بنية لصهر البيانات متعددة الوسائط

: رسالة مقدمة لـ

قسم علوم الحاسب

كلية الحاسبات والذكاء اإلصطناعي – جامعة الفيوم

كجزء من متطلبات الحصول على درجة الدكتوراه في فلسفة علوم الحاسب

:مقدمه من

هدير مصطفي سيد طلبة

مدرس مساعد بقسم علوم الحاسب

كلية الحاسبات والذكاء اإلصطناعي – جامعة الفيوم

:تحت اشراف

أ.د. هشام الديـــب

أستاذ بمعهد بحوث االلكترونيات

رئيس جامعة القاهرة الجديدة التكنولوجية

أ.د شيرين علي طايع

استاذ بقسم علوم الحاسب

وكيل كلية الحاسبات والذكاء االصطناعي

لشؤون الدراسات العاليا والبحوث

جامعة الفيوم

كلية الحاسبات والذكاء اإلصطناعي - جامعة الفيوم

جمهورية مصر العربية

2023

ملخص الرسالة

في عصرمعالجة وتحليل البيانات الكبيرة، نعمل على مجموعات متنوعة من البيانات من مصادر ومجالات مختلفة والتي تصف حدثًا واحدًا محددًا. تتألف مجموعات البيانات هذه من العديد من الأنماط ذات التمثيلات والتوزيعات والمقاييس والكثافات المختلفة. ومن المعلوم أن الذكاء الإصطناعي يهدف إلى محاكاة فهم الإنسان للظواهر. وبما أن العقل البشري يستطيع تحليل المعلومات بأشكال متنوعة ويستطيع استخلاص تفسيرات واستنتاجات من نواتج التحليل، فقد دعت الحاجة إلى أن يتمتع الذكاء الإصطناعي بالقدرة على تفسير وتحليل عدة أنماط من الإشارات معًا لاتخاذ القرار. وتعرف عملية دمج مجموعات البيانات غير المتجانسة وذات أنماط مختلفة لإنتاج معلومات أكثر عملية وشمولية باسم "صهر البيانات متعددة الوسائط".

ويعد التعرف السمعي البصري على الكلام بمثابة تقنية تسعى الى التعرف على الكلمات والعبارات المنطوقة من خلال تحليل المؤثرات الصوتية والبصرية معًا. وتعتمد أنظمة التعرف السمعي البصري على إخراج الصوت الصادر من صوت المتحدث بالإضافة إلى المؤثرات مثل حركات الشفاه وتعابير الوجه ولغة الجسد لتعزيز دقة التعرف على الكلام. فمن خلال دمج الصوت والصورة يمكن لأنظمة التعرف السمعي البصري أن تحقق مستويات دقة أعلى بكثير من الأنظمة التقليدية للتعرف على الكلام والتي تعتمد فقط على إدخال الصوت وتحليله.

في بداية هذه الرسالة، تم عرض تقييم كامل لأحدث تقنيات دمج البيانات متعددة الوسائط في السنوات الأخيرة. وبالإضافة إلى ذلك، تم عرض فرق الكفاءة بين الاستراتيجيات المختلفة لدمج البيانات متعددة الوسائط. ومن ثمّ، استغلت هذه الرسالة مشكلة التعرف السمعي البصري كدراسة حالة للتحقق من قوة التعلم العميق في دمج البيانات متعددة الوسائط بناءً على التغييرات في الدقة وعدد التحديات المتناولة في كل نموذج.

في هذه الأطروحة، يتم اقتراح ثلاث نماذج لصهر البيانات متعددة الوسائط وهم BiVAE و BiVAE_SoA و BiVAE_SeA. وتهدف هذه النماذج إلى دمج البيانات السمعية والبصرية للحصول على دمج فعال في سياق المهام السمعية – البصرية. يستغل نموذج BiVAE قدرات خوارزمية VAE في تعلم التمثيلات الكامنة بسلاسة وإنتاج بياناتجديدة. وهذا يمكّنه من إنشاء تمثيل موحَّد يجمع بين سمات الصوت والبصر. وتعد نماذج BiVAE_SoA و BiVAE_SeA نسخ محسنة من النموذج BiVAE والتي تدمج آاليات الانتباه التي تُستخدم لتطوير نماذج قادرة على إعطاء أولوية للإشارات التي تحتوي على محتوى أكثر إفادة أثناء مهمة التعرف، بدلاً من الاعتماد على جميع الإشارات المتكاملة بنفس القدر. حيث تقوم هذه النماذج بإعطاء وزنًا أعلى للإشارة ذات الصلة مقارنةً بالإشارة الأخرى وفقًا لنسبة الإشارة إلى الضوضاء.

وتهدف تجارب الأداء إلى تقييم دقة التعرف على الكلام اعتمادًا على الاشارات السمعية البصرية المدمجة باستخدام النماذج المقترحة مقارنة بدقة التعرف على الكلام اعتمادًا على الاشارات الصوتية فقط . واشتملت التجربة على استخدام تقنيتين لاستخراج السمات الصوتية وهما MFCC و GFCC. كما تم استخدام ثلاث مصنفات مختلفة وهم LSTM و ANN و SVM لتقييم مدى جودة وفاعلية التمثيل الموحد لسمات الاشارات السمعية البصرية المتكاملة، هذا التمثيل الناتج عن النماذج المقترحة. وتضمنت التجربة فرضيتين وهما: توافر جميع الاشارات المتكاملة أثناء التدريب والأختبل الموجه وكذلك غياب احدى الأشارات أثناء التدريب والأختبار الموجه أيضًا.

وقد أوضحت النتائج أن النموذج المقترح BiVAE_SeA هو الأفضل مقارنة بالنماذج المقترحة الأخرى. ومن خلال رؤية أكثر شمولية للنتائج تبين أن مهام التعرف على الكلام اعتمادا صهر الاشارات السمعية البصرية باستخدام النماذج المقترحة BiVAE و BiVAE_SoA و BiVAE_SeA تفوقت على مهام التعرف على الكلام اعتمادا على الأشارات الصوتية فقط بفارق متوسط دقة يصل إلى ٤٠.١٨% و ٤٠.٧٧% و ٤١.٠٧% على التوالي. إضافة إلى ذلك، فقد أظهرت النتائج تشابه الأداء في مهام التعرف السمعي البصري فيما بين السمات الصوتية المستخرجة

باستخدام خوارزمية MFCC و والمستخرجة باستخدام خوارزمية GFCC رغم التباين الملحوظ في أداء مهام التعرف السمعي اعتمادًا على كليهما، مما يشير إلى قدرة النماذج المقترحة على تعميم أدائها في مدى جودة وفعالية التمثيل الموحد لسمات الاشارات المدمجة التي تنتجها بصرف النظر عن مدى كفاءة التقنية المستخدمة في استخراج سمات الأشارات. كما أظهرت النتائج تفوق مصنف SVM على المصنفات الأخرى في مهام التعرف على الكلام السمعي فقط والسمعي البصري.

كما تمكنت نماذج الدمج المقترحة من التفوق على النماذج السابقة بفارق متوسط دقة يصل إلى ٤.٨٤% و ١٦.٢٦% في حالتي الصوت النقي والمشوش على التوالي. وإضافة إلى ذلك، فقد تفوقت النماذج المقترحة على النماذج الحديثة في حالة غياب أحدى الأشارات المتكاملة بفارق متوسط دقة يصل إلى ٣.٤٦% و ١٧.٥٤% في حالة توافر السمات الصوتية فقط وحالة توافر السمات المرئية فقط على التوالي.

تتكون الرسالة من ستة فصول منظمة كالآتي:

الفصل الأول

يحتوي هذا الفصل على مقدمة عن موضوع الرسالة والدوافع لهذه الرسالة والهدف منها كما يعرض الهيكل التنظيمي للرسالة.

الفصل الثانى

يقدم هذا الفصل نظرة عامة على التقنيات المستخدمة لدمج البيانات من وسائط مختلفة. كما يناقش الصعوبات التي تواجه هذه العملية مع تقديم تقييم شامل لأحدث الأبحاث حول دمج البيانات المتعددة. بالإضافة إلى ذلك، يتضمن مقارنة بين الأساليب المختلفة ونقاط القوة والضعف الخاصة بها. يتابع هذا الفصل دراسة التعرف على الكلام المرئي السمعي ويستعرض أحدث نماذج دمج البيانات المرئية السمعية.

الفصل الثالث

في هذا الفصل، يتم تقديم المفاهيم الأساسية للنهج والأساليب المعتمدة في الأطروحة، ليتم مناقشتها بإيجاز في الفصل التالي. يتم تنفيذ هذه النهج والتقنيات للمراحل الرئيسية في النظام المقترح، والتي تتضمن مرحلة معالجة المسبقة البيانات، ومرحلة دمج الميزات، ومرحلة تقييم الميزات المدمجة، من خلال مهام التعرف على الكلام السمعي البصري.

الفصل الرابع

يستعرض هذا الفصل تفاصيل المراحل الأساسية للنظام المقترح، والتي تتضمن مرحلة المعالجة المسبقة للبيانات الصوتية والمرئية، ومرحلة صهر/دمج الميزات، ومرحلة تقييم الميزات المدمجة. حيث تم في هذا الفصل اقتراح ثلاث نماذج صهر وهم BiVAE و BiVAE_SoA و BiVAE_SeA لصهر البيانات السمعية البصرية.