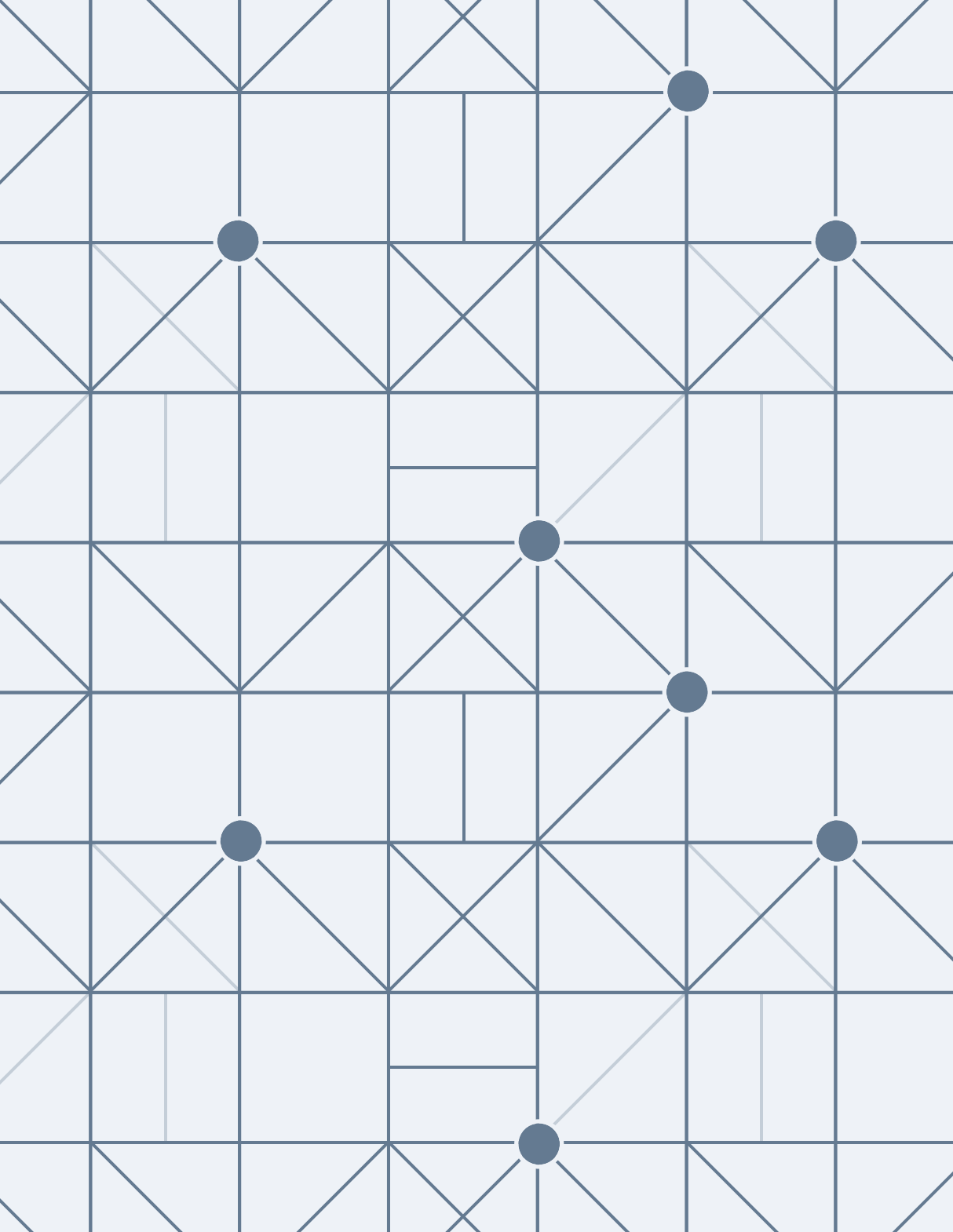

Assignment - 2

1. A single perceptron can create linear decision boundary only, but XOR is not linearly separable.

   with multi layer perceptrons, non-linear activations allows MLPs to warp the input space, enabling them to solve non-linear problems like XOR.

Q.2 Linear layers stacked:
product of 2 linear matrices is another linear matrix, stacking them adds no extra complexity.

→ In deep networks, gradients are multiplied repeatedly during back propogation; if weights are small, these gradients vanish.

→ ReLU vs sigmoid:-
Sigmoid saturates at 0 or 1 where its derivative is near zero, while ReLU has a constant gradient of 1 for all +ve values, preventing vanishing.

Q.3 Transformers process all words simultaneously and do not have order, positional Encoding instills sequence information so the model knows word positions.

- Sinusoidal vs Absolute:- Absolute PE assigns a fixed vector to each index, while sinusoidal uses sine/cosine waves to allow the model to generalize to unseen sequence lengths.

- Why RoPE helps: Rotary Embeddings use rotation matrices to capture relative distances between tokens, making it much more efficient for very long contexts.

Q.4 - Query is what we are looking for, key is the label of what we have and value is the actual information we extract.

- scaling by $\sqrt{d_k}$ prevents dot products from growing too large, which would otherwise push the softmax into flat regions with zero gradients.
- highest diagonal values represent a word's attention to itself, because a word is most similar to itself, its dot pdt is the highest.

Q.5 Splitting into heads allows the model to focus simultaneously on different types of relationships.

d_{model} → total embedding size.
h → no. of heads.
d_{head} → dimension of each individual head (d_{model}/h)

Q.6 - Greedy only picks the single best next word, while Beam search tracks multiple high-probability paths to find a better overall sequence.

- In translation, Greedy might pick a common word first that makes the rest of the sentence mathematically impossible or grammatically incorrect.

SOLVING:-

Q.1 $d_{head} = 768/12 = 64$

no. of parameters $= 3 \times 768 \times 768 = 1769472$

Q.2 Softmax :- $e^2 = 7.39$, $e^1 = 2.72$, $e^0 = 1$

Sum $= 7.39 + 2.72 + 1 = 11.11$

Divide $\Rightarrow [0.665, 0.245, 0.090]$