CS 585 Project Proposal - Team Catastrophe

1. Description of the Problem and Proposed Solution

Our team is implementing the LRCN: Layer-residual Co-Attention Networks for Visual Question Answering (VQA) paper.

- The primary problem is improving the performance of Visual Question Answering systems, which require deep understanding and interaction between visual content and textual information.
- VQA systems typically face challenges in integrating multimodal data, especially when the information transfer between the layers of deep models is inefficient.
- To overcome these challenges, the paper proposes implementing the Layer-Residual Mechanism (LRM), a plug-and-play solution that optimizes the multimodal feature transfer and stabilizes training in deeper layers.
- The solution involves applying the LRM to the Encoder-Decoder, Pure-Stacking, and Co-Stacking structures to enhance model performance across different VQA tasks.

2. Overview of How the Solution Can Be Achieved

The implementation will proceed in the following steps:

- **Dataset Selection:** We will use the **VQA v2 and CLEVR** datasets to train and evaluate our models. Both datasets are popular benchmarks for VQA tasks and will allow us to comprehensively evaluate our proposed solution.
- **Model Architecture:** We will begin with existing **transformer-based models (Encoder-decoder, pure stacking, and Co-stacked)** and integrate the LRM. The LRM adds residual connections between layers to enhance the transfer of information and mitigate the vanishing gradient problem in deep networks.
- **Training and Evaluation:** After modifying the architecture, we will train the model using standard hyperparameters and evaluate it on the datasets using standard VQA metrics.

3. Relevant Papers and Links

1. LRCN: Layer-residual Co-Attention Networks for Visual Question Answering. Link
2. CLEVR Dataset Link
3. VQA V2 Dataset Link

4. Existing Code to Build On

We do not have any existing code, though detailed implementation instructions are in the paper.

5. Datasets

We will use two datasets:

- **VQA v2 Dataset**
- **CLEVR Dataset**

Dataset Statistics Table:

Dataset	Train Images	Validation images	Test Images	Disk Space	Training Questions | Answers	Validation Questions | Answers	Test Questions
VQA v2	204,721	40,504	81,434	30GB	443,757 | 4,437,570	214,354 | 2,143,540	447,793 | 4,477,930
CLEVR	70,000	15,000	15,000	18GB	700,000 | 700,000	150,000 | 150,000	150,000 | 150,000

6. Contributions by Team Members

1. **Gagan Singhal:** Responsible for overseeing the theoretical aspects of the LRM and model architecture modifications. Will implement the encoder-decoder architecture and integrate the LRM.
2. **Astha Rastogi:** Handles dataset preprocessing, including data augmentation and splitting. Will be responsible for evaluating the model using VQA and CLEVR datasets.
3. **Chaitanya Chakka:** Implement the training pipeline, including setting up the model, training procedures.
4. **Satya Akhil Galla:** Model hyperparameter tuning and model inference.