Attention (P-8) (LLM notes)

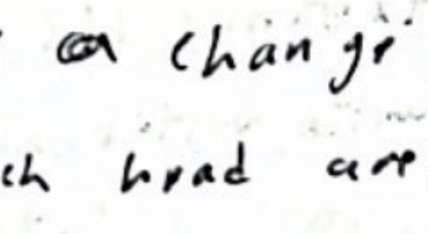
- Multi headed attention. Call ran in parrallell

• in the ex's from P1-P7, Like the Ex of adjectives updating nouns is one single head of attention, but there's many more heads needed to add context Like for ex updating words that are spelled the same but have different meaning, so we do what we did in single headed attention 10'000 times giving LLM's the level of context they really need.

Ex a upside-down fluffy blue creature

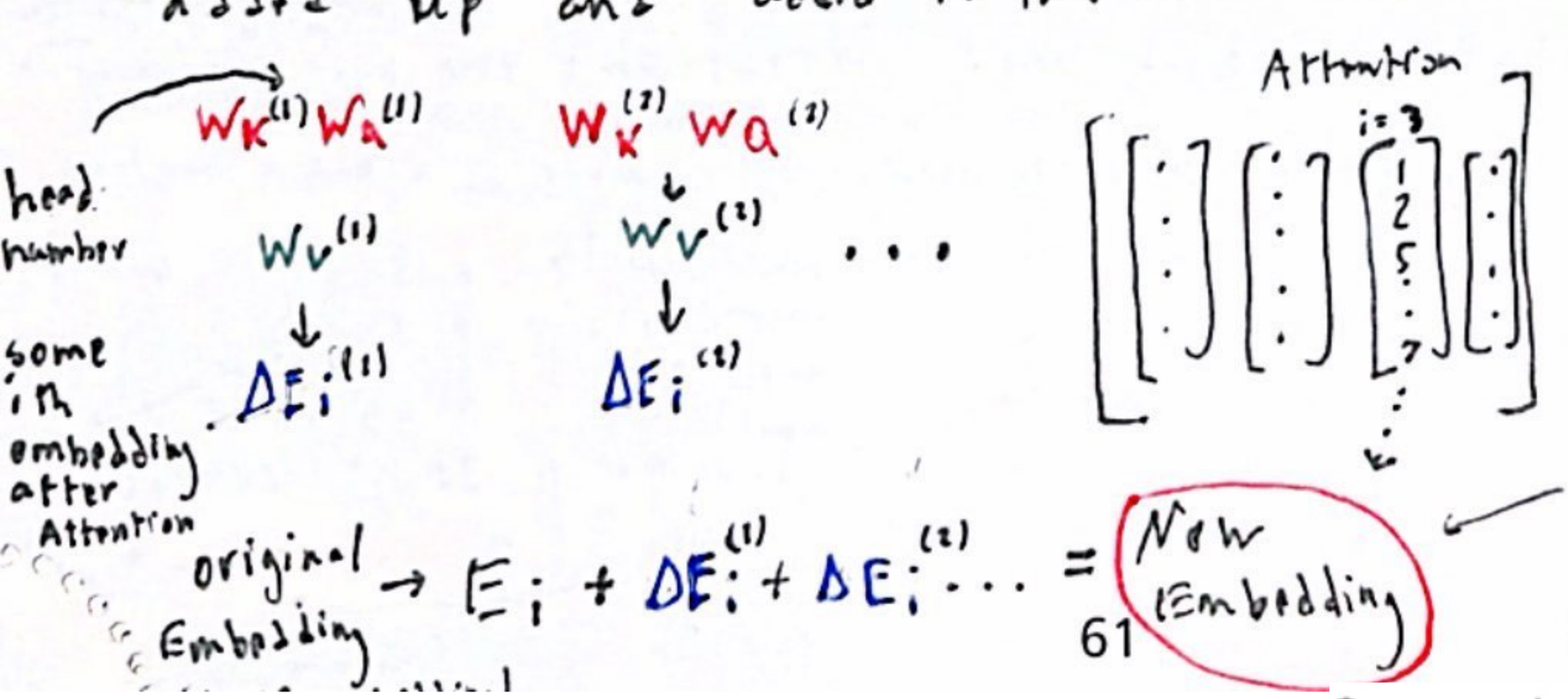
- is another head of attention telling the position and orientation of the creature

- is one head of attention, saying creature is fluffy and blue



• This together is multi headed attention, so this + 10'000 = full context

★ Each head has its own Key, Query and Value matricies these matricies parameters dictate the context its adding Like adjectives updating nouns, and is learned in training, there are 10000's of these heads each learning its "attention patterns" in training. ★ each attention head proposes a change to the embedding of the word, these changes from each head are added up and added to the original embedding in that attention Block



- after all these heads act on E_i it comes out of Attention Block as a refined embedding, so this happens for all embeddings

Scanned with CamScanner