Ali Al-kinani
S214345186@deakin.edu.au

Final Report: Building a Question-Answering Model for Simpson's Dataset

Table of Contents

Ali Al-kinani
S214345186@deakin.edu.au

Introduction

The Deakin Simpsons Challenge 2023 stands as an engaging platform at the confluence of computer vision and natural language processing. In this competition, participants are tasked with a dual mission: interpreting images from the Simpsons universe and responding to accompanying natural language queries. The primary goal is to forge a robust machine learning and deep learning model that can not only grasp the visual intricacies of the provided images but also craft coherent, accurate, and contextually relevant responses to the posed questions. This synthesis of image understanding, and linguistic articulation encapsulates the essence of the competition, promoting innovation in AI-driven comprehension of visual content and its communication through language.

Model Architecture Design and Overfitting Challenges

Our model design drew inspiration from a significant research paper introducing the "deeper LSTM Q" model. This approach elegantly combined Convolutional Neural Networks (CNN) for images and Recurrent Neural Networks (RNN) for text analysis. However, as we delved into the model's performance, a complex issue emerged: overfitting. While the model demonstrated remarkable mastery over the training data, it struggled with novel validation data. This dilemma prompted us to embark on a thorough exploration and enhancement process.

Strategies to Address Overfitting

Multiple strategies were embraced in our quest to neutralize overfitting's detrimental impact. The integration of dropout layers introduced an element of randomness during training, compelling the model to be more adaptive. Additionally, L2 regularization was adopted to penalize large weights, restraining the model from overemphasizing specific features. Although these measures exhibited progress, it became evident that an amalgamation of techniques would be indispensable.

Data Augmentation as a Generalization Technique

In our pursuit to bolster model generalization, data augmentation techniques were incorporated. The introduction of image shifting, a form of transformation, augmented the dataset without compromising its integrity. This augmentation diversifies the training data, enabling the model to learn invariant features across different visual contexts.

Finding the Optimal Vocabulary Size and Embedding Dimension:

The textual questions, forming an integral component of the model input, underwent a meticulous assessment of vocabulary size and embedding dimensions. The vocabulary size influenced the richness of the linguistic representation, while embedding dimensions impacted the ability to capture semantic nuances. Calibration of these dimensions was pivotal in achieving an equilibrium between model expressiveness and generalization.

CNN and RNN Fusion for Enhanced Performance

In response to the intricate interplay between images and textual questions, a profound architectural refinement was initiated. This involved adopting CNN for image input processing and RNN (SimpleRNN) for textual input processing. This fusion accentuated the model's capability to concurrently comprehend image and text, resulting in a synergistic augmentation of performance.

Training and Validation

Iterative training and validation constituted the heartbeat of our model development. Through multiple epochs, we diligently monitored the loss and accuracy metrics on both training and validation sets. This dynamic evaluation provided insights into the model's learning trajectory and its propensity to generalize on unseen data.

Challenges and Lessons Learned

The journey was fraught with challenges. Overcoming overfitting required persistent experimentation with an amalgamation of strategies, reinforcing the adage that a singular solution might not suffice. Balancing model complexity to avert overfitting and underfitting proved intricate, necessitating a deep understanding of architecture intricacies. The tuning of hyperparameters, from learning rates to dropout probabilities, emerged as an artful yet essential process. Striking the right balance between CNN and RNN layers, coupled with the judicious selection of activation functions, fortified the model's overall robustness.

Ali Al-kinani
S214345186@deakin.edu.au

Conclusion

In my quest to conquer overfitting in the context of the Simpson's challenge, I adopted an iterative and adaptive approach. I began by designing an architecture inspired by the "deeper LSTM Q" model, intertwining Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN) for image and text processing. Yet, overfitting emerged as a challenge.

To combat this, I fine-tuned the architecture. I introduced dropout layers and L2 regularization, curbing the model's inclination to over memorize while fostering generalization. Despite these measures, overfitting persisted.

A pivotal moment arrived when I acknowledged the value of data diversity. Leveraging data augmentation, particularly image shifting, I diversified the training dataset to enhance its representation. Although progress was made, remnants of overfitting remained.

I delved into architectural adjustments, downsizing dense layers, embeddings, and dropout rates. These changes, informed by insights from discussions, aimed to strike a balance between complexity and overfitting resistance.

The infusion of CNN and SimpleRNN layers enhanced the model's fusion of visual and textual data processing. Hyperparameter tuning underscored the journey, meticulously fine-tuning learning rates and dropout probabilities to achieve convergence speed and overfitting mitigation.

This pursuit was marked by a dynamic approach, incorporating diverse advice. Despite multifaceted attempts, the overfitting challenge lingered, spurring continued exploration of activation functions, dropout rates, and architecture.

In summary, this journey against overfitting underscores the fusion of architectural design, hyperparameter optimization, and data pre-processing. It highlights the dedication, adaptability, and methodical exploration inherent in refining a machine learning model, aiming for a harmonious balance between performance and generalization.