التعرف على الكلمات العربية باستخدام الشبكات العصبونية التلافيفية التكرارية

طالبة الدكتوراه هزار النجم

كلية الهندسة المعلوماتية – جامعة البعث

إشراف الدكتورة يسر السيد سليمان الأتاسي

الملخص

إن التعرف على الكلمات العربية في الصور من المسائل المهمة في مجال رؤية الحاسوب لمنفعتها في تطبيقات الأتمتة وتحليل الصور وهي من المسائل الصعبة وذلك بسبب اختلاف الخطوط العربية وطول الكلمة وتنوع أشكال الحرف وفقاً لموقعه في الكلمة، سنقدم في هذه المقالة بنية شبكة عصبونية لحل هذه المسألة وسنفصل كل جزء من هذه البنية ودوره في عملية التعرف وسنعرض بعض الأمثلة عن تطبيق هذه البنية على عدد من الصور والتي تحوي على كلمات عربية

الكلمات مفتاحيه: رؤية الحاسوب – التعرف على الكلمة العربية – التعلم العميق – الشبكات العصبونية التلافيفية – الشبكات التكرارية – التصنيف الزمني الموصل

Recognition of the Arabic Words using recurrent convolutional neural networks

Abstract

Recognition of the Arabic words in the image is one of the important issues in computer vision for its usefulness in automation applications and image analysis, and it is one of the difficult issues due to the difference in Arabic fonts, word length and the variety of letter shapes according to their position in the word, in this article we will present a neural network architecture to solve this issue and we will separate each part of this architecture and its role in the recognition process, and we will present some examples of applying this architecture to a number of images that contain Arabic words

Keywords:

Computer Vision–Arabic Word Recognition–Deep Learning–Convolutional Neural Network – Recurrent Neural Network – Temporal Classification Connectionist

مقدمة

أصبحت نماذج الشبكات العصبونية العميقة محط اهتمام مؤخراً في مسائل تصنيف الأغراض والتعرف على النصوص ومنها الشبكات العصبونية التلافيفية التكرارية (Recurrent Convolutional Neural Network)

والتي لها ميزات الشبكات العصبونية والشبكات التكرارية وتعامل الكلمة على اعتبارها سلسلة من البيانات ومهمتها عنونة كل عنصر من هذه السلسلة على حدى لذا تعد مسألة التعرف على الكلمة مسألة عنونة سلسلة (sequence labelling problem).

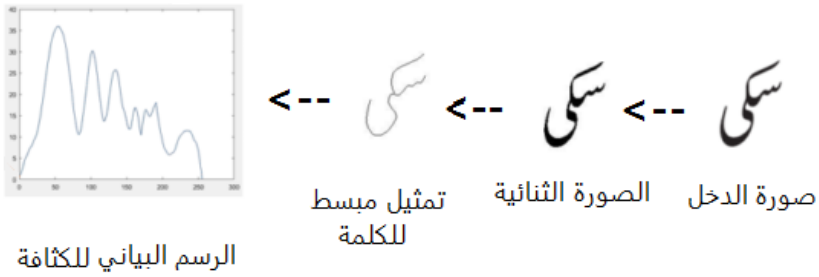
1- الهدف من البحث

إن الهدف من البحث هو حل مشكلة تقسيم الكلمة إلى أحرف حيث كان يتطلب التعرف على الكلمة في صورة تقسيم الصورة إلى أبعاد متساوية (صور أصغر) ومحاولة التعرف على كل حرف في هذه الصورة والذي يعد صعباً في اللغة العربية بسبب اختلاف أبعاد الحرف وأشكاله وموقعه.

3 – الدراسات السابقة للتعرف على الكلمة

اقترحت إحدى الدراسات التعرف على كلمات اللغة الأردية [3] تطبيق معالجة مسبقة للصورة كتحويلها إلى صورة ثنائية ومن ثم تحويل الكلمة إلى تمثيل مبسط عنها وهو ما يسمى (Topological skeleton) وبعدها يتم استخراج الرسم البياني للكثافة (thickness graph) للكلمة وهو عدد البكسلات في كل عمود في

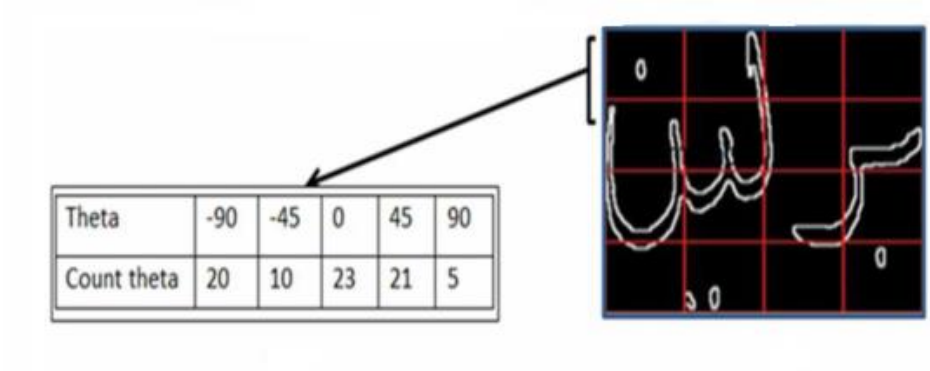
الصورة، ومن ثم ندخل الرسم البياني للشبكة العصبونية عوضاً عن أن يكون الدخل هو الصورة بوضعها الطبيعي (raw image) كما يوضح الشكل 3-1:



الشكل 3-1 معالجة مسبقة لصورة لاستخراج الرسم البياني للكثافة

من مساوئ هذه الطريقة أنها تتطلب استخدام كل الكلمات في اللغة لتحليلها وتصنيفها وبذلك تكون طريقة غير عملية بسبب عدد الكلمات في اللغة العربية ويقدر بـ12.3 مليون كلمة.

هناك دراسة أخرى [1] اقترحت استخراج شعاع ميزات لكل كلمة فيها يتم تقسيم الصورة إلى 4*4 أجزاء، وتستخرج ميزات اتجاه الميل (gradient direction feature) لكل جزء على حدى على اعتبار أن الزوايا مقسمة لـ5 زوايا رئيسية وهي [-90, -45,0,45,90] كما يوضح الشكل 3-2:

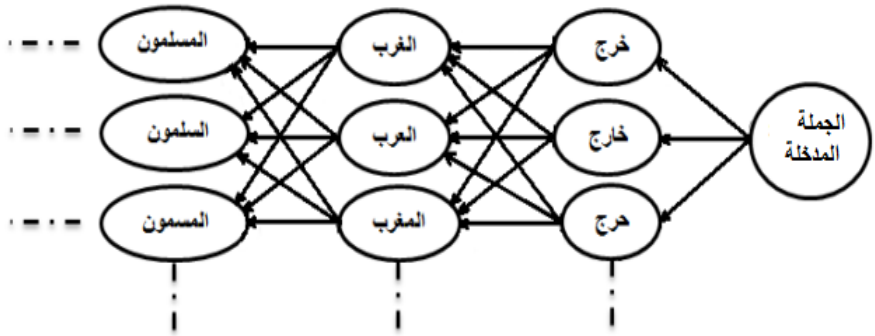


الشكل 3-2 استخراج ميزات اتجاه الميل لصورة

ويكون طول الشعاع الناتج (80)، كذلك يستخرج شعاع ميزات آخر بطول (128) ناتج عن تطبيق تحويل المويجة المنفصلة (Discrete Wavelet Transform) على الصورة عند المستوى (LL3) واستخراج الميزات البنيوية وقيم الانحراف المعياري للصورة الناتجة وتمثيلها على شكل شعاع بطول (128) يضاف هذا الشعاع إلى الشعاع السابق ليصبح الطول الكلي للشعاع هو (208) يدخل إلى مرحلة التصنيف باستخدام خوارزمية الجار الأقرب (k-nearest neighbors)، من مساوئ هذه الطريقة المعالجة المسبقة والتي تتطلب وقت وجهد.

أما في الدراسة [4] فتم استخدام عدد كبير من الكلمات العربية للتدريب واستخرجت ميزات تحويلات الجيب المنفصلة (Discrete Cosine Transform) لهذه الصور لتجميع الكلمات المتشابهة ضمن حزم (clusters) وفقاً لميزاتها، في مرحلة الاختبار يدخل السطر الذي يحوي نصاً وليكن (L) وفيه (n) كلمة على نظام التعرف ومن ثم يقسم لصور كلمات ($x_1,x_2,,,,,x_n$) تقارن ميزات كل صورة دخل (x_i) مع ميزات الحزم

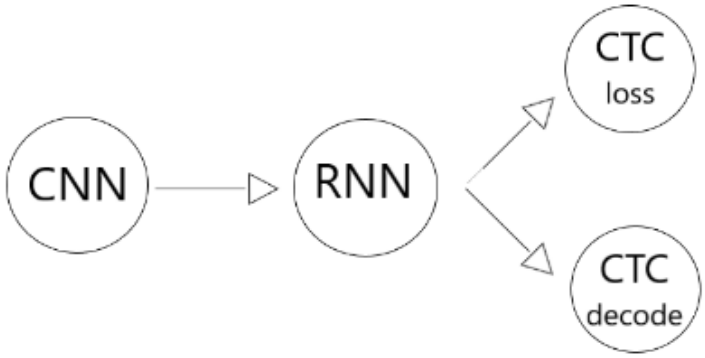
وتسند الصورة للحزمة التي تحقق تشابهاً معها أي تحقق أقل مسافة اقليدية (Euclidean Distance) ممكنة معها، نستخدم كلمات الحزم التي تقابل كل صورة(x_i) في السطر (L) لتشكل سلسلة من الخيارات المتاحة وباستخدام نموذج لغة (n-gram language) يمكن اختيار النص الصحيح كما في الشكل3-3 الذي يعرض مجموع من كلمات كل حزمة مثل (خرج – خارج –حرج...)



الشكل 3-3 مجموعة من الخيارات الممكنة والتي تقابل كلمات السطر (L)

4 – البنية المقترحة للشبكة العصبونية التلافيفية التكرارية:

تتألف هذه البنية من ثلاثة مكونات [5] وهي الطبقات التلافيفية (convolutional Layers) والطبقات التكرارية (Recurrent Layers) وطبقة التصنيف الزمني الموصل (Temporal Classification Connectionist Layer) كما يوضح الشكل 4-1

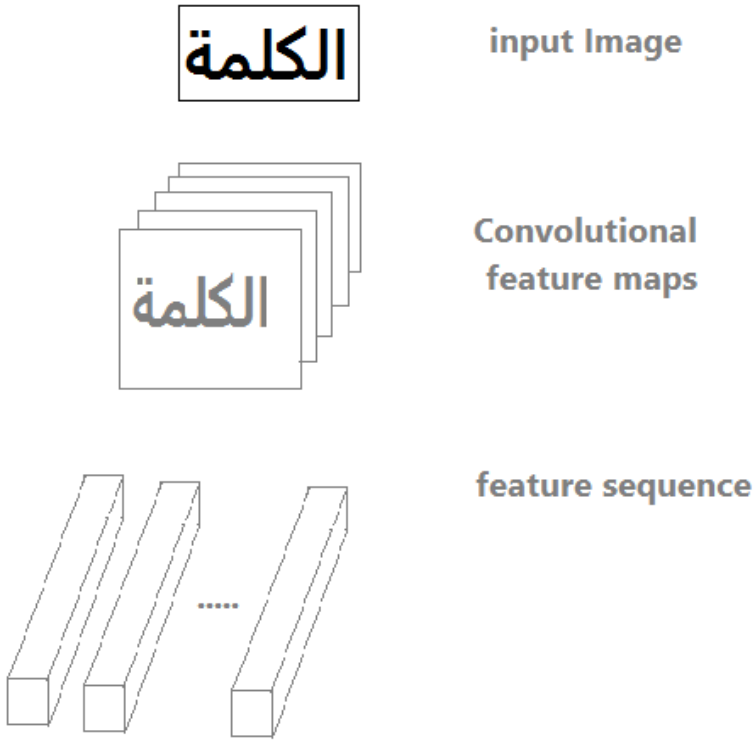


الشكل 4-1 شكل مبسط للبنية المقترحة

دخل هذه الشبكة هو صورة لكلمة ما أما خرجها فيكون سلسلة من العناوين (label sequence) أي سلسلة من الأحرف (العناوين) التي تشكل كلمة، سنفصل فيما يلي كل مكون فيها ودوره في عملية التعرف.

4-1-1 الطبقات التلافيفية (convolutional Layers)

تحتوي على طبقات تلافيفية وطبقات تجميع (max-pooling) مشابهة لنموذج الشبكة العصبونية التلافيفية القياسي ويستخدم لاستخراج ميزات صورة الدخل على شكل خريطة ميزات (Feature Map) ثم تحول هذه الخريطة لسلسلة من أشعة الميزات (sequence feature) وذلك باستخدام العمود i من كل عنصر من عناصر الخريطة ليشكل شعاع الميزة i أي أنه كل عنصر من سلسلة الميزات يقابل منطقة مستطيلة من الصورة الأصلية ويعد وصف لهذه المنطقة يوضح الشكل 4-2 آلية عمل الطبقات التلافيفية



الشكل 4-2 آلية عمل الطبقات التلافيفية

4-1-2 الطبقات التكرارية (Recurrent Layers):

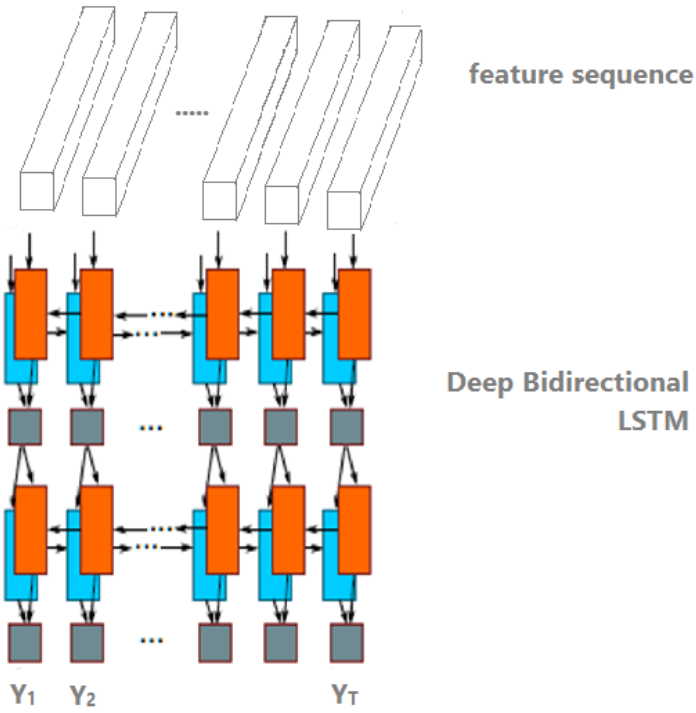
تحتوي على شبكة عصبونية تكرارية ثنائية الاتجاه (Bidirectional Recurrent Neural Network)

دخلها هو خرج المرحلة السابقة (سلسلة الميزات) وخرجها هو مصفوفة تسمى (character score) تعبر عن احتمالية وجود كل عنوان (label) في كل شعاع ميزة من سلسلة الميزات.

تتألف الشبكات العصبونية التكرارية من وحدات عصبونية تحوي الوحدة التقليدية (i) على طبقة مخفية وطبقة دخل وخرج، عندما تتلقى الوحدة قيمة x_i تتغير قيمة h_t (والتي تعبر عن حالة الوحدة الحالية) بتابع غير خطي باستخدام الدخل والحالة السابقة للوحدة h_{t-1} وفق المعادلة التالية [6]:

$$h_t = G(x_i , h_{t-1})$$

ويكون خرج الوحدة Y_i بالاعتماد على h_t تعاني الشبكات التكرارية من مشكلة التدرج المتلاشي (vanishing gradient problem) مما يؤثر سلباً على عملية التدريب لذا تم حل هذه المشكلة باستخدام شبكات الذاكرة طويلة المدى وهي نوع من الشبكات التكرارية والتي تجيد التعامل مع البيانات المتسلسلة لذا نستخدم شبكات الذاكرة طويلة المدى عوضاً عن الشبكات التكرارية التقليدية، يوضح الشكل 4-3 الطبقات التكرارية ثنائية الاتجاه المستخدمة وتحوي على شبكة ذاكرة طويلة المدى من اليسار إلى اليمين (forward) وشبكة أخرى من اليمين إلى اليسار (backward) والسبب في استخدام شبكة ثنائية الاتجاه لأن ذلك يمكنها من التعرف على شكل الحرف من اليمين إلى اليسار ومن اليسار إلى اليمين.



الشكل 4-3 شبكة عصبونية تكرارية ثنائية الاتجاه

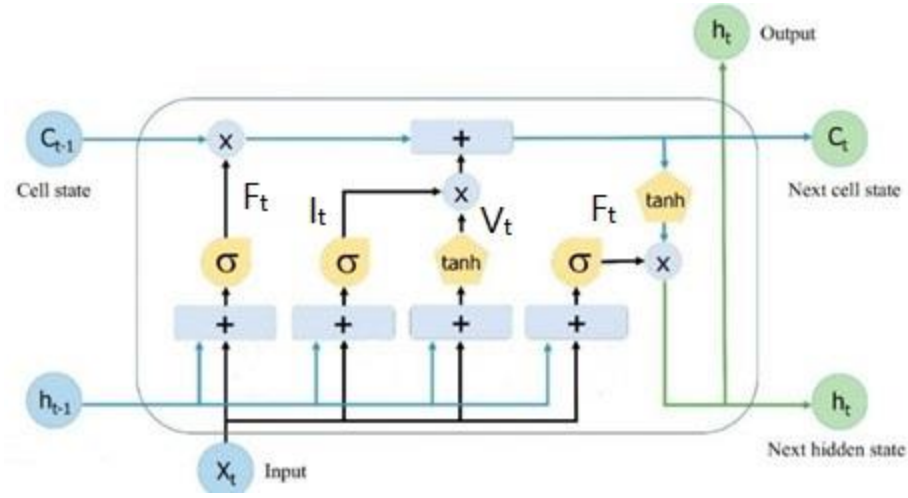
4-1-2-1 شبكات الذاكرة طويلة المدى (Long Short Term Memory networks) [3]

هي نوع خاص من الشبكات تحتوي على سلسلة من وحدات الشبكة العصبية المتكررة، في كل منها تم استخدام بوابات (gates) تقسم البوابات إلى ثلاثة أنواع:

- بوابة الدخل (input gate)
- بوابة النسيان (forget gate)
- بوابة الخرج (output gate)

بالإضافة لما يسمى حالة الخلية (cell state) وهو عبارة عن ناقل للبيانات عبر السلسلة

يوضح الشكل 4-4 نموذجاً من هذه الوحدات



الشكل 4-4 شبكات الذاكرة طويلة المدى

أول خطوة في شبكات الذاكرة طويلة الأمد هو اختيار أي من المعلومات ستهمل في حالة الخلية C_t، يعود هذا الخيار لبوابة النسيان حيث فيها يتم تمرير الدخل X_t وخرج الخلية السابقة h_{t-1} لتابع sigmoid ويكون الخرج F_tهو شعاع قيمها بين 0 و 1 بالنسبة لـ C_{t-1} القيم الأقرب (0) في F_t تهمل والقيم الأقرب للـ(1) تحفظ .

الخطوة التالية هي اختيار أي من المعلومات سيتم حفظها باستخدام بوابة الدخل فيها يتم تمرير الدخل x_t والحالة المخفية h_{t-1} لتابع sigmoid لتقرير أي من المعلومات سيتم تحديثها لتولد I_t، ثم تدخل نفس البيانات

(x_t,h_{t-1}) لتابع tanh لتشكل شعاع من القيم المرشحة الجديدة V_t والتي تكون بين -1 و 1 والتي يمكن أن تضاف لحالة الخلية، تستخدم القيم $(F_t،V_t,I_t)$ لتحدث معلومات حالة الخلية وفقاً للمعادلة التالية [7]

$$C_t=F_t \times C_{t-1} + I_t \times V_t$$

الخطوة الأخيرة هي اختيار الخرج (ويسمى أحياناً Hidden State يحوي معلومات عن الإدخالات السابقة X و يستخدم للتوقع) بالاعتماد على حالة الخلية حيث نقرر أي جزء من حالة الخلية سيتم تمريره كخرج عن طريق تابع tanh وفقاً للمعادلة التالية [7]

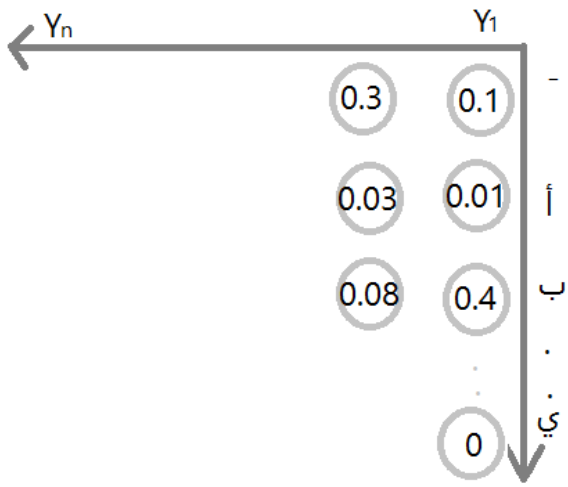
$$H_i=\tanh(C_i) \times F_i$$

بالتعرف على بنية شبكات الذاكرة الطويلة الأمد نلاحظ كيف أنها تحل مشكلة التعلم البطيء والذاكرة القصيرة الأمد للشبكات التكرارية بتخزين معلومات الادخال السابقة وتعديل حالة الخلية بشكل مستمر، في الشبكات العصبونية التلافيفية التكرارية يكون دخل كل وحدة هو شعاع ميزة يحوي معلومات عن جزء من الصورة والخرج هو مصفوفة أحادية البعد تحوي توزع احتمالات وجود حرف ما ضمن شعاع الميزات المدخل لها وتشكل مجموعة هذه المصفوفات (character –score).

4-1-3 طبقة التصنيف الزمني الموصل (Temporal Classification Connectionist Layer):

دخل هذه الطبقة هو خرج المصفوفة السابقة character–score وخرجها هو الكلمة المتوقعة في الصورة ولها مهمتان أولاً حساب قيمة الخسارة لتدريب الشبكة العصبونية ثانياً فك ترميز (decoding) المصفوفة وذلك باختيار الحرف الذي له احتمالية أكبر من كل عمود ضمن المصفوفة لتشكل مجموعة هذه الأحرف الكلمة المناسبة.

بدراسة اللغة العربية كمثال فإن عدد العناوين (الأحرف) هو (28) ويضاف له ما يسمى بالعنوان الفارغ (blank) ويرمز بـ(–) ويمثل أي خرج ليس ضمن الأحرف العربية وبذلك يكون عدد العناوين المحتملة مقابل كل عمود من character–score هو (29)، في الشكل 4–5 مثال عن character–score فيه كل شعاع Y_i يعبر عن احتمالات وجود الأحرف العربية والعنوان الفارغ ضمن عمود في الصورة



الشكل 4–5 مثال عن character–score

يتضمن فك التشفير البحث عن عدد من الخيارات الممكنة بالاعتماد على احتمالية وجودها في الصورة وبما أن حجم المفردات غالباً ما يكون مئات ومئات من الكلمات لذا تعد مسألة البحث من المسائل الصعبة الحل، هناك طرق عديدة لفك التشفير منها طريقة البحث الشره (Greedy Search Decoder) وطريقة البحث عن سلاسل الكلمة (Word Beam Search)

4-1-3-1 طريقة البحث الشره Greedy Search

تقوم هذه الطريقة على مبدأ اختيار الحرف الذي له الاحتمالية الأكبر في كل خطوة كمثال من الشكل 4-5 من الشعاع Y_1 نختار المحرف (ب) الذي يقابل القيمة (0.4) وهي أكبر قيمة ضمن الشعاع ثم نختار (-) أي محرف فارغ لأن له الاحتمالية الأكبر وهكذا حتى تنتهي من المصفوفة ونحصل على نص مقابل، من منافع هذه الطريقة أنها سريعة جداً وفعالة لكن الخرج قد لا يكون مثالياً.

4-1-3-2 بحث عن سلاسل الكلمة Word Beam Search [2]

مبدأ هذه الطريقة هي اختيار أعلى قيمة بين جداءات الاحتمالات الممكنة من العلاقة التالية:

$$p(s) = p(0) * p(1) * \dots * p(T)$$

S : السلسلة

T: طول السلسلة

P(i): احتمال محرف ما في السلسلة i

تقدم عدد من الاحتمالات الممكنة كلما زاد هذا العدد كلما تحسن خرج النموذج .

باستخدام إحدى طرق فك التشفير السابقة يكون الخرج لإحدى الصور على الشكل:

ا---لل-ككك--لل-ممم----ة---

بحيث يكون كل محرف مقابل لعمود (1 بكسل) من الصورة، بحذف الأحرف المتكررة ضمن كل مجموعة يصبح الخرج:

ا---ل-ك--ل-م----ة---

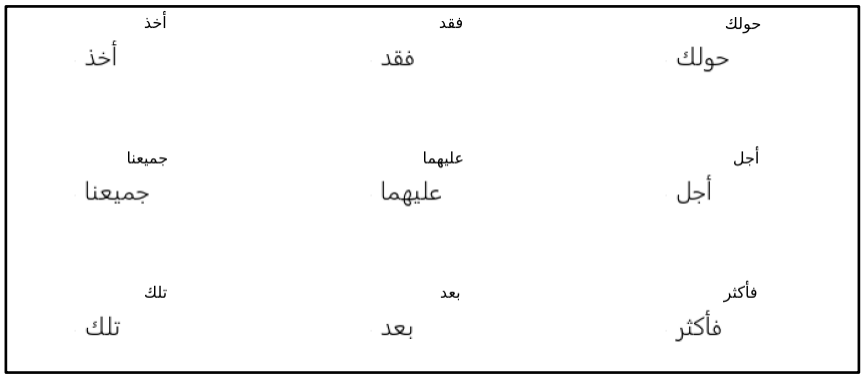
ثم حذف المحرف الفارغ (-) ويصبح الخرج (ا ل ك ل م ة) --> الكلمة.

من ميزات الشبكات العصبونية التلافيفية التكرارية أولاً أن لها القدرة على استخراج المعلومات من الأغراض (النصوص) التي لها بنية محددة في السلسلة، ويساعد هذا في التعامل مع كل محرف وفقاً لطوله مثلاً حرفي الـ "ك" و "ب" يمكن التمييز بينها وفقاً للطول، ثانياً يمكن لهذه البنية استخدام خاصية التغذية الرجعية (back-propagation) لتعديل الأوزان و تقليل الأخطاء مما يسمح له بتدريب الطبقات التكرارية والتلافيفية، ثالثاً تستطيع هذه الشبكة العمل على سلاسل (نصوص) مختلفة الطول.

5- النتائج العملية:

بتطبيق البنية السابقة باستخدام لغة البايثون على قاعدة بيانات أعددناها سابقاً يدوياً وتم تشكيلها من توليد صور عدد من كلمات اللغة العربية وتحوي على 1700 صورة عربية

وباستخدام بيئة Google Colab تم تدريب الشبكة واختبارها وكانت النتائج كما في الشكل 5-1 كل صورة من الصور تحوي محتواها النصي أعلاها.



الشكل 5-1 النتائج العملية

WRR Word Recognition Rate معدل التعرف على الكلمة

يوضح الجدول 5-1 مقارنة بين نتائج دقة التعرف على الكلمة (Word Recognition Rate)

حيث يعرف WRR وفق المعادلة:

$$\frac{\text{عدد الكلمات تم التعرف عليها بشكل صحيح}}{\text{عدد الكلمات الكلي}}$$

في الدراسات السابقة المذكورة في الفصل الثالث ونتائج دراساتنا

دقة الدراسة السا	8
دقة الدراسة السا	5

دقة الدراسة السا	6
دقة دراستنا	9

الجدول 5-1

كانت نتائج الدقة في الدراسة الأولى ممتازة لأنها تدرب الشبكة العصبونية على تمثيل بياني مبسط للكلمة أما الدراسة الثانية فكانت نتائجها أقل من سابقتها لأنها تستخدم مصنف الجار الأقرب عوضاً عن الشبكة العصبونية، نتائج الدراسة الثالثة هي الأقل بسبب عدد التجمعات (cluster) الكبير حيث يحوي كل تجمع الكلمات العربية المتشابهة في اللغة العربية أما في دراستنا فقد كانت الدقة ممتازة بسبب الشبكة الهجينة المستخدمة التي تولد ميزات الصورة وتعالجها بالشبكات العصبونية التكرارية ثنائية الاتجاه لتولد احتمالات وجود كل حرف في موقع ما ضمن الصورة.

تطلب تدريب الشبكة السابقة (56 دقيقة و 43 ثانية) وزمن الاختبار (28.2 ثانية).

6 – الأعمال المستقبلية:

يمكن توسيع هذه الدراسة لتشمل كشف والتعرف على الكلمات العربية في الصورة باستخدام شبكات أعقد من بنية الشبكات العصبونية التلافيفية التكرارية مثل الشبكة العصبية التلافيفية التي تعتمد على القناع Mask R-CNN وتقوم بمهام التعرف وكشف الأغراض في الصورة باستخدام مجموعة من النوافذ الممكنة.

7 – الخاتمة:

إن التعرف على الكلمات العربية باستخدام الشبكات العصبونية التلافيفية التكرارية هو مجال بحث نشط يحتاج دائماً إلى تحسين في الدقة، في القسم العملي أظهرنا أن دقة التصنيف كانت واعدة بمعدل 98.9٪ على صور الاختبار، نأمل في المستقبل أن يتحسن معدل الدقة ليشمل الكلمات العربية مهما كان شكلها ونوع الخط في الصورة.

8- المراجع:

[1] عباس م.ع ، عبدالكريم ع. ,2018 , Proposed Multi Feature Extraction Method For Off-line Arabic Handwriting Word Recognition, مجلة المنصور, Vol. 2018,No. 30, P.P. 17

[2]H. Scheidl, S. Fiel and R. Sablatnig, "Word Beam Search: A Connectionist Temporal Classification Decoding Algorithm," 2018 16th International Conference on Frontiers in Handwriting Recognition (ICFHR), 2018, pp. 253-258, doi: 10.1109/ICFHR-2018.2018.00052.

[3] Naseer A. , Zafar K., 2018, Comparative analysis of raw images and meta feature based Urdu OCR using CNN and LSTM,International Journal of Advanced Computer Science and Applications, Vol. 9,No. 1, P.P. 419-424

[4] Nashwan F.M.A. , Rashwan M.A.A. , Al-Barhamtoshy H.M. , Abdou S.M. , 2018, A holistic technique for an Arabic OCR

system,Journal of Imaging, Vol. 4,No. 1, P.P. 1–11

[5] Safarzadeh V.M. , Jafarzadeh P., 2020, Offline Persian Handwriting Recognition with CNN and RNN–CTC,2020 25th International Computer Conference, Computer Society of Iran, CSICC 2020, No. February, P.P. 1–13

[6] Shi B. , Bai X., 2017, An End–to–End Trainable Neural Network for Image–Based Sequence Recognition and Its Application to Scene Text Recognition,IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 39,No. 11, P.P. 2298–2304

[7] Vinet L. , Zhedanov A., 2011, A “missing” family of classical orthogonal polynomials,Journal of Physics A: Mathematical and Theoretical, Vol. 44,No. 8, P.P. 156–157