(ranging from 0.71 to 1.00) despite exhibiting very poor cluster separation. Figure 9 (d) provides the visual description for this mechanism. A negative score, as in *aircraft*, confirms its samples are, in average, geometrically closer to a neighboring cluster. However, the high F1-score is correlated with the second metric: high readout alignments. This high alignment score means the correct readout vector is strongly aligned with the centroid of its true samples. Therefore, the classification succeeds not because the clusters are geometrically separate (which the Silhouette score proves they are not), but because the readout vector is so precisely aimed at the center of its correct (but messy) cluster that it still "claims" its samples. The projection of these samples onto the correct readout vector is still higher than their projection onto any other vector, resulting in a correct classification. This shows a case where the network's alignment mechanism can succeed despite a failure to create geometrically pure clusters.

Other low-frequency intents, such as *quantity* and *ground_fare*, show other combinations of metric failures. While these represent complex failure modes, the four primary patterns identified and visualized cover the most distinct and interpretable modes of model success and failure.

This case study on ATIS shows the robustness and explanatory power of our framework. It reveals how dataset properties like class imbalance are not abstract statistical issues but are directly encoded into the geometric and topological structure of the network's learned dynamics.

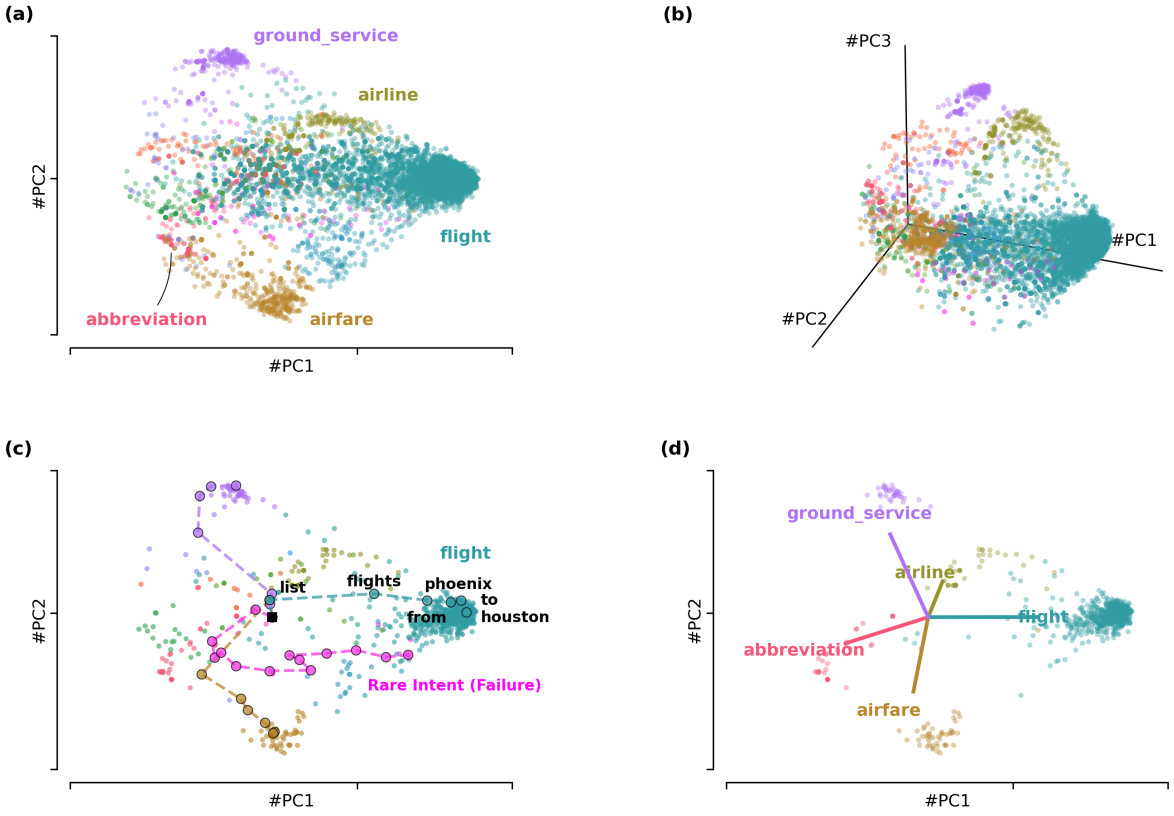

Fig. 8. State space visualizations of the GRU model trained on the imbalanced ATIS dataset. **(a)** 2D projection of the final hidden states, showing a few large distinct clusters for high-frequency intents (e.g. flight, airfare). **(b)** 3D projection, confirming the geometric separation of the major clusters. **(c)** Example 2D trajectories. A "success" (teal) steers to the flight cluster, while a "failure" (magenta) for a rare intent wanders into the wrong cluster. **(d)** Alignment of readout vectors (colored lines) with their corresponding high-frequency clusters (colored dots)

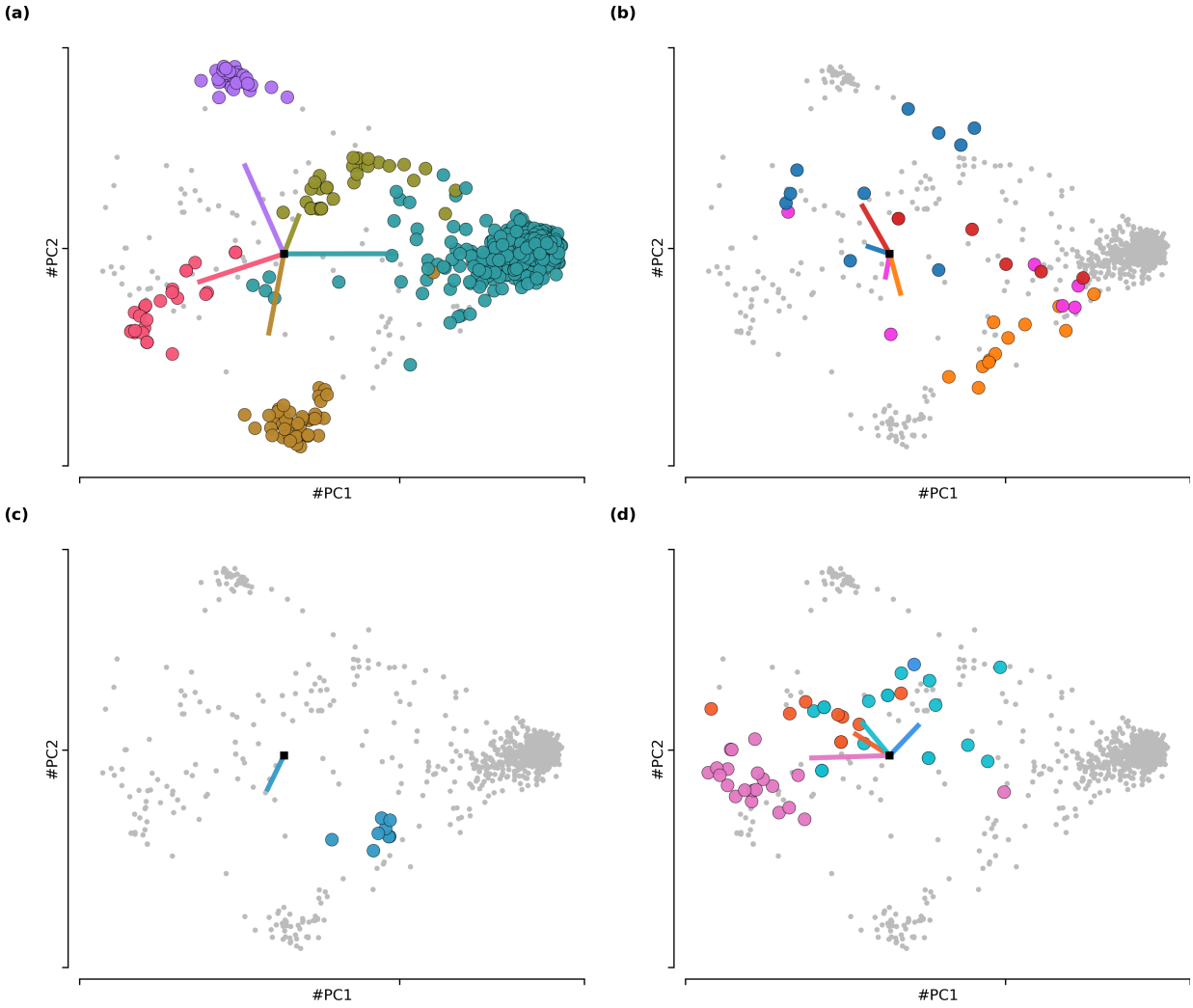

Fig. 9. The four patterns of model behavior in the imbalanced ATIS PCA state space. Final states for analyzed intents (colored dots), all other final states (Gray dots) and readout vectors for analyzed intents (colored lines). **(a)** Pattern 1: Convergent High Performance. High separation and strong alignment. **(b)** Pattern 2: Geometric Collapse. Poor separation and weak alignment. **(c)** Pattern 3: Alignment Failure. Good separation but poor alignment. **(d)** Pattern 4: Alignment-Driven Classification. Poor separation but strong alignment.

8. Conclusion

Intent detection remains a challenging problem that has yet to be fully solved. Conceptually, one can think of intent as evolving dynamically: as more words are processed, the potential intent becomes more constrained, as if moving between interpretations. Our approach embraces this notion, modeling the search for the final intent as a dynamical process within the state space of a Recurrent Neural Network.

In this paper, we applied reverse engineering techniques to study the computational mechanisms of RNNs trained for intent detection. Our analysis of the balanced SNIPS dataset revealed that networks converge to an elegant and highly interpretable solution: they partition their state space into a low-dimensional manifold of distinct, well-separated clusters corresponding to each intent. We showed that sentences evolve along structured trajectories, steering the network's hidden state toward the correct cluster. To test the generalizability of this framework, we extended our analysis to the imbalanced ATIS dataset. This more challenging, real-world scenario revealed how this ideal geometric solution is distorted by class imbalance. Our two-part diagnostic framework (Geometric Separation and

Readout Alignment) allowed us to move beyond simple accuracy scores and identify four distinct, mechanistic patterns of model success and failure, explaining why certain intents perform well while others fail.

While other interpretability methods can identify which input token are salient, our dynamical systems approach provides a unique, mechanistic understanding of how the network's internal state evolves over time and arrive at a decision. This perspective opens several promising avenues for future research. This geometric framework can be extended to address related, high-stakes problems in conversational AI. Our finding in the ATIS dataset provides a concrete empirical basis for this. For example, in out-of-scope (OOS) detection, utterances would likely produce trajectories that fail to converge to any of the established, high-frequency clusters, terminating instead in a dense, undifferentiated central region. This geometric distance from a final state to the nearest cluster could thus serve as a robust signal for OOS detection. The framework could also be adapted to analyze the joint intent detection and slot filling task. This would make it possible to explore how the state space dynamics for intents and slots interact and mutually influence one another. Additionally, the geometric separation of clusters could serve as a new metric for model robustness, or be used to probe for demographic biases in how different user utterances are processed [49]. Finally, a critical and exciting line of work involves adapting this dynamical systems framework from RNNs to the modern Transformer architectures. Understanding the state-space geometry and attractors within these more complex models is a key challenge for the field of interpretability.

References

[1] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard, M. Kudlur, J. Levenberg, R. Monga, S. Moore, D.G. Murray, B. Steiner, P. Tucker, V. Vasudevan, P. Warden, M. Wicke, Y. Yu and X. Zheng, TensorFlow: a system for large-scale machine learning, in: *Proc. Conference on Operating Systems Design and Implementation (USENIX)*, 2016, pp. 265–283.

[2] W.A. Abro, G. Qi, M. Aamir et al., Joint intent detection and slot filling using weighted finite state transducer and BERT, *Applied Intelligence* **52** (2022), 17356–17370.

[3] K. Aitken, V.V. Ramasesh, A. Garg, Y. Cao, D. Sussillo and N. Maheswaranathan, The geometry of integration in text classification RNNs, in: *International Conference in Learning Representation (ICLR)*, 2021.

[4] D. Bahdanau, K. Cho and Y. Bengio, Neural Machine Translation by Jointly Learning to Align and Translate, in: *International Conference in Learning Representation (ICLR)*, 2015.

[5] E. Bastianelli, A. Vanzo, P. Swietojanski and V. Rieser, SLURP: A Spoken Language Understanding Resource Package, in: *Proc. Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020, pp. 7252–7262.

[6] F. Camastra and A. Vinciarelli, Estimating the intrinsic dimension of data with a fractal-based method, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **24**(10) (2002), 1404–1407.

[7] N. Capuano, Transfer learning techniques for cross-domain analysis of posts in massive educational forums, in: *Intelligent Systems and Learning Data Analytics in Online Education*, Academic Press, 2021, pp. 133–152.

[8] A. Ceni, P. Ashwin and L. Livi, Interpreting Recurrent Neural Networks Behaviour via Excitable Network Attractors, *Cognitive Computation* **12**(2) (2020), 330–356.

[9] K. Cho, B. Van Merrienboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk and Y. Bengio, Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation, in: *Proc. Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2014, pp. 1724–1734.

[10] A. Coucke, A. Saade, A. Ball, T. Bluche, A. Caulier, D. Leroy, C. Doumouro, T. Gisselbrecht, F. Caltagirone, T. Lavril, M. Primet and J. Dureau, Snips Voice Platform: an embedded Spoken Language Understanding system for private-by-design voice interfaces, 2018.

[11] P. Dangeti, *Statistics for machine learning: techniques for exploring supervised, unsupervised, and reinforcement learning models with Python and R*, O'Reilly, 2017.

[12] M. Ester, H.-P. Kriegel, J. Sander and X. Xiaowei, A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise, in: *Proc. of the International Conference on Knowledge Discovery and Data Mining (KDD)*, 1996, pp. 226–231.

[13] J. FitzGerald, C. Hench, C. Peris, S. Mackie, K. Rottmann, A. Sanchez, A. Nash, L. Urbach, V. Kakarala, R. Singh, S. Ranganath, L. Crist, M. Britan, W. Leeuwis, G. Tur and P. Natarajan, MASSIVE: A 1M-Example Multilingual Natural Language Understanding Dataset with 51 Typologically-Diverse Languages, 2022.

[14] M. Golub and D. Sussillo, FixedPointFinder: A Tensorflow toolbox for identifying and characterizing fixed points in recurrent neural networks, *Journal of Open Source Software* **3**(31) (2018), 1003.

[15] I. Goodfellow, Y. Bengio and A. Courville, *Deep learning*, Adaptive computation and machine learning, The MIT Press, Cambridge, Massachusetts, 2016.

[16] D. Hakkani-Tür, G. Tur, A. Celikyilmaz, Y.-N. Chen, J. Gao, L. Deng and Y.-Y. Wang, Multi-Domain Joint Semantic Frame Parsing Using Bi-Directional RNN-LSTM, in: *Proc. InterSpeech*, 2016, pp. 715–719.

[17] R. Hamon, H. Junklewitz and I. Sanchez, Robustness and explainability of Artificial Intelligence: from technical to policy solutions., Technical Report, Joint Research Center. European Commission, 2020.