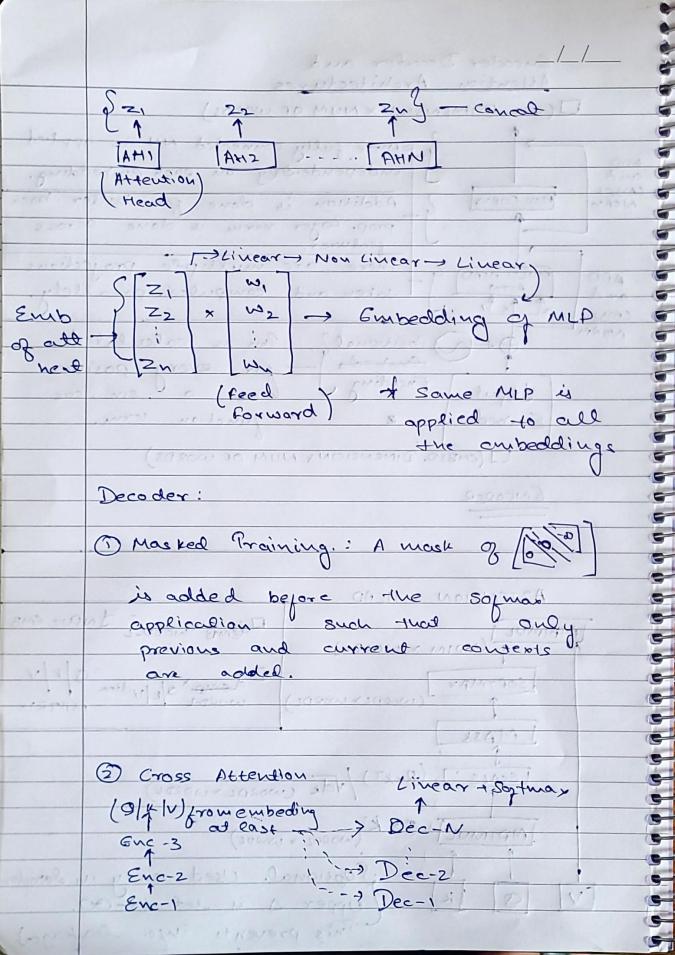
Encoder Decoder and Attention Architectures DIVAL DIMENSION X NUM OF WORDS) 2- Same fully connected NN is applied OPA independently on each embedding and CAYET FEED PORWA Addition is done NONV prop. Layer norm is done There are multiple projections ADD -taken and and CAYER NORM Phis is calculated Positional? Embeds every position Juput Embeds [(EMBED. DIMENSIONX NUM OF WORDS) ENCODER ATTENTION HEAD (VDIM & WORDL) dono DE Vam = 9 K/V SOFTMAN (WORPLX WORDL) MATRIX MASK (BXKT) / Volk (WORDL > WORDL) SCALS 1--- Saptional. Used only in decoder A is set to-00. This prevent info. leakages}



200 Junes Linear and of black Add Norn Feod I Add Norak 91 m 1 v 3 decouly, no embeddings are Encoder Add Norw required. Attu 6. Attention with mark of oill forward items St that Output Emb (Right Shifted) + Pas encoding