



ok\_embd [32\*128, n\_vocab]

↓ ge\_roads(n\_batch)

inp\_embd [32\*128, n\_batch]

↓ cp

la\_er\_embd [32\*128, n\_batch]

↓

Self Attention  
heads: 32  
head\_embd: 128

↓

sa\_embd [32\*128, n\_batch]

