The paper "Deja Vu: Contextual Sparsity for Efficient LLMs at Inference Time" is closely related to the Transformer architecture, which is the foundation of modern large language models (LLMs) like GPT-3, PaLM, and OPT. Here's how the paper connects to Transformers:

1. **Focus on Transformer Components**: The paper specifically targets the efficiency of the two main components in Transformer layers:
   - Multi-head attention blocks
   - MLP (feed-forward) blocks

2. **Contextual Sparsity in Attention**: The paper discovers that for a given input, many attention heads can be skipped (up to 80% sparsity) without affecting output quality. This is particularly interesting because:
   - Different inputs activate different subsets of attention heads
   - The paper shows attention heads behave like clustering algorithms (similar to mean-shift clustering)

3. **MLP Sparsity**: The paper finds even higher sparsity (up to 95%) in MLP layers due to ReLU/GeLU activations creating natural sparsity.

4. **Residual Connections**: The work leverages the residual connection property of Transformers (where embeddings change slowly across layers) to enable asynchronous prediction of which components can be skipped.

5. **Implementation Optimizations**: The proposed DEJAVU system provides hardware-aware optimizations specifically for Transformer inference, achieving 2-6× speedups over standard Transformer implementations.

The key innovation is showing that Transformers exhibit "contextual sparsity" - where different inputs naturally activate different sparse pathways through the model - and developing methods to exploit this for efficient inference while maintaining the Transformer's in-context learning capabilities.