ESTUDO DA TÉCNICA GENERATIVA *NORMALISING FLOWS* PARA SÍNTESE DE GESTOS GUIADA POR ÁUDIOS DE FALA

Palavras-Chave: *Normalising Flows, Deep Learning*, Modelos generativos, Gestos

Autores(as):
RITA BRAGA SOARES DA SILVA, IMECC – UNICAMP
Profª. Drª. PAULA DORNHOFER PARO COSTA (orientadora), FEEC - UNICAMP
Me. RODOLFO TONOLI (co-orientador), FEEC - UNICAMP

INTRODUÇÃO:

Modelos generativos, baseados em modelos de aprendizado de máquina profundo ou *deep learning*, permitem gerar dados sintéticos a partir de uma distribuição aprendida a partir de dados reais, como imagens, textos, áudios ou vídeos. Esses modelos têm aplicações em diversas áreas como tradução automática, geração de texto, de imagens e, de particular interesse do grupo de pesquisas no qual o presente projeto se insere, síntese de fala e síntese de gestos.

Um dos grandes desafios da área de pesquisas em modelos generativos é a formação de jovens pesquisadores. Isso ocorre porque a compreensão das diferentes abordagens adotadas por modelagens generativas requer muito mais que uma sólida base matemática, exigindo também uma excursão por tópicos avançados de modelagem probabilística. Por outro lado, as aplicações de modelos generativos, tais como a síntese de fala a partir de texto, ou a síntese de gestos a partir de áudio, requerem experiência com o processamento de dados multimodais.

Pesquisas recentes têm se concentrado em modelos generativos para produzir gestos automaticamente, sem intervenção humana, quando um novo conteúdo, como fala, é detectado. Trabalhos como o de Fares, Pelachaud e Obin (2022) procuram melhorar a qualidade do gesto gerado, aproveitando-se de inputs multimodais, com modelos generativos que usam tanto características de texto quanto de fala. Esses modelos dependem de entradas como áudio de fala ou texto para realizar um mapeamento do discurso oral para comportamentos não-verbais e usam dados de captura de movimento para orientar o processo de geração (Vide Figura 1).

O projeto de iniciação científica aqui apresentado tem por escopo a análise da técnica generativa conhecida como “Normalising Flows” na síntese de gestos com base em áudio da fala, visando a concepção de agentes conversacionais personificados, utilizando o artigo de Alexanderson et al. (2020) como referência. Os objetivos abrangem a compreensão da problemática, a edificação de uma base sólida em modelos generativos, a especialização no estudo da abordagem referida como “Normalising Flows”.

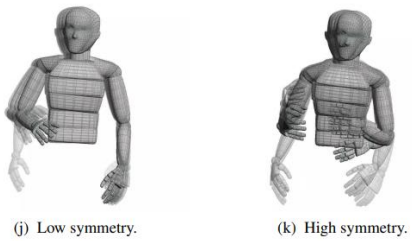

Figura 1 - Gestos performados por Agentes Conversacionais gerados por um modelo de deep learning. Fonte: (ALEXANDERSON et al., 2020)

METODOLOGIA:

O projeto de iniciação científica está sendo desenvolvido utilizando os materiais e infraestrutura já disponíveis no DCA/FEEC, incluindo um sistema computacional compatível com as ferramentas necessárias. A metodologia adotada foi estruturada em algumas fases.

A primeira fase tem como objetivo a integração ao ecossistema do tema de pesquisa e a construção de uma base teórica sólida sobre o tema. A partir de um estudo aprofundado sobre a geração de gestos através da fala, dos trabalhos relevantes como os de Kucherenko et al. (2019), Nyatsanga et al. (2023) e Alexanderson et al. (2020). Além disso, os conceitos matemáticos essenciais e referências básicas sobre redes generativas profundas serão estudados, utilizando capítulos de Tomczak (2022). A partir desses trabalhos, citações importantes também serão analisadas, utilizando a abordagem de "bola de neve". Essa técnica envolve a análise das referências citadas em um artigo e o rastreamento dessas citações para identificar outras fontes pertinentes. O processo é repetido com os novos artigos encontrados, criando um efeito de "bola de neve" onde uma referência leva a outra, ampliando significativamente o conjunto de literatura relevante (WOHLIN, 2014).

Na fase seguinte, o ambiente de desenvolvimento será analisado e com a replicação de experimentos existentes. Esta fase inclui o estudo do repositório disponível em https://github.com/simonalexanderson/StyleGestures e a replicação do ambiente de desenvolvimento para tentar reproduzir os resultados originais. O treinamento na ferramenta de desenvolvimento PyTorch (PASZKE et al., 2019) também é parte fundamental desta fase.

A fase final será dedicada à documentação e disseminação dos resultados, contando com uma documentação detalhada do uso da técnica e possíveis adaptações para diferentes experimentos. Com essa metodologia, o projeto garante uma abordagem completa, desde a integração inicial e estudo teórico até a implementação prática e documentação, proporcionando uma formação abrangente e aprofundada no tema de estudo.

RESULTADOS E DISCUSSÃO:

Para os estudos iniciais, foram realizadas leituras de artigos e materiais relacionados aos principais tópicos que norteiam o artigo de referência de Alexanderson et al. (2020).

MoGlow (HENTER; ALEXANDERSON; BESKOW, 2020) que é um modelo de síntese de movimento probabilístico e controlável, estende o conceito de Glow (KINGMA; DHARIWAL, 2018), para a síntese de movimento. Sendo assim, foi um grande tópico de estudo para este projeto, uma vez que modela a distribuição condicional dos movimentos dado um condicionamento (como características de áudio) usando uma sequência de *Normalising Flows*.

O artigo de referência também propõe uma abordagem independente do nível de abstração do espaço de controle desejado, chamada de controle de estilo. Outros trabalhos, como os de Yang et al. (2023) e Yoon et al. (2021), foram estudados para entender outras técnicas de controle de estilo de gestos, que são o grande diferencial do modelo. Para avaliar a performance do movimento, o artigo de Alexanderson et al. (2020) utiliza participantes humanos para avaliar a semelhança humana, a adequação do movimento, os efeitos do controle de estilo e a síntese de corpo inteiro. O trabalho de revisão de Wolfert, Robinson e Belpaeme (2022) foi estudado para entender melhor as diversas abordagens de avaliação de gestos e movimento, destacando a importância de métodos de avaliação uniformes. Além disso, para entender o processamento de sinais de fala e as variáveis acústicas, foram estudados artigos de revisão como o de Adiga e Prasanna (2019) e bibliotecas como `torchaudio` e `librosa` para processamento de sinal e áudio. Dessa forma, foi possível construir uma base inicial de formação teórica para o entendimento do problema.

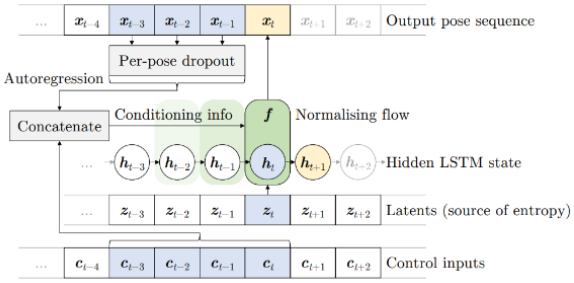

Figura 2 Arquitetura MoGlow. Fonte: (HENTER; ALEXANDERSON; BESKOW, 2020)

Para familiarização com a parte prática, foram utilizados repositórios disponíveis online de artigos sobre *Normalising Flows* para entender seu funcionamento geral como os trabalhos de Rezende e Mohamed (2015), Wehenkel e Louppe (2021), Papamakarios, Pavlakou e Murray (2017) e Cao, Aziz e Titov (2020). Note que, o repositório *StyleGestures* é referência a dois artigos: "MoGlow: Probabilistic and controllable motion synthesis using normalising flows" e "Style-controllable speech-driven gesture synthesis using normalising flows". Mesmo que este projeto tenha como principal referência um dos artigos, ambos se fazem necessários para o entendimento completo do código e sua aplicação. Isso se deve ao fato de que a arquitetura do MoGlow (vide Figura 2) é muito semelhante à do modelo apresentado no artigo de referência (vide Figura 3). Essa base compartilhada reforça a conexão entre os dois trabalhos, demonstrando como os avanços e técnicas desenvolvidas em um artigo foram adaptados e expandidos no outro. Portanto, o estudo e a replicação dos resultados exigem uma compreensão integrada dos dois artigos.

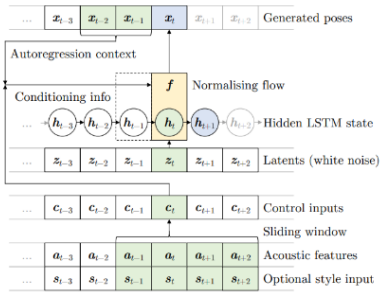

Figura 3 - Arquitetura do modelo do artigo de referência. Fonte: (ALEXANDERSON et al., 2020)

É importante destacar que o ambiente (enviroment) do repositório não está bem documentado e isso dificulta a aplicação prática, visto que a combinação das versões certas de cada pacote é essencial para que o código consiga rodar corretamente. A ausência de versões específicas dos pacotes pode causar vários problemas, como: (a) a incompatibilidade de versões, sem especificar as versões, o conda instalará as versões mais recentes dos pacotes disponíveis. No entanto, versões mais recentes podem introduzir mudanças que não são compatíveis com o código existente, resultando em erros de execução ou comportamento inesperado; (b) dependências transitivas, que também variam de versão, e (c) a própria reprodutibilidade, sem versões específicas, reproduzir o ambiente exato em que os resultados originais foram obtidos se torna extremamente difícil, o que compromete a validade dos experimentos a serem replicados. Para mitigar esses problemas, foi necessário um esforço considerável para identificar as versões corretas dos pacotes que permitiriam a execução do código conforme esperado. Esse processo envolveu a consulta de documentações adicionais, a realização de testes iterativos e a colaboração com outros pesquisadores para alinhar as versões das bibliotecas utilizadas. Por exemplo, determinar a versão correta do PyTorch que é compatível com o cudatoolkit=10.2 exigiu múltiplas tentativas até que a combinação funcional fosse encontrada. No entanto, a própria comunidade do GitHub, plataforma em que o código está disponível publicamente, forneceu indicativos cruciais sobre a versão correta do PyTorch (1.5) e do Python (3.7), o que facilitou significativamente os testes subsequentes de versões.

CONCLUSÕES:

É observado que a falta de documentação adequada do ambiente do repositório dificulta a execução correta do código devido à necessidade de combinações específicas de versões dos pacotes. A ausência dessas informações causa problemas de incompatibilidade e compromete a reprodutibilidade dos experimentos.

A próxima fase da pesquisa consiste na tentativa de reprodução dos resultados originais para verificar a replicabilidade e validar os resultados anteriores. Esse processo visa assegurar a eficácia dos métodos empregados e a validade das descobertas em um novo contexto linguístico, se a adaptação se mostrar possível.

BIBLIOGRAFIA

ADIGA, N.; PRASANNA, S. R. M. **Acoustic features modelling for statistical parametric speech synthesis: A review**. IETE Technical Review, Taylor Francis, v. 36, n. 2, p. 130–149, 2019.

ALEXANDERSON, S. et al. **Style-controllable speech-driven gesture synthesis using normalising flows**. In: WILEY ONLINE LIBRARY. Computer Graphics Forum. [S.l.], v. 39, n. 2, p. 487–496, 2020.

CAO, N. D.; AZIZ, W.; TITOV, I. **Block neural autoregressive flow. In: PMLR. Uncertainty in artificial intelligence**. [S.l.], p. 1263–1273, 2020.

FARES, M.; PELACHAUD, C.; OBIN, N. **Transformer Network for Semantically- Aware and Speech-Driven Upper-Face Generation**. 2022.

HENTER, G. E.; ALEXANDERSON, S.; BESKOW, J. **Moglow: Probabilistic and controllable motion synthesis using normalising flows**. ACM Transactions on Graphics (TOG), ACM New York, NY, USA, v. 39, n. 6, p. 1–14, 2020.

KINGMA, D. P.; DHARIWAL, P. **Glow: Generative flow with invertible 1x1 convolutions**. Advances in neural information processing systems, v. 31, 2018.

KUCHERENKO, T. et al. **Analyzing input and output representations for speech-driven gesture generation.** In: Proceedings of the 19th ACM International Conference on Intelligent Virtual Agents. [S.l.: s.n.], p. 97–104. 2019.

NYATSANGA, S. et al. **A comprehensive review of data-driven co-speech gesture generation**. arXiv preprint arXiv:2301.05339, 2023.

PAPAMAKARIOS, G.; PAVLAKOU, T.; MURRAY, I. **Masked autoregressive flow or density estimation. Advances in neural information processing systems**, v. 30, 2017.

PASZKE, A. et al. Pytorch: **An imperative style, high-performance deep learning library**. In: Advances in Neural Information Processing Systems 32. Curran Associates, Inc., p. 8024–803. 2019.

REZENDE, D.; MOHAMED, S. **Variational inference with normalizing flows**. In: PMLR. International conference on machine learning. [S.l.], p. 1530–1538. 2015.

TOMCZAK, J. M. **Deep generative modeling**. [S.l.]: Springer, 2022.

WEHENKEL, A.; LOUPPE, G. **Graphical normalizing flows**. In: PMLR. International Conference on Artificial Intelligence and Statistics. [S.l.], p. 37–45. 2021.

WOHLIN, C. **Guidelines for snowballing in systematic literature studies and a replication in software engineering**. In Proceedings of the 18th International Conference on Evaluation and Assessment in Software Engineering. Association for Computing Machinery, USA, 38, 1–10. 2014.

WOLFERT, P.; ROBINSON, N.; BELPAEME, T. **A review of evaluation practices of gesture generation in embodied conversational agents.** IEEE Transactions on Human-Machine Systems, v. 52, n. 3, p. 379–389, 2022

YANG, S. et al. Diffusestylegesture: **Stylized audio-driven co-speech gesture Generation with diffusion models.** arXiv preprint arXiv:2305.04919, 2023.

YOON, Y. et al. Sgtoolkit: **An interactive gesture authoring toolkit for embodied conversational agents**. In: The 34th Annual ACM Symposium on User Interface Software and Technology. New York, NY, USA: Association for Computing Machinery, (UIST '21), p. 826–840. ISBN 9781450386357. Disponível em: <https://doi.org/10.1145/3472749.3474789>. 2021.