





Day 2: Embeddings and Vector Stores/Databases Resources mentioned in today's livestream: Jinhyuk Lee's Google Scholar profile: <https://scholar.google.com/citations?...>  
The original transformer paper: Attention Is All You Need <https://arxiv.org/abs/1706.03762>  
BERT paper explaining bidirectional attention: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding <https://arxiv.org/abs/1810.04805> A recent paper from NVidia explaining how to adapt decoder-only language model for embedding generation: NV-Embed: Improved Techniques for Training LLMs as Generalist Embedding Models <https://arxiv.org/abs/2405.17428> Whitepaper explaining "Native Integration of the ScaNN Algorithm into AlloyDB Database Internals": <https://services.google.com/fh/files/...>  
Assignments:  [Optional] Listen to the summary podcast episode ( • [Whitepaper Companion Podcast - Embedd...](#)) for this unit (created by NotebookLM, <https://notebooklm.google.com/>).  Read the "Embeddings and Vector Stores/Databases" whitepaper - <https://kaggle.com/whitepaper-embeddi...>  Complete these code labs on Kaggle: Build a RAG question-answering system over custom documents - <https://www.kaggle.com/code/markisher...> Explore text similarity with embeddings - <https://www.kaggle.com/code/markisher...> Build a neural classification network with Keras using embeddings - <https://www.kaggle.com/code/markisher...>