A residual GRU-Attention hybrid architecture for high-fidelity Chinese classical poetry generation

Ji Li[1*]

[1]Department of Computer Science, University of Utah, Salt Lake City, UT, USA

Abstract. Automatic generation of classical poetry is a highly challenging constrained text generation task, demanding strict adherence to stylistic rules like meter, rhyme, and parallelism alongside high-level semantic coherence. While hybrid architectures combining Recurrent Neural Networks (RNNs) and Attention mechanisms show promise, they often suffer from significant training instability, hindering their performance. This paper systematically investigates these stability challenges in RNN-Attention hybrid architectures and demonstrates through comparative experiments the decisive impact of architectural integration strategies on model convergence and generation quality. We validate our findings on Chinese classical poetry generation, a domain characterized by its intricate constraints. Experimental results show that our proposed model (V3), which employs a novel integration strategy, achieves a comprehensive quality score of 8.1, marking a significant improvement of 24.6% over a strong GRU baseline (V2). Furthermore, ablation studies confirm that a naive concatenation of modules leads to mode collapse, highlighting that the proposed integration strategy—rather than the simple stacking of components—is the key to ensuring model stability and achieving superior performance. Our work offers valuable insights for designing robust hybrid models for other complex text generation tasks.

1 Introduction

1.1 Constrained Poetry Generation

Chinese classical poetry is a highly constrained text generation task, requiring adherence to strict rules governing structure, rhyme, and tonal patterns [1]. The multi-layered and stringent nature of these constraints establishes it as an ideal proving ground for constrained text generation, demanding that a model demonstrate not only linguistic fluency and semantic coherence but also consistent adherence to intricate, non-local regulations throughout the generative process [2].

Early methods based on Recurrent Neural Networks (RNNs) [3] were effective at learning local poetic structures but struggled with long-range dependencies due to the vanishing gradient problem, often resulting in theme drift. The Transformer architecture [4], with its self-attention mechanism, largely solved this by modeling global context. Nevertheless,

* Corresponding author: u1497986@umail.utah.edu

given the inherent advantages of RNNs' sequential inductive bias, exploring hybrid architectures that combine the strengths of both models remains a promising research direction.

1.2 Instability of Hybrid Architectures: A Principled Integration Method

Although the combination of RNNs and attention mechanisms has been applied in many NLP tasks, the impact of their integration strategy on model stability has not been systematically studied. In our exploration, we identified a critical issue: an intuitive and naive approach to architectural fusion—simply concatenating the output features of an RNN with those of an attention module—leads to severe training instability and triggers a catastrophic Mode Collapse [5], rendering the model completely ineffective.

This observation points to a significant gap in the literature regarding the architectural principles essential for avoiding such catastrophic failures. Addressing this, this paper systematically investigates the training stability issues in RNN-Attention hybrid architectures to demonstrate the decisive impact of integration strategies on model convergence. Through targeted ablation studies, we prove that a principled integration methodology, rather than simple superposition of components, is the key to success. We validate our proposed stable architecture on the highly constrained task of Chinese classical poetry generation, showing significant performance improvements over strong baselines.

2 Models and Methods

2.1 Dataset and Preprocessing

The corpus for this study is an open-source dataset containing 57,580 Tang dynasty quatrains. Our preprocessing pipeline, adapted from [6], involved character-level tokenization with <START> and <EOP> tokens added to each poem. All sequences were vectorized using a character-to-index mapping and left-padded to a fixed length using a padding token, which was ignored during loss calculation.

2.2 Model Architectures for Comparison

To systematically evaluate our proposed hybrid architecture, we compared it against two representative RNN baseline models.

Our first baseline, a standard LSTM model [7], exhibited severe repetition and mode collapse (e.g., generating nonsensical loops like "花落花开日, 花落柳条春"), exposing the core deficiency of standard RNNs. This result exposed the core deficiency of standard RNN models in generating diverse and creative content without additional mechanisms.

The second baseline is an optimized model where we replaced the LSTM units with GRU units [8]. GRU is comparable in performance to LSTM but has a simpler structure and fewer parameters, which generally leads to easier convergence. After extensive training, this model performed excellently, completely resolving the repetition issues of the V1 model and generating semantically coherent and aesthetically pleasing verses. Therefore, we positioned the V2 model as a strong baseline, representing the state-of-the-art for a well-tuned, purely recurrent network and providing a challenging comparison for our proposed V3 model

2.3 Proposed Hybrid Architecture (V3)

Our proposed V3 architecture aims to combine the sequential modeling capability of GRU with the global context-capturing ability of self-attention. During development, we explored two different integration strategies, which form our core ablation study. By ensuring stable gradient flow and simplifying optimization, this design effectively avoids the mode collapse observed in the naive variant. The schematic for this model is presented in Fig. 1.

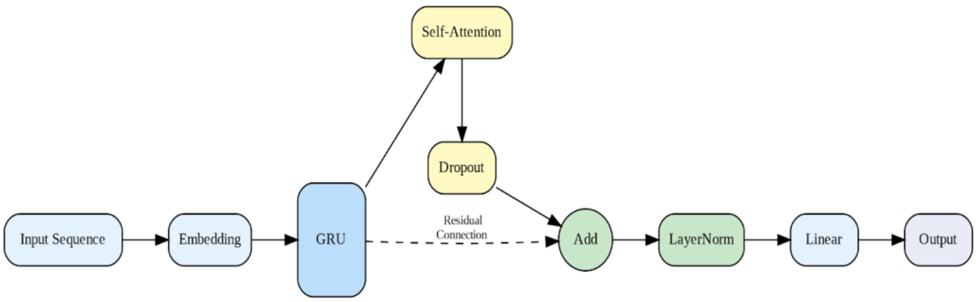

Fig. 1. V3 Hybrid Model Architecture

The Final Principled Design (V3-Optimized) Inspired by the design of the Transformer architecture, particularly its core "Add & Norm" module, we re-architected our hybrid model.

In this optimized design, the output of the self-attention module is no longer concatenated with the GRU's output. Instead, it is "added" back to the GRU's output via a residual connection (an addition operation), followed by stabilization through a Layer Normalization operation.

Theoretical Basis The addition operation implements the residual connection proposed by He et al. [9]. Residual connections offer two major theoretical advantages in deep networks. They provide a "gradient highway" to mitigate the vanishing gradient problem and make it easier for the model to learn an identity mapping.If the contextual information provided by the attention module is not beneficial at a certain timestep, the model can easily revert to the original GRU output by driving the weights of the attention module's output towards zero, which greatly simplifies the optimization process.

Another core principle is Layer Normalization, proposed by Ba et al. [10], which stabilizes training dynamics by normalizing activations within a layer. This is particularly important in RNNs as it helps mitigate Internal Covariate Shift—the phenomenon where the distribution of inputs to subsequent layers continuously changes as the parameters of previous layers are updated during training. This forces layers to constantly adapt to new data distributions, slowing down training and affecting stability.

These two components work in synergy to form a powerful stabilization framework. The residual connection provides a direct information pathway, while layer normalization ensures that the fused signal (gru_outputs + attention_out) remains within a well-behaved, stable distribution. This makes the optimization process smoother, thereby avoiding the exploding or vanishing activations observed in the naive concatenation model and ultimately preventing mode collapse.

2.4 Experimental Setup

All models were trained using the PyTorch framework with the AdamW optimizer [11] and a CosineAnnealingLR learning rate scheduler.The number of training epochs for each model was determined by its convergence: V1 was trained for 69 epochs, V2 for 80 epochs, while V3 in 99 epochs.

Given the complexity of evaluating creative text generation tasks, we employed a hybrid evaluation methodology. We designed a multi-dimensional scoring system covering eight aspects: fluency, creativity, imagery, word choice, thematic relevance, metrical regularity, repetition avoidance, and emotional expression. The scores were weighted with 50% from human experts and 50% from large language models (25% from Claude and 25% from Gemini), aiming to combine deep human understanding with the scalability of LLMs. We avoided n-gram metrics like BLEU or ROUGE, as they are ill-suited for creative tasks that prioritize semantic quality over exact word matches.

We conducted side-by-side comparisons of generated samples from all models under the same prompts to provide concrete, interpretable evidence of performance. To comprehensively assess the models' generative capabilities, we employed three different decoding strategies: Greedy Search, Beam Search (k=5), and Temperature Sampling (T=0.8), to explore the trade-off between determinism and diversity.

3 Results and Analysis

3.1 Model Performance Assessment

We first conducted a quantitative evaluation of the three core models (V1-LSTM, V2-GRU, V3-Hybrid) using our multi-dimensional scoring system. The results are shown in Table 1.

Table 1: Multi-dimensional Quality Evaluation of Models

Scoring Dimension	V1 Baseline Model	V2 GRU Model	V3 Optimized Model
Fluency	3.0	6.5	8.5
Creative Diversity	2.0	6.0	8.0
Imagery	2.5	6.5	8.5
Word Choice Accuracy	4.0	6.0	8.0
Thematic Relevance	3.5	6.5	7.5
Metrical Regularity	5.0	7.0	8.5
Repetition Avoidance	1.5	7.0	8.0
Emotional Expression	2.5	6.5	8.0
Comprehensive Score	3.0	6.5	8.1

As shown in Table 1, the proposed V3 hybrid architecture (comprehensive score of 8.1) significantly outperforms the strong GRU baseline V2 (6.5), marking a 24.6% improvement. The largest relative gains were in "Creative Diversity" and "Repetition Avoidance," which indicates that the stabilized self-attention mechanism effectively captures global context to break the local dependency limitations of pure RNN models.

Table 2: Model Performance under Different Decoding Strategies

Generation Strategy	V1 Baseline Model	V2 GRU Model	V3 Optimized Model
Greedy Search	2.5	6.0	7.5
Temperature Sampling (T=0.8)	3.0	6.5	8.5
Beam Search (k=5)	3.5	6.8	8.3

The results in Table 2 show that V3's superiority remains robust across different decoding strategies. Notably, its score increases to 8.5 under temperature sampling (T=0.8), suggesting that the model has learned a smoother and richer probability distribution. While increased randomness tended to expose noise in the baseline models, V3 effectively leveraged it to explore novel yet coherent imagery—a key trait for high-quality creative generation.

3.2 Qualitative Analysis and Case Study

To more intuitively demonstrate the differences between the models, we conducted a comparative analysis of their generated samples for the same prompts, as shown in Table 3.

Table 3: Comparative Analysis of Generated Poetry Samples (Greedy Search)

Model	Prompt: "白日依山尽" (The white sun sets behind the mountains)	Prompt: "春夏秋冬" (Spring, Summer, Autumn, Winter - Acrostic)
V1-LSTM	白日依山尽，青山有所依。 白云生白日，白日照青苔。 *(Translation):* The white sun sets behind the mountains, the green mountains have something to rely on. The white clouds produce the white sun, the white sun shines on the green moss.	春风吹落花，夏日照东西。 秋风吹落叶，冬日落寒衣。 *(Translation):* Spring winds blow fallen flowers, summer sun shines east and west. Autumn winds blow fallen leaves, winter sun brings cold clothes.
V2-GRU	白日依山尽，朱门对郁牢。 江南杨柳色，云隔故交烧。 *(Translation):* The white sun sets behind the mountains, the red gate faces the gloomy fortress. The color of the willows in Jiangnan, clouds separate old friends being burned.	春色满朝阳，夏时气候长。 秋风至朝日，冬日徧阳阳。 *(Translation):* Spring colors fill the morning sun, summer's climate is long. Autumn wind arrives at dawn, winter sun spreads its warmth.
V3-Naive (Concatenation)	白日依山尽，春风春风来。 白日青山近，青山白日开。 *(Translation):* The white sun sets behind the mountains, spring winds, spring winds come. The white sun is near the green mountains, the green mountains and the white sun open.	春风吹，夏风吹，秋风吹，冬风吹。 (Translation): Spring winds blow, summer winds blow, autumn winds blow, winter winds blow.
V3-Hybrid	白日依山尽，黄河涨海流。 累年边地断，今日汉宫愁。 *(Translation):* The white sun sets behind the mountains, the Yellow River swells into the sea. For years the borderlands have been cut off, today sorrow fills the Han palace.	春风吹马毳花骢，夏色如眉照绮楼。 秋光斗艳初呈撮，冬晚如今未出头。 *(Translation):* Spring winds blow on the dappled steed's felt-like coat, summer hues like eyebrows reflect in the ornate tower. Autumn's radiance competes in beauty at its first showing, winter's evening has not yet revealed itself.

For the prompt "白日依山尽" (The white sun sets behind the mountains), the V1-LSTM model shows clear repetition, while the V2-GRU model is more coherent but imagistically obscure. In stark contrast, the V3-Hybrid's continuation constructs a grand narrative chain from a natural landscape to border warfare and finally to courtly sorrow, showcasing superior imagery and thematic coherence.

The acrostic poem task further highlights these differences. The V3-Naive (concatenation) model's output—degenerating into "Spring winds blow, summer winds blow..."—visually demonstrates the catastrophic mode collapse resulting from naive integration. The V3-Hybrid model, however, successfully completes this highly constrained task with rich and ornate language.

3.3 Ablation Study: The Critical Role of Principled Integration

The core argument of this research is that the success of a hybrid architecture lies not in simply stacking modules, but in the method of their integration. The performance difference between V3-Naive (concatenation) and V3-Hybrid (Add & Norm) constitutes the key ablation study supporting this claim.

As experimental logs show, the V3-Naive model experienced severe mode collapse during training. When prompted with "床前明月光" (Moonlight before my bed), the model fell into an infinite loop of "一夜一声声" (one sound after another all night); when generating the "春夏秋冬" acrostic, it degenerated into a crude imitation of "Spring winds blow, summer winds blow...". This phenomenon is a known failure mode in generative models where the output diversity plummets, converging on a few high-frequency patterns.

In stark contrast, the V3-Hybrid model, which employs residual connections and layer normalization, completely eliminated the mode collapse phenomenon. The high-quality outputs in Table 3 are a clear testament to its stability. This result powerfully proves our central hypothesis: residual connections and layer normalization are not merely nice-to-have optimizations but are fundamental enabling technologies that ensure the GRU-attention hybrid architecture can be trained stably and realize its potential. This ablation study clearly shows that simply adding an attention module on top of an RNN is insufficient; the integration methodology itself is the decisive innovation.

4 Conclusion

This paper investigated the training stability of RNN-Attention hybrid architectures, demonstrating that the integration strategy—not the mere combination of components—is the decisive factor for success. Our proposed V3 model, which uses a principled "Add & Norm" framework to fuse a GRU with self-attention, significantly outperforms a strong GRU baseline in generating high-quality Chinese classical poetry. A key ablation study further confirmed that while naive feature concatenation leads to catastrophic mode collapse, our integration method ensures stable training and unlocks the architecture's full potential.

This work highlights that for complex creative tasks, stable architectural design should be considered a core prerequisite for realizing generative capabilities, not a secondary optimization.

Limitations of this study include the dataset's confinement to Tang dynasty quatrains and the inherent subjectivity of poetry evaluation. Future work could focus on the controllable generation of specific styles and emotions, the integration of external knowledge bases, and the application of this stable hybrid architecture to more complex poetic forms.

References

1. Z. Yang, P. Cai, Y. Feng, F. Li, W. Feng, E. S.-Y. Chiu, H. Yu, Generating Classical Chinese Poems from Vernacular Chinese. in Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP) (2019)

2. Q. Wang, T. Luo, D. Wang, Can Machine Generate Traditional Chinese Poetry? A Feigenbaum Test. arXiv preprint arXiv:1606.05829 (2016)
3. X. Zhang, M. Lapata, Chinese Poetry Generation with Recurrent Neural Networks. in Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), 670-680 (2014)
4. A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin, Attention Is All You Need. arXiv preprint arXiv:1706.03762 (2017)
5. I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative Adversarial Nets. in Advances in Neural Information Processing Systems 27 (NIPS 2014), 2672-2680 (2014)
6. WilliamX1, AI-poet. GitHub repository (2021) https://github.com/WilliamX1/AI-poet
7. S. Hochreiter, J. Schmidhuber, Long Short-Term Memory. Neural Computation, 9(8), 1735-1780 (1997)
8. K. Cho, B. van Merrienboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, Y. Bengio, Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation. in Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), 1724-1734 (2014)
9. K. He, X. Zhang, S. Ren, J. Sun, Deep Residual Learning for Image Recognition. in Proceedings of the IEEE conference on computer vision and pattern recognition, 770-778 (2016)
10. J. L. Ba, J. R. Kiros, G. E. Hinton, Layer Normalization. arXiv preprint arXiv:1607.06450 (2016)
11. I. Loshchilov, F. Hutter, Decoupled Weight Decay Regularization. arXiv preprint arXiv:1711.05101 (2019)