The model is explained in the paper by Santos, Vaswani, Gonnet & Brevet titled “Attention Is All You Need” as a newer architecture applied to sequence transduction, particularly in natural language processing. Therefore, the paper does not recur or use convolution and exploits attention mechanisms for dependencies in sequences. It can also afford a high level of parallelism during the training, which results in efficiency and performance. From the above discussion, it can be seen that the proposed model can work with sequences and also focuses on the encoder-decoder architecture, which emphasizes the associations between words.

The architecture of the proposed model can be studied as follows:

* Self-Attention & Multi-Head Attention: These mechanisms do not just finish at word order, but they encode long-distance dependencies with the connection between any two constituents of sequence.
* Positional Encoding & Feed-Forward Networks: Unlike recurrent models that require memory to store information, the Model does not use memory. These techniques compensate for this by encoding word order and prologue, which include more at the processing end.
* Layer Normalization & Residual Connections: These aid in stabilizing the training process and the rate of convergence the model goes through.
* Parallelization: It is natural to train on multiple processors simultaneously, much quicker than training on the previous models.

Training Model models is a compute-intensive task that leverages massive datasets like the WMT English-to-German translation tasks. Here's a breakdown of the critical aspects:

* Data Preparation: Millions of sentence pairs are meticulously prepared. Byte-pair encoding creates a shared vocabulary and similar lengths of batch sentences to optimize training efficiency.
* Hardware Acceleration: Powerful GPUs like NVIDIA P100s are used to handle heavy computations. Training times vary depending on model size, ranging from 12 hours for smaller models to 3.5 days for larger ones.
* Adaptive Learning: The Adam optimizer uses a warmup-decay learning rate. This approach gradually increases the learning rate at the beginning (warmup) and then smoothly reduces it as training progresses.
* Regularization Techniques: Techniques like dropout help prevent the model from overfitting on the specific training data, allowing it to generalize better to unseen examples.

The results demonstrate that the Model performs state-of-the-art translation tasks, overpowering previous models while requiring comparatively less training time and computational resources.

In a gist, the model's architecture and training methodology represents a significant advancement in sequence modeling, showcasing the power of attention mechanisms in handling complex language tasks efficiently.