Required Dependencies:

```
# Core Packages
pip install streamlit pandas numpy torch transformers accelerate bitsandbytes

# LangChain Ecosystem (Updated modular imports)
pip install langchain-core langchain-text-splitters langchain-community langchai

# Vector Store & Embeddings
pip install faiss-cpu sentence-transformers

# Document Loaders
pip install PyMuPDF python-docx
```

Advanced RAG System

Upload documents, build a knowledge base, and ask questions!

Document Upload & Processing

Upload documents ?

Drag and drop files here

Limit 200MB per fil…

Browse files

1706.03762v7.pdf 2.1MB ✕

Uploaded Files

1706.03762v7.pdf (application/pdf)

Process Documents

Query Interface

Knowledge base is ready! Ask your questions below.

Enter your question:

Formula for attention

Ask Question

Answer

The formula for attention, as described in section 3.2.1 of Document 0, is:

Attention(Q, K, V) = softmax(QK^T / sqrt(dk))V

Document Chunks Preview

Select chunk to preview

Chunk 1

Chunk 1 Content

Provided proper attribution is provided, Google hereby grants permission to reproduce the tables and figures in this paper solely for use in journalistic or scholarly works.
Attention Is All You Need
Ashish Vaswani∗

```
{
  "source" : "1706.03762v7.pdf"
  "start_index" : 0
}
```

where Q, K, and V are matrices representing queries, keys, and values, respectively, with dimensions dk x m (where m is the batch size), and dk is the dimension of the keys and queries. The softmax function is applied to the scaled dot products between the queries and keys, and the resulting weights are used to compute a weighted sum of the values. This weighted sum is then returned as the output of the attention mechanism.

In section 3.2.2, it is mentioned that this attention mechanism is scaled by a factor of 1/sqrt(dk) to address the issue of growing dot products as dk increases, which can lead to small gradients during training. Without this scaling, an alternative attention mechanism called additive attention was found to outperform dot-product attention for larger values of dk, but with slower practical performance due to its higher theoretical complexity.

Retrieved Documents

› Document 1

› Document 2

› Document 3

› Document 4

› Document 5

Save Query

How to Use:

1. **Initialize Models**: Click 'Initialize Models' in the sidebar
2. **Upload Documents**: Use the file uploader to add PDF, DOCX, or TXT files
3. **Process Documents**: Click 'Process Documents' to chunk and embed your files
4. **Ask Questions**: Enter questions in the query interface

Free Deployment Options:

- **Streamlit Cloud**: Connect your GitHub repo with this app
- **Hugging Face Spaces**: Upload as a Streamlit Space
- **Railway/Render**: Deploy with minimal configuration