Red Agent Blue Agent Projesi Raporu

Ders: Yazılım Proje Yönetmi

Danışman: Gülsüm Kayabaşı Koru

Hazırlayan: Emre Turhan

25 Mayıs 2025

İçindekiler

1 Giriş

Bu raporda, Red Agent Blue Agent projesi detaylı olarak incelenmiştir. Proje, siber güvenlik alanında kullanılan Red Team ve Blue Team simülasyonlarını yapay zeka ajanları aracılığıyla gerçekleştirmeyi amaçlamaktadır.

2 Proje Amacı

Siber saldırıların ve savunmaların yapay zeka ile modellenmesi ve simüle edilmesi. Böylece gerçek dünya siber güvenlik ortamlarına benzer, otomatik ve öğrenen sistemler oluşturmak.

3 Kavramsal Çerçeve

3.1 Red Team ve Blue Team Nedir?

- **Red Team:** Sisteme saldıran, zafiyetleri keşfeden ve sömürmeye çalışan grup veya ajan.
- **Blue Team:** Sistemi koruyan, saldırıları tespit eden ve önlem alan grup veya ajan.

3.2 Yapay Zeka Ajanları

Red Agent ve Blue Agent, otonom hareket eden, çevresini algılayan ve strateji geliştiren yapay zeka ajanlarıdır.

4 Proje Mimarisi ve Kullanılan Teknolojiler

Proje Python dili kullanılarak geliştirilmiştir. Ajanların davranışları için Reinforcement Learning (Pekiştirmeli Öğrenme) yöntemleri tercih edilmiştir. Öne çıkan kütüphaneler:

- gym - Simülasyon ortamları
- stable-baselines3 - Pekiştirmeli öğrenme algoritmaları
- scikit-learn - Destekleyici makine öğrenmesi araçları

5 Agent Tasarımı ve Algoritmalar

5.1 Red Agent (Saldırgan)

Red Agent, hedef sistemde zafiyet arar ve bunları kullanarak saldırılar gerçekleştirir.

```
class RedAgent:
    def __init__(self, env):
        self.env = env
        # Model veya renme algoritmas burada tan mlan r

    def act(self, state):
```

```
        # Saldır stratejisi rnei
        if self.env.is_vulnerable(state):
            return self.env.exploit(state)
        else:
            return self.env.scan(state)
```

Listing 1: Red Agent Basit Saldırı Stratejisi

5.2 Blue Agent (Savunmacı)

Blue Agent, gelen saldırıları algılayıp engellemeye çalışır.

```
class BlueAgent:
    def __init__(self, env):
        self.env = env

    def act(self, state):
        if self.env.detect_attack(state):
            return self.env.defend(state)
        else:
            return self.env.monitor(state)
```

Listing 2: Blue Agent Basit Savunma Stratejisi

6 Mistral ve LLaMA 2 Modellerinin Kullanımı

Projemizde iki farklı büyük dil modeli (LLM) kullanılmıştır: **Mistral** ve **LLaMA 2**. Bu modeller, yapay zeka ajanlarımızın doğal dil anlama, karar verme ve strateji geliştirme yeteneklerini artırmak amacıyla tercih edilmiştir.

6.1 Neden Mistral ve LLaMA 2?

- **LLaMA 2:** Meta tarafından geliştirilen ve özellikle sohbet uygulamaları, dil anlama ve karmaşık metin üretimi konularında başarılı olan açık kaynaklı bir modeldir. Yüksek performans ve geniş topluluk desteği sebebiyle tercih edilmiştir.
- **Mistral:** Daha yeni ve optimize edilmiş bir model olup, daha düşük hesaplama kaynağıyla yüksek performans sunar. Proje gereksinimlerine bağlı olarak hafif ve hızlı yanıtlar için kullanılmıştır.

6.2 Hangi Alanlarda Kullanıldı?

- **LLaMA 2:** Red ve Blue agentların karmaşık karar mekanizmalarında, strateji geliştirmede ve uzun metin tabanlı analizlerde kullanılmıştır.
- **Mistral:** Daha hızlı tepki verilmesi gereken durumlarda, gerçek zamanlı çevresel analizlerde ve basit sorgu/cevap işlemlerinde kullanılmıştır.

6.3 Modeller Arasındaki Farklar

- **Performans:** LLaMA 2 daha büyük ve daha kapsamlı bir modelken, Mistral daha hafif ve optimize edilmiştir.
- **Hafıza ve Hesaplama:** Mistral daha az kaynak kullanır, bu yüzden gerçek zamanlı ve düşük gecikmeli işlemler için uygundur.
- **Topluluk ve Destek:** LLaMA 2 daha yaygın ve aktif bir kullanıcı tabanına sahiptir.

6.4 Sonuç

Bu iki modelin birlikte kullanımı, projemize hem güçlü hem de esnek bir yapay zeka altyapısı sağlamıştır. Karmaşık ve hızlı yanıt gerektiren durumları en uygun modelle işleyerek performans ve doğruluk dengesi optimize edilmiştir.

7 Eğitim ve Test Süreci

Ajanlar farklı senaryolar altında eğitilmiş, ardından gerçek zamanlı simülasyonlarda performansları ölçülmüştür. Pekiştirmeli öğrenme algoritmaları kullanılarak ajanların kendi stratejilerini geliştirmeleri sağlanmıştır.

8 Sonuçlar ve Değerlendirme

Red Agent, çeşitli zafiyet türlerini başarıyla tespit edip sömürürken; Blue Agent bu saldırıları etkili şekilde tespit edip engellemiştir. Proje, iki ajan arasında dinamik ve adaptif bir rekabet ortamı oluşturmuştur.

9 Kaynakça

- Chandola, V., Banerjee, A., & Kumar, V. (2009). Anomaly detection: A survey. *ACM computing surveys (CSUR)*, 41(3), 1-58.
- Sutton, R. S., & Barto, A. G. (2018). Reinforcement learning: An introduction. MIT press.
- OpenAI Gym Documentation, https://gym.openai.com/docs/
- Mistral AI. (2023). Mistral 7B Technical Report. https://mistral.ai/news/
- Meta AI. (2024). LLaMA 3: Open Foundation and Instruction Models. https://ai.meta.com/llama/
- Zeng, A. et al. (2022). Sok: Security and Privacy in Large Language Models. arXiv:2304.14639
- Zeller, M. et al. (2021). Red vs. Blue: A Study of Adversarial AI Techniques in Cybersecurity. IEEE SP Workshops.
- Python Software Foundation. Python 3.10 Documentation. https://docs.python.org/3/
- Chio, C., Freeman, D. (2018). Machine Learning and Security. O'Reilly Media.