**Summary of Transformer Architecture Study**

**Introduction**

The study explores variations on the Transformer architecture, focusing on its application in English-to-German translation and English constituency parsing. Transformers, known for their efficiency and effectiveness in sequence transduction tasks, leverage multi-headed self-attention mechanisms to achieve superior performance over traditional recurrent and convolutional models.

**Variations on the Transformer Architecture**

**Metrics and Parameters:**

**N (Layers)**: Number of layers in the Transformer model.

**d (Model Dimension)**: Dimensionality of input and output.

**d\_ff (Feedforward Dimension)**: Dimension of the feedforward network.

**h (Heads)**: Number of attention heads.

**d\_k, d\_v (Key, Value Dimensions)**: Dimensions for key and value vectors.

**P\_drop (Dropout Probability)**: Dropout rate.

**ε\_ls (Label Smoothing)**: Label smoothing factor.

**Train Steps**: Number of training steps.

**PPL (Perplexity)**: Per-wordpiece perplexity.

**BLEU (Bilingual Evaluation Understudy)**: Score measuring translation quality.

**Params (Parameters)**: Number of parameters in the model.

**Key Variations:**

**Base Model**: 6 layers, 512 model dimension, 2048 feedforward dimension, 8 heads, 64 key and value dimensions, 0.1 dropout, 0.1 label smoothing, 100K training steps. Perplexity: 4.92, BLEU: 25.8, Params: 65M.

**Row A**: Varying number of attention heads and attention dimensions.

**Row B**: Reducing attention key size, resulting in lower model quality.

**Row C & D**: Bigger models with higher dropout to avoid overfitting.

**Row E**: Replacing sinusoidal positional encoding with learned positional embeddings, resulting in similar performance to the base model.

**English Constituency Parsing**

To test the generalization of Transformers, experiments were conducted on English constituency parsing using the Wall Street Journal (WSJ) dataset. The results demonstrate the model's capability to handle tasks with strong structural constraints and longer outputs.

**Training and Results:**

**4-layer Transformer with 1024 model dimension**.

**Datasets**: WSJ (40K sentences) and a larger semi-supervised corpus (17M sentences).

**Vocabulary**: 16K tokens for WSJ only, 32K tokens for semi-supervised setting.

**Inference**: Beam size of 21, α = 0.3.

**Results**: Outperformed previous models except for the Recurrent Neural Network Grammar. Achieved an F1 score of 91.3 with WSJ only and 92.7 with semi-supervised training.

**Attention Visualizations**

**Figure 3: Long-Distance Dependencies**

**Layer 5 of 6**: Attention mechanism in the encoder self-attention layer identifies long-distance dependencies, completing phrases like 'making...more difficult'.

**Visualization**: Different colors represent different attention heads, focusing on the word 'making'.

**Figure 4: Anaphora Resolution**

**Layer 5 of 6**: Two attention heads involved in resolving anaphora.

**Top Visualization**: Full attentions for head 5.

**Bottom Visualization**: Isolated attentions from the word 'its' for heads 5 and 6.

**Figure 5: Sentence Structure**

**Layer 5 of 6**: Attention heads exhibit behavior related to sentence structure, with different heads learning to perform specific tasks.

**Conclusion**

The Transformer model, entirely based on attention mechanisms, demonstrates significant improvements in sequence transduction tasks, particularly in machine translation and parsing tasks. The ability to train faster and achieve state-of-the-art results underscores the potential of attention-based models for a wide range of applications. Future work will extend these models to handle diverse input and output modalities and investigate more efficient attention mechanisms.

**Visual Representation (Sample Charts)**

1. **BLEU Score Comparison**:

|  |  |
| --- | --- |
| **Model Variation** | **BLEU Score** |
| Base | 25.8 |
| A | 24.9-25.8 |
| B | 25.5 |
| C | 25.4-26.0 |
| D | 25.3-26.2 |
| E | 24.5-25.7 |

2. **Perplexity Comparison**:

|  |  |
| --- | --- |
| **Model Variation** | **Perplexity** |
| Base | 4.92 |
| A | 5.01 |
| B | 5.00 |
| C | 4.66-5.12 |
| D | 4.75-5.77 |
| E | 4.33-5.77 |

By summarizing the detailed study of the Transformer architecture, this document aims to provide a comprehensive understanding of the variations and their impact on model performance, illustrated with relevant visualizations and metrics.