Your 598 ID: 5107

Title of paper: Attention is All You Need by Polosukhin et al.
What is their primary result? The authors propose a novel network architecture tiled "the Transformer," which is based solely on attention mechanisms, and dispensing with recurrence and convolutions found in the more prolific sequence transduction models.
Why is this important? The authors provide experimental data suggesting that this model is superior in quality, more parallelizable, and requires less training time.
What are their key ideas? The Transformer follows the overall architecture using stacked self-attention and point-wise, fully connected layers for both the encoder and decoder, but
What are the limitations, either in performance or applicability?
What might be an interesting next step based on this work?
What's the architecture?
How did they train and evaluate it?
Did they implement something?
Grader's 598 ID: