𝗧𝗵𝗶𝘀 𝗶𝘀 𝗵𝗮𝗻𝗱𝘀 𝗱𝗼𝘄𝗻 𝗼𝗻𝗲 𝗼𝗳 𝘁𝗵𝗲 𝗕𝗘𝗦𝗧 𝘃𝗶𝘀𝘂𝗮𝗹𝗶𝘇𝗮𝘁𝗶𝗼𝗻 𝗼𝗳 𝗵𝗼𝘄 𝗟𝗟𝗠𝘀 𝗮𝗰𝘁𝘂𝗮𝗹𝗹𝘆 𝘄𝗼𝗿𝗸. ⬇️  
  
𝘓𝘦𝘵'𝘴 𝘣𝘳𝘦𝘢𝘬 𝘪𝘵 𝘥𝘰𝘸𝘯:  
𝗧𝗼𝗸𝗲𝗻𝗶𝘇𝗮𝘁𝗶𝗼𝗻 & 𝗘𝗺𝗯𝗲𝗱𝗱𝗶𝗻𝗴𝘀:  
- Input text is broken into tokens (smaller chunks).   
- Each token is mapped to a vector in high-dimensional space, where words with similar meanings cluster together.   
  
𝗧𝗵𝗲 𝗔𝘁𝘁𝗲𝗻𝘁𝗶𝗼𝗻 𝗠𝗲𝗰𝗵𝗮𝗻𝗶𝘀𝗺 (𝗦𝗲𝗹𝗳-𝗔𝘁𝘁𝗲𝗻𝘁𝗶𝗼𝗻):   
- Words influence each other based on context — ensuring "bank" in riverbank isn’t confused with financial bank.   
- The Attention Block weighs relationships between words, refining their representations dynamically.   
  
𝗙𝗲𝗲𝗱-𝗙𝗼𝗿𝘄𝗮𝗿𝗱 𝗟𝗮𝘆𝗲𝗿𝘀 (𝗗𝗲𝗲𝗽 𝗡𝗲𝘂𝗿𝗮𝗹 𝗡𝗲𝘁𝘄𝗼𝗿𝗸 𝗣𝗿𝗼𝗰𝗲𝘀𝘀𝗶𝗻𝗴)   
- After attention, tokens pass through multiple feed-forward layers that refine meaning.   
- Each layer learns deeper semantic relationships, improving predictions.   
  
𝗜𝘁𝗲𝗿𝗮𝘁𝗶𝗼𝗻 & 𝗗𝗲𝗲𝗽 𝗟𝗲𝗮𝗿𝗻𝗶𝗻𝗴  
- This process repeats through dozens or even hundreds of layers, adjusting token meanings iteratively.   
- This is where the "deep" in deep learning comes in — layers upon layers of matrix multiplications and optimizations.   
  
𝗣𝗿𝗲𝗱𝗶𝗰𝘁𝗶𝗼𝗻 & 𝗦𝗮𝗺𝗽𝗹𝗶𝗻𝗴  
- The final vector representation is used to predict the next word as a probability distribution.   
- The model samples from this distribution, generating text word by word.   
  
𝗧𝗵𝗲𝘀𝗲 𝗺𝗲𝗰𝗵𝗮𝗻𝗶𝗰𝘀 𝗮𝗿𝗲 𝗮𝘁 𝘁𝗵𝗲 𝗰𝗼𝗿𝗲 𝗼𝗳 𝗮𝗹𝗹 𝗟𝗟𝗠𝘀 (𝗲.𝗴. 𝗖𝗵𝗮𝘁𝗚𝗣𝗧). 𝗜𝘁 𝗶𝘀 𝗰𝗿𝘂𝗰𝗶𝗮𝗹 𝘁𝗼 𝗵𝗮𝘃𝗲 𝗮 𝘀𝗼𝗹𝗶𝗱 𝘂𝗻𝗱𝗲𝗿𝘀𝘁𝗮𝗻𝗱𝗶𝗻𝗴 𝗵𝗼𝘄 𝘁𝗵𝗲𝘀𝗲 𝗺𝗲𝗰𝗵𝗮𝗻𝗶𝗰𝘀 𝘄𝗼𝗿𝗸 𝗶𝗳 𝘆𝗼𝘂 𝘄𝗮𝗻𝘁 𝘁𝗼 𝗯𝘂𝗶𝗹𝗱 𝘀𝗰𝗮𝗹𝗮𝗯𝗹𝗲, 𝗿𝗲𝘀𝗽𝗼𝗻𝘀𝗶𝗯𝗹𝗲 𝗔𝗜 𝘀𝗼𝗹𝘂𝘁𝗶𝗼𝗻𝘀.   
  
Here is the full video from 3Blue1Brown with exaplantion. I highly recommend to read, watch and bookmark this for a further deep dive: [**https://lnkd.in/dAviqK\_6**](https://lnkd.in/dAviqK_6)  
  
𝗜 𝗲𝘅𝗽𝗹𝗼𝗿𝗲 𝘁𝗵𝗲𝘀𝗲 𝗱𝗲𝘃𝗲𝗹𝗼𝗽𝗺𝗲𝗻𝘁𝘀 — 𝗮𝗻𝗱 𝘄𝗵𝗮𝘁 𝘁𝗵𝗲𝘆 𝗺𝗲𝗮𝗻 𝗳𝗼𝗿 𝗿𝗲𝗮𝗹-𝘄𝗼𝗿𝗹𝗱 𝘂𝘀𝗲 𝗰𝗮𝘀𝗲𝘀 — 𝗶𝗻 𝗺𝘆 𝘄𝗲𝗲𝗸𝗹𝘆 𝗻𝗲𝘄𝘀𝗹𝗲𝘁𝘁𝗲𝗿. 𝗬𝗼𝘂 𝗰𝗮𝗻 𝘀𝘂𝗯𝘀𝗰𝗿𝗶𝗯𝗲 𝗵𝗲𝗿𝗲 𝗳𝗼𝗿 𝗳𝗿𝗲𝗲: [**https://lnkd.in/dbf74Y9E**](https://lnkd.in/dbf74Y9E)