Projeto de Formatura – 2024

PCS - Departamento de Engenharia de Computação e Sistemas Digitais

Engenharia de Computação

Tema: **Análise Formal de Segurança Crítica (*safety*) de Inteligência Artificial em Sistemas de Detecção de Arritmias Cardíacas: Diferenciação entre Batimentos Saudáveis e Não Saudáveis.**

Introdução

O objetivo do trabalho é realizar a **avaliação formal de segurança crítica (*safety*)** de **Sistemas de Detecção de Arritmias Cardíacas (SDACs)** baseados em redes neurais através da **diferenciação do domínio dos batimentos cardíacos classificados como saudáveis do domínio das classificações de arritmia**. Com essa avaliação, é esperado que haja **maior previsibilidade das respostas do sistema** e que ela alavanque seu uso em aplicações críticas em relação a segurança, como em **diagnóstico médico e marca-passos**.

Método de Trabalho

1 Pesquisa Bibliográfica

Inicialmente, foi realizada uma investigação do estado da arte de detecção de arritmias cardíacas utilizando técnicas de IA para embasar teoricamente a pesquisa. Após isso, foi realizada outra investigação sobre o estado da arte das ferramentas de verificação formal de Redes Neurais para o planejamento da arquitetura da solução proposta.

2 Arquitetura da Solução Proposta

Com base nos achados da pesquisas bibliográfica e da análise comparativa de ferramentas, foi definida uma arquitetura de solução baseada em três etapas: Tradução, Sobreaproximação e Verificação. Além disso, foram definidas as ferramentas relevantes para o desenvolvimento da pesquisa.

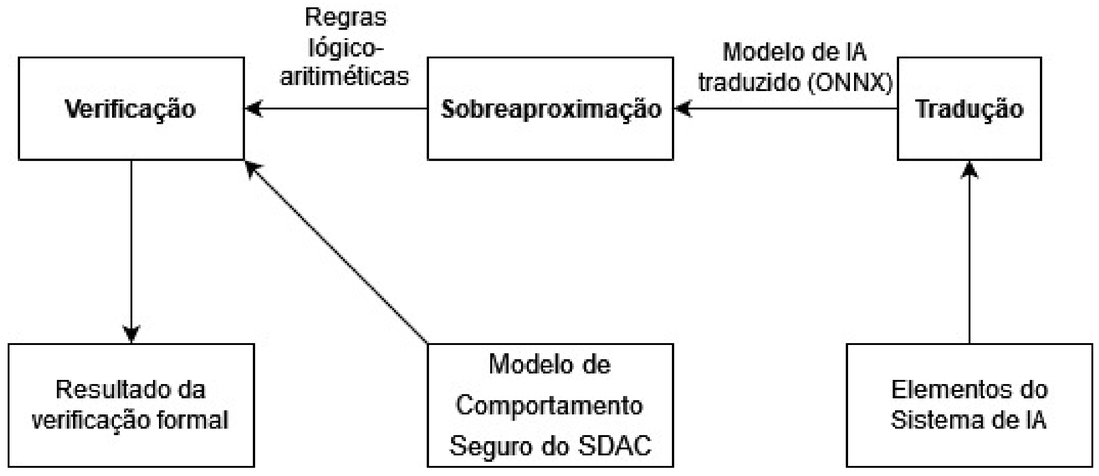
3 Implementação do Processo de Tradução

Utilizando o arcabouço ***Tf2Keras***, foi implementada a tradução do sistema de redes neurais para a especificação **ONNX** (*Open Neural Network Exchange*), que é utilizada como padrão *de facto* no estado da arte de verificação de redes neurais. O sistema traduzido foi confirmado através de uma comparação tripla, validando que ele é equivalente ao sistema alvo da análise.

4 Verificação baseada em Ataques Adversários

Com base na ferramenta *alpha-beta-crown*, foi realizada uma verificação de comportamento da rede neural baseada em ataques adversários.

Arquitetura da Solução



A verificação formal é composta por **três etapas**:

- **Tradução:** Nesta etapa é realizada a tradução do sistema de redes neurais da especificação original em Keras para a especificação ONNX. A necessidade desse passo dá-se pela compatibilidade da ferramenta de verificação formal com essa especificação.
- **Sobreaproximação:** O sistema de redes é sobreaproximado por funções matemáticas que permitam a investigação da verificação formal de parâmetros de segurança crítica nelas.
- **Verificação:** Por último, através da sobreaproximação realizada, é verificado se o sistema satisfaz a modelagem de comportamento seguro mediante análise de seu conjunto imagem.

Considerações sobre Sobreaproximação

Verificou-se que a ferramenta *alpha-beta-crown* possui restrições que impedem a sobreaproximação da rede neural alvo da análise. Essas restrições envolvem limitações em hiperparâmetros de algumas camadas da rede.

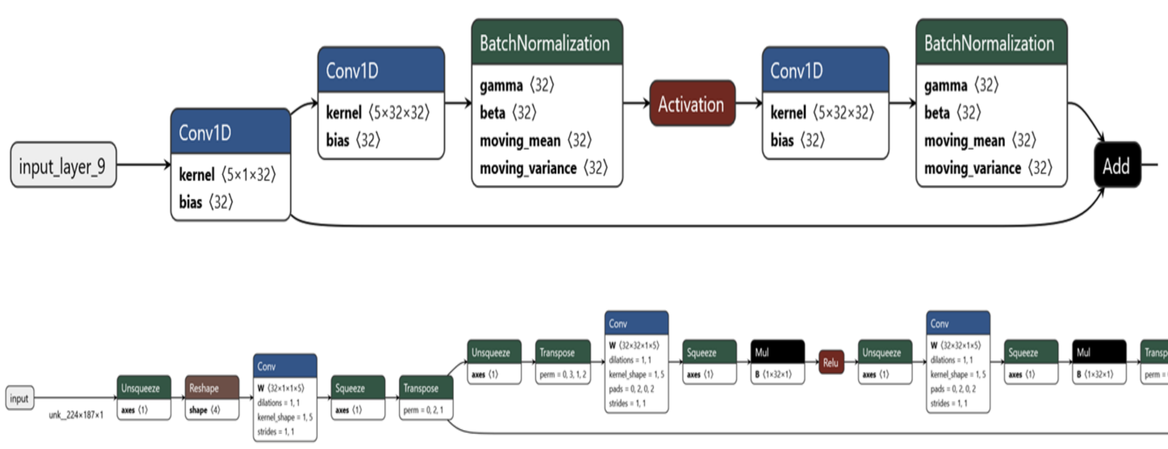
A estratégia adotada no lugar da sobreaproximação foi através da estimulação da rede por ataques adversários.

Ataques Adversários

A estratégia se baseia em um algoritmo de *Projected Gradient Descent* (PGD) para buscar pela menor perturbação na entrada da rede que provoca uma confusão na classificação.

Tradução do Sistema de Redes Neurais

Essa etapa do projeto foi implementada traduzindo a rede em especificação Keras para **ONNX**. Para verificar a correspondências dos modelos, foram avaliadas a arquitetura das redes neurais, sua resposta a exemplos da base de dados e os pesos atribuídos aos neurônios através dos tensores.



Conclusão

Os resultados obtidos pavimentam a verificação formal de segurança do sistema de detecção de arritmias cardíacas, mas não são suficientes para esse fim. Isso se deve às limitações da ferramenta alpha-beta-crown e dos ataques adversários. Para trabalhos futuros em **Mestrado**, consideram-se: (i.) ajustes para análise formal da rede neural, (ii.) expansão para arritmias e (iii.) comparação com normas de segurança e equipamentos comerciais.

Integrante: Gabriel Stephano Santos

Professor Orientador: Prof. Dr. Paulo Sérgio Cugnasca

Coorientador: Prof. Dr. Antonio Vieira da Silva Neto

Apoio: PCS Departamento de Engenharia de Computação e Sistemas Digitais · POLI USP · c²d centro de ciência de dados · FUSP · GAS Grupo de Análise de Segurança · Escola Politécnica da USP · itaú