جامعة المنوفية

كلية الهندسة الإلكترونية

قسم هندسة و علوم الحاسبات

تأمين حقوق النشر لنماذج الشبكات العصبية العميقة المدربة باستخدام العلامات المائية الرقمية

رسالة مقدمة للحصول على درجة الدكتوراه في العلوم الهندسية

تخصص اختر القسم باللغة الإنجليزية

مجال الرسالة: علوم الحاسب

قسم اختر القسم باللغة الإنجليزية

من المهندسة

آلاء محمد أحمد فكيرين هلال

بكالوريوس الهندسة الإلكترونية – قسم هندسة و علوم الحاسبات – كلية الهندسة الإلكترونية بمنوف – جامعة المنوفية 2011

ماجستير العلوم الهندسية – قسم هندسة و علوم الحاسبات

كلية الهندسة الإلكترونية بمنوف – جامعة المنوفية 2018 –

لجنة الإشراف

أ.د. أيمن السيد أحمد عميره

أستاذ بقسم اختر القسم باللغة الإنجليزية

كلية الهندسة الإلكترونية – جامعة المنوفية

أ.د. جمال محروس علي عطيه

أستاذ بقسم هندسة و علوم الحاسبات

كلية الهندسة الإلكترونية – جامعة المنوفية

د. مروه أحمد شومان

أستاذ مساعد بقسم هندسة و علوم الحاسبات

كلية الهندسة الإلكترونية – جامعة المنوفية

أيلول 2024

تأمين حقوق النشر لنماذج الشبكات العصبية العميقة المدربة باستخدام العلامات المائية الرقمية

الملخص

في الآونة الأخيرة، حققت تقنيات التعلم العميق مستويات مذهلة من الدقة وتُستخدم في مجالات مهمة مثل الرعاية الصحية، السيارات ذاتية القيادة، ومعالجة اللغات. تدريب الشبكات العصبية العميقة (DNNs) يتطلب الكثير من الوقت، البيانات، والقدرة الحاسوبية. نتيجة لذلك، تُباع النماذج المدربة مسبقًا، لكنها تكون عرضة للنسخ والمشاركة غير المصرح بها. تستكشف هذه الأطروحة استخدام العلامات المائية الرقمية لحماية نماذج الشبكات العصبية العميقة من النسخ غير المصرح به.

بعد مقارنة الطرق المختلفة لتحسين الدقة، توصلت الأطروحة إلى أن المحسن Stochastic Gradient Descent (SGD)هو الأكثر فعالية.

تم أيضا اقتراح نظام حماية هجين ذو مستويين، حيث تم تقديم خمس مقترحات متسلسلة وتم اختباره ضد عدة هجمات. أظهرت النتائج أن هذا النظام ينجح في حماية النماذج ويصمد أمام أنواع مختلفة من الهجمات، متفوقًا على الأساليب الأخرى الموجودة.

يمكن تلخيص الإسهامات الرئيسية للأطروحة كما يلي :اقترحنا استراتيجيتين لتحقيق هذا الهدف .

يمكن تلخيص الفكرة الأساسية للاستراتيجية الأولى في ثلاثة عناصر:

1. إجراء دراسة مقارنة للتقنيات الحديثة التي تركز على ضمان حماية حقوق الملكية للنماذج العصبية العميقة.(DNNs)
2. تقديم تحسين في الدقة.

3. تقييم عدة تجارب على إطار العمل المقترح باستخدام مجموعتين مختلفتين من البيانات MNISTوCIFAR10-CNN

أما مفهوم الاستراتيجية الثانية فيمكن تلخيصه في ثلاثة عناصر رئيسية:

1. استخدام الهجمات العدائية كعلامات مائية لحماية ملكية نماذج الشبكات العصبية العميقة.
2. إنشاء نظام حماية هجين ذو مستويين، لضمان بقاء أحد المستويات صامدًا في حال فشل الآخر. تم تطوير هذا النظام من خلال تطبيق خمس مقترحات متسلسلة.
3. تقييم نظام العلامات المائية عن طريق تعريضه لسبعة أنواع مختلفة من الهجمات: هجوم طريقة التدرج السريع، هجوم تدرج الإسقاط التلقائي، هجوم التدرج التلقائي المترافق، هجوم الطريقة التكرارية الأساسية، هجوم التدرج التكراري باستخدام الزخم، الهجوم المربع، والهجوم التلقائي.

العمل المقدم في هذه الأطروحة منظم في خمسة فصول، كما يلي :

الفصل الأول يقدم مقدمة عامة للأطروحة، يشرح أهمية موضوعها والغرض منها، بالإضافة إلى الشكل التنظيمي لبقية الأطروحة .

الفصل الثاني يقدم الدراسات السابقة، تبدأ بمقدمة عن التعلم العميق والأبحاث ذات الصلة، ثم يتناول العلامات المائية الرقمية، ويختتم بمناقشة تضمين العلامات المائية في الشبكات العصبية العميقة (DNNs) لحمايتها من الاستخدام غير المصرح به .

الفصل الثالث يناقش العلامات المائية الرقمية كوسيلة لحماية نماذج الشبكات العصبية العميقة (DNN) والحفاظ على الملكية الفكرية. كما يقدم الفصل تحليلًا مقارنًا للتقنيات الحديثة في العلامات المائية، بالإضافة إلى دراسة حول تأثير المحسنات المختلفة على دقة النماذج، بناءً على تجارب مع مجموعتي بيانات MNIST و CIFAR10

الفصل الرابع يقدم نظام حماية هجين ذو مستويين لحماية نماذج الشبكات العصبية العميقة المدربة مسبقًا من التوزيع غير المصرح به. يضمن النظام الحماية والقوة من خلال تقديم مستويين من الحماية، حيث يبقى أحد المستويين في حال فشل الآخر. يشمل النظام خمس مقترحات متسلسلة: أولًا، تضمين الهجمات العدائية، ثانيًا، إعادة تصنيف العينات، ثالثًا، تطبيق التقليم (Pruning)، رابعًا، تحسين حماية المستوى الثاني ضد الهجمات، خامسًا، تحسين حماية المستوى الأول ضد الهجمات. وأخيرًا، اختبار هذا النظام الهجين ذو المستويين ضد سبعة أنواع من الهجمات. تم اثبات ان النظام ذو كفاءة عالية في حماية النماذج .

الفصل الخامس يختتم الأطروحة، يقدم اقتراحات ويضيف نقاط بحث مفتوحة للعمل المستقبلي.