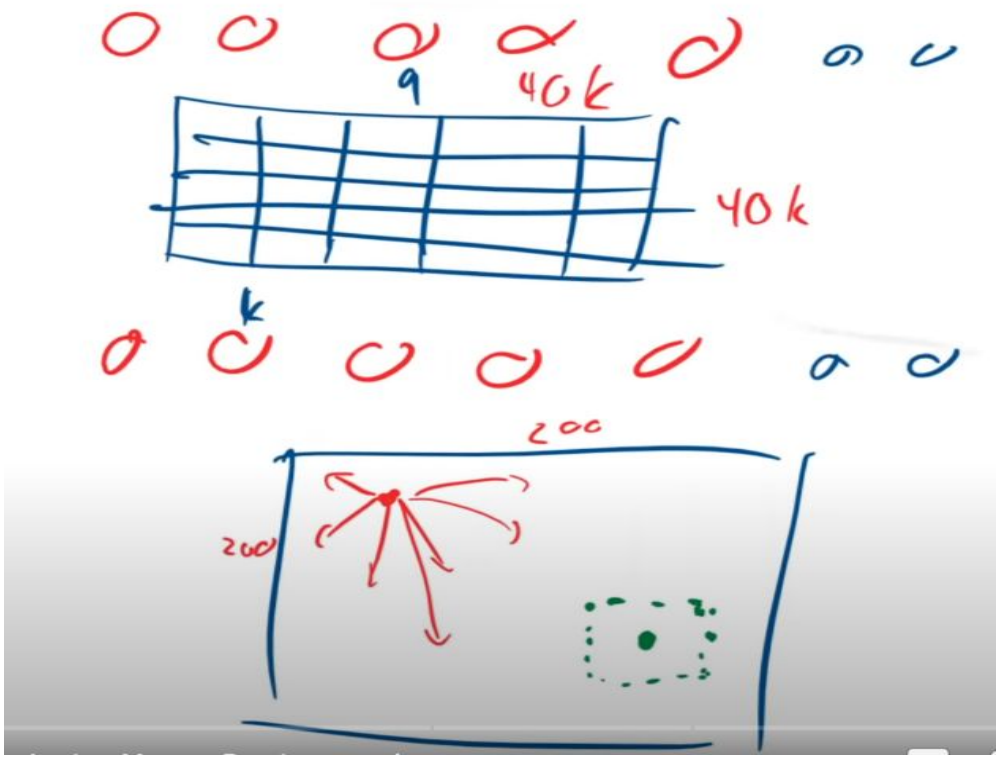
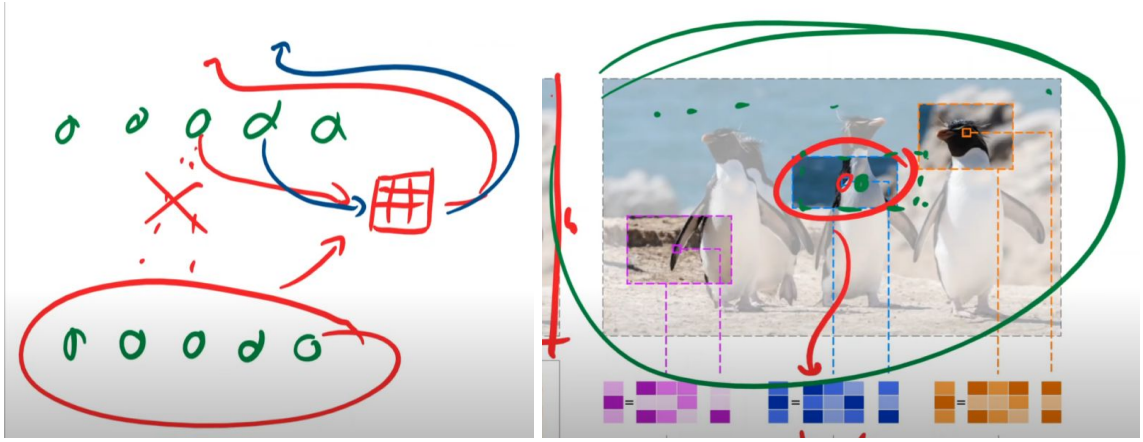
- Lambda layers capture contextual info of some input without using expensive attention maps and can process any sequence length.
- Attention layer can be seen as a generalised version of FFN. For High Res images, local attention was used (due to memory constraints) which is a generalised version of convolution filter.

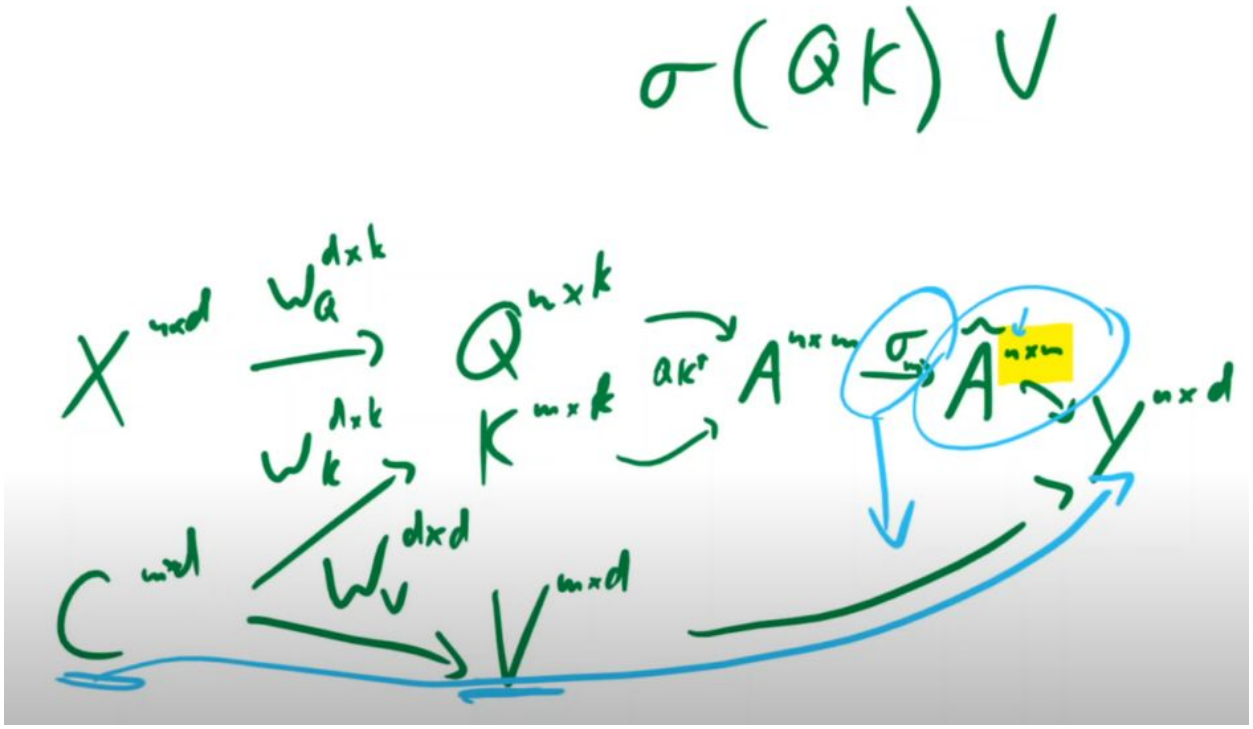


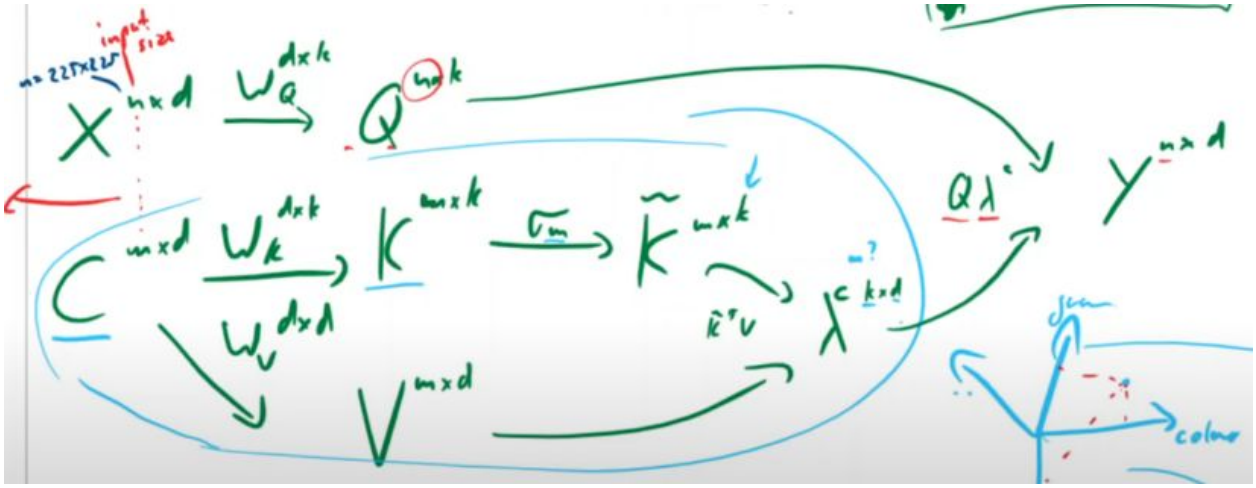
- Local context is transformed into a **linear** func which is independent of input (equivalent to matrix mul) which is applied on input pixel to get final representation. But in attention, each input pixel was multiplied with each context pixel.
    - In attention, direct mul with all context pixels, but in lambda, different func for diff pixels.

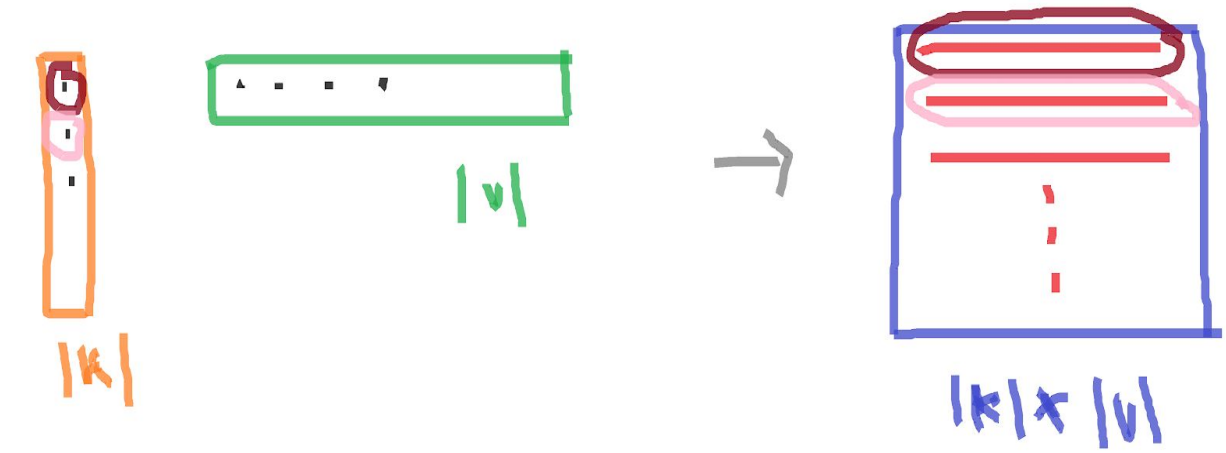


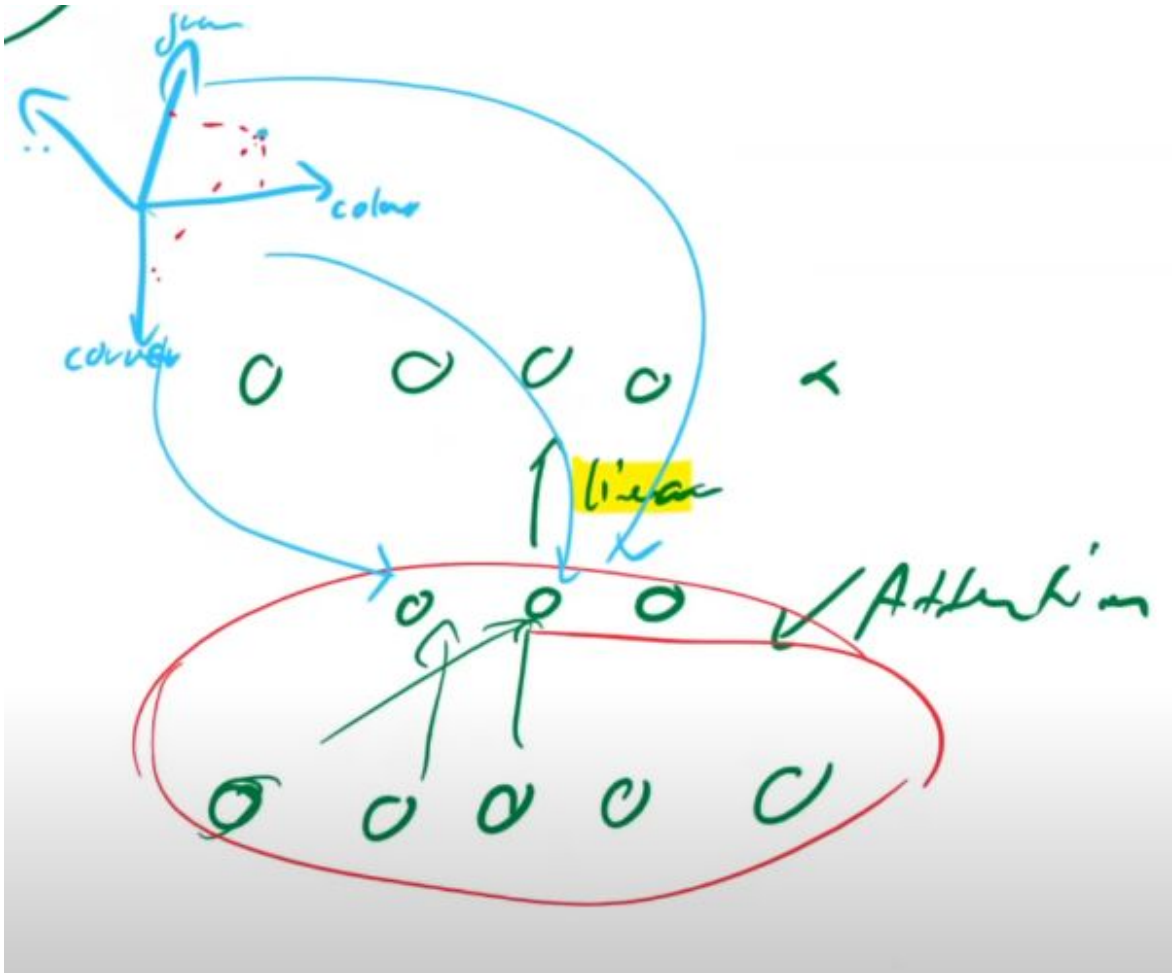
- Interactions : Content based => q_n and c_m . Position based => q_n and (n,m).

- The W_K matrix is like learned queries. They are multiplied with context to get K. Softmax is applied over m dim of K matrix to get k softmax distributions. These are used to give weightage to each value vector for that dimension (intuitively property).
- |k| is a vector space of properties of the context. This info about properties is aggregated to get summary. In attention query decides importance of a context pixel for each property but here importance of each context pixel for a property is fixed.
- The m dim is contracted to get k x d summary of context (content lambda). Similarly position lambda is calculated for each query. Final lambda = context lambda + position lambda.

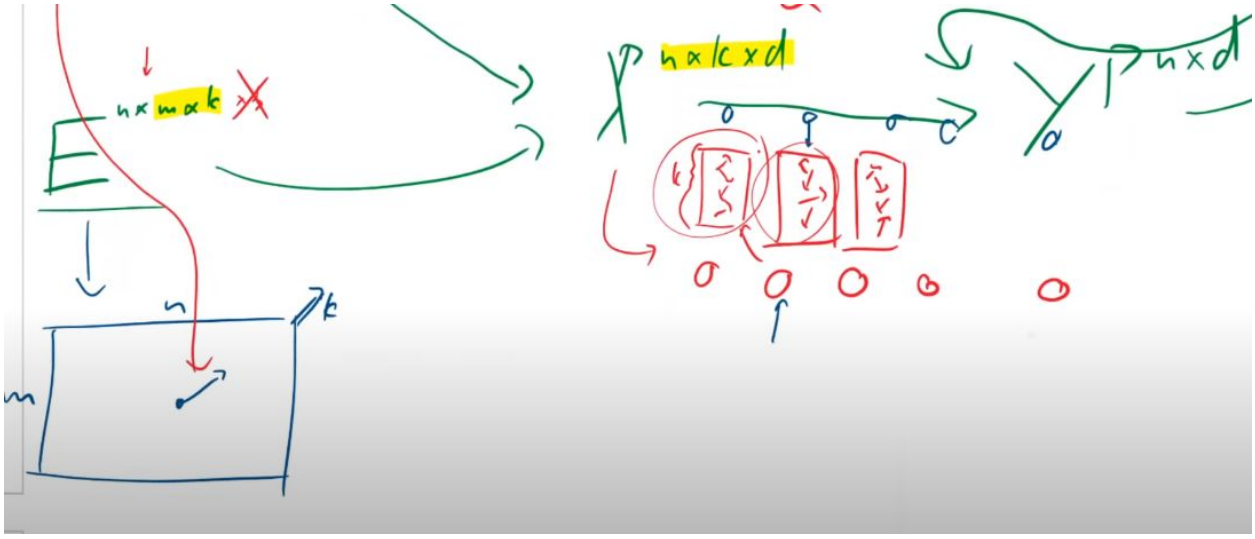





|V|
|K|
|K| x |V|


linear

- Positional Embeddings are fixed (independent of query) and aggregate info from context in a diff way for each query position. This gives unique info for each query.



- Content lambda is permutation invariant w.r.t context pixels as it does not depend on exact positions. Batch Normalisation applied after queries and values calculation was helpful.
- The inductive bias of a learning algorithm is the set of assumptions that the learner uses to predict outputs of given inputs that it has not encountered.
- Transitional equivariance is maintained by taking relative positional embeddings. https://towardsdatascience.com/translational-invariance-vs-translational-equivariance-f9fbc8fca63a