جمهورية العراق
وزارة التعليم العالي والبحث العلمي
جامعة تكريت
كلية التربية للعلوم الإنسانية
قسم الجغرافية

دور الذكاء الاصطناعي في الهواتف الذكية الحديثة – المرحلة الثانية

م .م رنا مزاحم جهاد

١٤٤٧هـ ٢٠٢٥م- ٢٠٢٦م

المبحث الأول: المقدمة والخلفية العلمية للحوسبة الإدراكية المحمولة

١.١ تحول جذري

شهدت الأنظمة المدمجة في الهواتف الذكية تحولاً جذرياً، حيث انتقلت من دور المعالجة الخطية للبيانات إلى دور الأنظمة الإدراكية التكيفية. تاريخياً، كانت خوارزميات الذكاء الاصطناعي المعقدة، مثل الشبكات العصبية العميقة، تُنفذ حصرياً على خوادم سحابية عملاقة تمتلك قدرات حوسبية وطاقة غير محدودة. ومع ذلك، فإن هذا النموذج السحابي واجه ثلاثة تحديات بنيوية جعلت استمراره في الهواتف الذكية مستحيلاً:

- **زمن الاستجابة** :تتطلب تطبيقات مثل الواقع المعزز والترجمة الفورية زمن استجابة منخفضاً للغاية، وهو ما لا يمكن ضمانه عبر شبكات الاتصال اللاسلكية التقليدية بسبب تقلبات النطاق الترددي للشبكة.
- **استهلاك الطاقة والبيانات** :إرسال تدفقات مستمرة من البيانات العالية الدقة (مثل فيديو عالي الدقة بمعدل إطارات سريع) إلى السحاب يستهلك طاقة الإرسال اللاسلكي وعبر شبكات الهاتف بشكل يستنزف البطارية بكثافة.
- **الخصوصية والأمن** :القوانين التشريعية الحديثة لحماية البيانات والنزعة العامة للمستخدمين ترفض رفع البيانات الحيوية والشخصية كبصمات الوجه والصوت إلى خوادم خارجية.

من هنا برزت الحاجة إلى نقل الذكاء الاصطناعي إلى الحافة، وهو ما يُعرف بـ **الذكاء الاصطناعي الطرفي(Edge AI)** ، حيث تُجرى عمليات الاستدلال والتوصيف محلياً بالكامل على الجهاز.

١.٢صياغة المشكلة البحثية

تكمن المشكلة البحثية الأساسية في كيفية موازنة "مفارقة العتاد والطاقة". إن الشبكات العصبية الحديثة تتطلب بليارات العمليات الحسابية وتستهلك سعة ذاكرة ضخمة، في حين أن الهاتف الذكي محكوم ببيئة ذات موارد مقيدة للغاية:

- البطارية محدودة السعة وتخضع لقيود الحجم الفيزيائي للهاتف.

- تبديد الحرارة يعتمد على التبريد الساكن دون مراوح، مما يضع حداً حرارياً صارماً لا يمكن تجاوزه لتجنب تلف الشريحة أو اختناق الأداء.

المبحث الثاني: هندسة العتاد وتسريع الشبكات العصبية المدمجة

٢.١معمارية وحدة المعالجة العصبية(NPU Architecture)

تعتمد وحدات المعالجة المركزية (CPUs) على معمارية معالجة تسلسلية عامة، بينما تعتمد وحدات معالجة الرسوميات (GPUs) على معالجة متوازية واسعة النطاق لكنها تستهلك طاقة مرتفعة. لحل هذه المعضلة، دمجت مصانع أشباه الموصلات **وحدة المعالجة العصبية (NPU)** كجزء أساسي من شريحة النظام.(SoC)

تتكون وحدة المعالجة العصبية في جوهرها من مصفوفة ثنائية الأبعاد من عناصر المعالجة المصممة خصيصاً لتنفيذ عمليات الضرب والتراكم المتكررة بكثافة عالية جداً. هذه المصفوفات تتخلى عن معمارية "فون نيومان" التقليدية لتجنب الانتقال المتكرر للبيانات بين المعالج والذاكرة، وهو ما يعرف بـ "عنق زجاجة الذاكرة."

٢.٢أنماط تدفق البيانات في المعالجات الطرفية

لتوفير الطاقة الناتجة عن قراءة وكتابة الأوزان والميزات من والى الذاكرة العشوائية الخارجية، تعتمد المعالجات الحديثة على استراتيجيات تدفق بيانات ذكية:

- **نمط تثبيت الأوزان :(Weight-Stationary)** حيث يتم شحن أوزان الشبكة العصبية داخل خلايا الذاكرة المحلية السريعة والقريبة جداً من عناصر المعالجة وتثبيتها هناك، بينما تتدفق ميزات الصورة أو النص الجاري معالجته عبر هذه العناصر. هذا النمط يقلل من قراءة الأوزان من الذاكرة الخارجية التي تستهلك طاقة كبيرة.
- **الحوسبة داخل الذاكرة :(In-Memory Computing)** دمج الدوائر المنطقية الحسابية مباشرة داخل بنية خلايا الذاكرة السريعة، بحيث يتم حساب النتائج أثناء قراءة البيانات في نفس الموقع الفيزيائي، محققاً كفاءة طاقة قصوى وقدرة على معالجة تريليونات العمليات في الثانية لكل واط.

المبحث الثالث: الهندسة البنيوية لضغط وتكييف النماذج العصبية

النماذج البحثية الكبيرة يتم تدريبها بدقة عائمة عالية الحجم. لنقل هذه النماذج إلى بيئة الهاتف المحدودة، يجب إخضاعها لعمليات هندسية تقلل من حجمها ومجهودها الحسابي دون التضحية بدقتها.

٣.١تكميم الشبكات العصبية(Model Quantization)

التكميم هو عملية تحويل أوزان الشبكة العصبية وميزاتها من صيغة الأعداد العائمة ذات الحجم الكبير (مثل ٣٢ بت) إلى صيغة الأعداد الصحيحة منخفضة الحجم (مثل ٨ بت أو ٤ بت).

- **التكميم المنتظم الساكن** :يعتمد على مسح نطاق البيانات وتحديد القيم العظمى والصغرى، ثم توزيع القيم العائمة على مستويات رقمية صحيحة محددة وثابتة. يتضمن ذلك حساب عامل قياس ونقطة صفر لضمان عدم تشويه القيم القريبة من الصفر.
- **التكميم الأثنائى والتكميم الفرعي** :في المستويات المتقدمة جداً، يتم ضغط بعض طبقات الشبكة العصبية إلى قيم لا تتعدى بتات معدودة، مما يخفض حجم النموذج الكلي بمعدل يصل إلى ٧٥٪، وهو ما يتيح فك الاختناق عن الذاكرة العشوائية للهاتف ويسرع من عملية الاستدلال بشكل ملحوظ.

المبحث الرابع: التصوير الحاسوبي المتقدم وأنظمة الرؤية الحاسوبية

لم تعد الصورة الملتقطة بكاميرا الهاتف الحديث تعكس الخصائص الفيزيائية البسيطة للمستشعر والعدسة، بل أصبحت نتاجاً لعملية إعادة بناء رقمي وحاسوبي كاملة تعتمد على الرؤية الحاسوبية والتعلم العميق.

٤.١معالج إشارات الصور العصبي(Neural ISP)

في التصوير التقليدي، يمر الضوء عبر العدسة إلى المستشعر، ثم تتولى رقاقة ثابتة معالجة الإشارات عبر خطوات تتابعية جامدة (مثل فك الفسيفساء، موازنة اللون الأبيض، وتقليل الضوضاء). في الهواتف الحديثة، تم استبدال هذا النظام بـ **معالج إشارات صور عصبي**.

- يتم تدريب شبكات عصبية عميقة التلافيف أو شبكات توليدية على ملايين الصور الخام والصور المقابلة لها الملتقطة بكاميرات احترافية.

- يقوم المعالج العصبي بتحويل البيانات الخام القادمة من المستشعر مباشرة إلى صورة نهائية، حيث يقوم بمهام فك الفسيفساء وتقليل الضوضاء الديناميكي وعزل الخلفية في خطوة برمجية موحدة وذكية.

المبحث الخامس: معالجة اللغات الطبيعية وتشغيل النماذج الكبيرة محلياً

انتقلت الهواتف الذكية من مرحلة المساعدات الصوتية القائمة على الأوامر النصية الثابتة والمحددة مسبقاً، إلى مرحلة تشغيل **نماذج اللغات الكبيرة التوليدية (On-Device LLMs)** بشكل محلي كامل ودون الحاجة للاتصال بالإنترنت.

٥.١آليات الانتباه المتقدمة والمصغرة

تعتمد نماذج المحولات (Transformers) على آلية الانتباه الذاتي، والتي تتطلب في حالتها الطبيعية مساحة ذاكرة تتزايد بشكل حاد (تربيعي) مع زيادة طول النص الإدخالي. لتخطي هذا العائق في الهواتف الذكية، يتم تطبيق خوارزميات إدارة الذاكرة المصغرة:

- **خوارزميات الانتباه الوميضي :(FlashAttention)** تقوم بتقسيم مصفوفات الانتباه الضخمة إلى كتل صغيرة جداً وتمريرها مباشرة داخل الذاكرة السريعة للمعالج دون الحاجة لكتابة وقراءة البيانات الوسيطة من الذاكرة العشوائية الرئيسية للهاتف، مما يضمن استمرار توليد النصوص بسرعة وبأقل استهلاك طاقة ممكن.
- **ذاكرة التخزين المؤقت للمفاتيح والقيم :(KV Caching)** يتم تخزين الحسابات السابقة للكلمات في الذاكرة لتجنب إعادة حسابها مع كل كلمة جديدة يتم توليدها، وتخضع هذه الذاكرة لتقنيات ضغط مكثفة لتقليل مساحتها.

٥.٢التوليد المستعاد المعزز الموضعي(Local RAG)

لكي يكون المساعد الذكي نافعاً، يجب أن يمتلك سياقاً حول المستخدم (مواعيده، رسائله، وتفضيلاته). بدلاً من رفع هذه البيانات للسحاب، يقوم الهاتف ببناء **رسم بياني للمعلومات محلي ومحمي.**

- يقوم الذكاء الاصطناعي بتحويل كافة نصوص ورسائل المستخدم المحلية إلى "تضمينات سياقية" (مصفوفات رقمية تفهم المعنى).

- عندما يطرح المستخدم سؤالاً، يبحث النظام محلياً عن التضمينات المشابهة ويدمجها كمستندات مرجعية داخل نموذج اللغة المحلي، مما يمنح المستخدم إجابة دقيقة وشخصية للغاية مع الحفاظ التام على سرية البيانات.

المبحث السادس: أمن النظام والخصوصية الرياضية الموزعة

تعد الهواتف الذكية الحديثة العقد الطرفية الأساسية في منظومة **التعلم الاتحادي Federated) (Learning**، وهي معمارية تتيح تدريب وتحسين النماذج بشكل جماعي وتوزيعي دون الحاجة لجمع بيانات المستخدمين في خادم مركزي.

٦.١آلية عمل التعلم الاتحادي الموزع

١. يقوم الخادم المركزي للشركة بإرسال نموذج ذكاء اصطناعي ذي أوزان قياسية ابتدائية إلى ملايين الهواتف حول العالم أثناء فترة شحن الهاتف وخموله ليلاً.
٢. يقوم كل هاتف بتطوير وتعديل هذا النموذج محلياً وبشكل مستقل، باستخدام بيانات الاستخدام الخاصة بصاحب الهاتف (مثل طريقة تصحيح الكلمات على لوحة المفاتيح أو التطبيقات المفضلة).
٣. بدلاً من إرسال البيانات الشخصية الخام إلى خوادم الشركة، يقوم الهاتف بحساب "التحديثات والتعديلات" التي طرأت على أوزان النموذج فقط، ويقوم بإرسال هذه التعديلات المشفرة منفردة.
٤. يقوم الخادم المركزي بجمع التحديثات القادمة من ملايين الأجهزة، ويقوم بعملية دمج وحساب متوسط هذه التحديثات لإنتاج نموذج عالمي جديد أكثر ذكاءً وقدرة على التنبؤ، ثم يُعاد إرساله للهواتف مجدداً.

٦.٢الخصوصية التفاضلية(Differential Privacy)

لحماية أوزان النماذج المرسلة من خوادم الهواتف من هجمات الهندسة العكسية (حيث يمكن للمخترقين أحياناً استنتاج البيانات الأصلية من خلال تحليل التغير في الأوزان)، يتم تطبيق **الخصوصية التفاضلية.**

- تقوم الخوارزمية بإضافة "ضوضاء عشوائية منضبطة ومدروسة" إلى التحديثات قبل إرسالها من الهاتف.