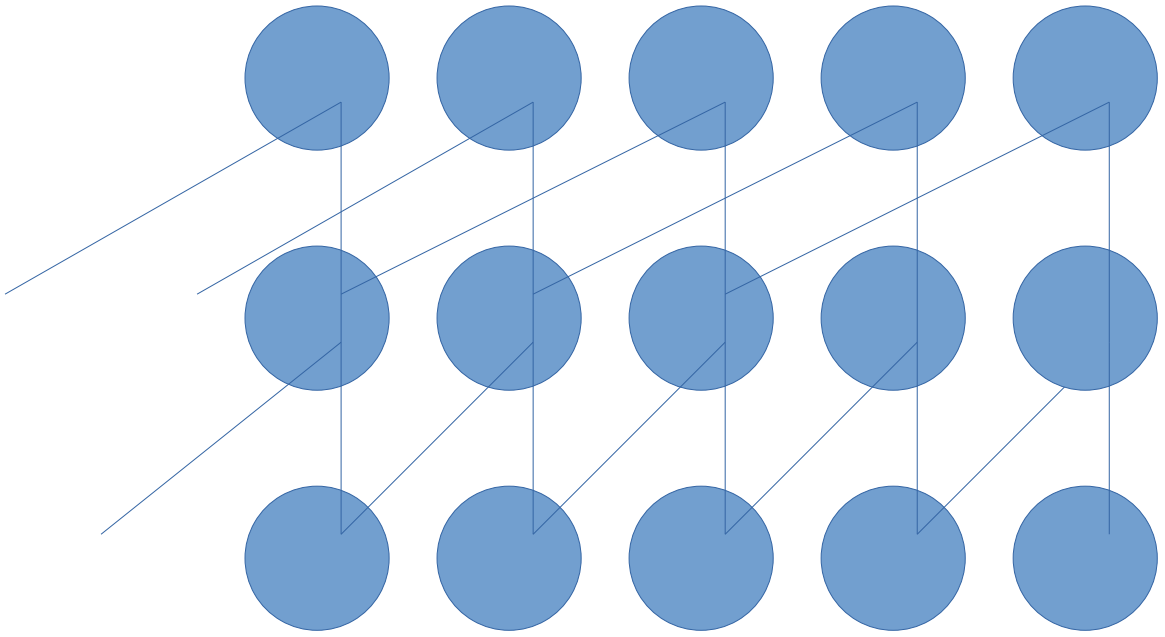
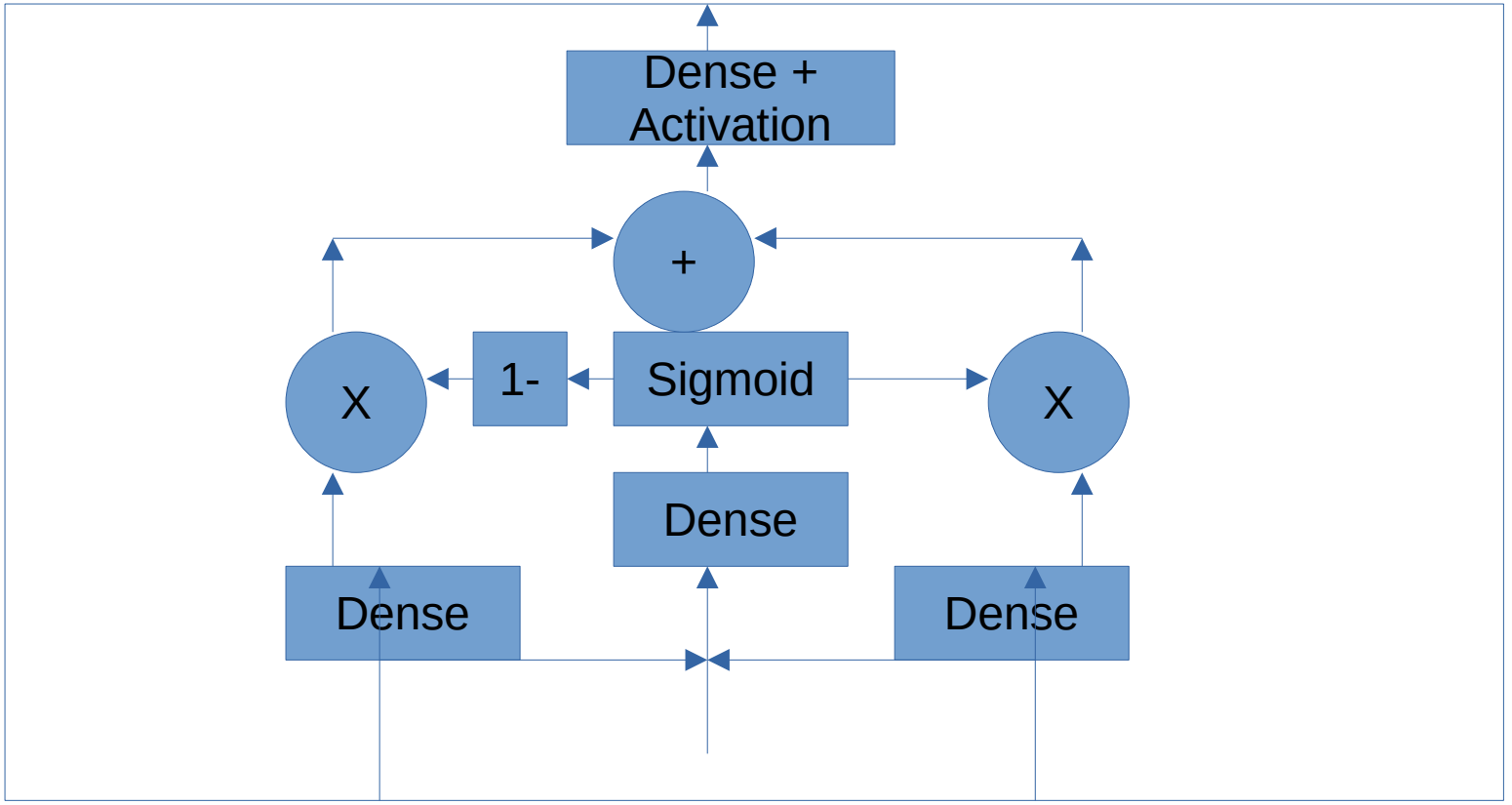


Binary attention each connection.



- If the left side is empty, set the sigmoid result to 1, meaning attending fully. If the right side is empty, the sigmoid result to 0.
- Go both directions, use this another set of binary attention

  block to weight the result’s value. Now, it means that

  all tokens have attended both forward and backward!
- Share weight at each layer. Do this until all tokens are covered by the last time step.
- Repeat this process N times.