Modèle CUTIE

En faisant correspondre la sortie de CUTIE avec les données étiquetées de la grille, le modèle apprend à générer l'étiquette pour chaque texte de la grille d'entrée en explorant les caractéristiques spatiales et sémantiques. Pour cette raison, la tâche de CUTIE ressemble à la tâche de segmentation sémantique dans le domaine de la vision par ordinateur, mais avec des distributions de données plus éparses. Plus précisément, la grille cartographiée contient des points de données éparses (fragments de texte), contrairement aux images remplies de pixels. Les textes clés cartographiés sur la grille sont soit proches soit éloignés les uns des autres en raison des différents types de mise en page des documents. Par conséquent, l'intégration d'une capacité de traitement du contexte à plusieurs échelles est bénéfique pour le réseau.

En fait, plusieurs méthodes ont été proposées dans le domaine de la segmentation sémantique pour capturer les contextes multi-échelles dans les données d'entrée. Les méthodes de la pyramide d'images et de la structure codeur-décodeur visent toutes deux à exploiter les informations multi-échelles. Les objets intéressés de différentes échelles deviennent proéminents dans les premiers réseaux en utilisant des données d'entrée à échelles multiples pour rassembler des caractéristiques multi-échelles. Les réseaux ultérieurs réduisent les cartes de caractéristiques pour agrandir les champs réceptifs et réduire la charge de calcul, puis capturent des détails plus fins en récupérant progressivement les informations spatiales des caractéristiques des couches inférieures. Cependant, la résolution spatiale est réduite au cours du processus d'encodage et le processus de décodage n'exploite que les caractéristiques de haute résolution mais de bas niveau pour récupérer la résolution spatiale, le processus d'encodage par enjambement consécutif décime les informations détaillées [11]. De plus, le processus de codage et de décodage applique des restrictions de forme au processus d'augmentation de la forme de la grille, comme présenté dans la section 3.1.

En revanche, le champ de vision des filtres peut être élargi de manière efficace et les contextes multi-échelles peuvent être capturés en combinant des caractéristiques multi-résolution [11] ou en appliquant une convolution atrus [2, 3, 4, 5]. Pour capturer la connexion à longue distance et éviter la perte potentielle d'information dans le processus d'encodage, nous proposons deux architectures de réseau différentes et comparons leurs performances dans la section 4. En fait, nous avons expérimenté plusieurs types de structures de modèles et nous n'en détaillons que deux ici pour éviter d'être un article fastidieux. Plus précisément, le réseau CUTIE-A proposé est un réseau neuronal convolutif à haute capacité qui fusionne les caractéristiques multirésolution sans perdre les caractéristiques haute résolution, le réseau CUTIE-B proposé est un réseau convolutif avec convolution atrous pour élargir le champ de vision et module ASPP (Atrous Spatial Pyramid Pooling) pour capturer les contextes multi-échelle. CUTIE-A et CUITE-B mènent toutes deux un processus d'encodage du sens sémantique avec une couche d'incorporation des mots au tout début. Le Dropout est appliqué à la couche d'incorporation pour améliorer la capacité de généralisation de CUTIE. La fonction de perte d'entropie croisée est appliquée pour comparer la grille de classe de tokens prédite et la grille de vérité terrain.

1 CUTIE-A

CUTIE-A évite la perte d'information dans le processus d'encodage tout en tirant parti des encodeurs en combinant les résultats de l'encodage aux représentations haute résolution maintenues tout au long du processus de convolution. Comme pour le HRNet proposé dans [11], un réseau haute résolution sans stridulation est utilisé comme réseau principal et plusieurs sous-réseaux haute-basse résolution sont progressivement ajoutés et connectés au réseau principal. Pendant le processus de

connexion du réseau principal et des sous-réseaux, les caractéristiques multi-échelles sont fusionnées pour générer des représentations riches.

2 CUTIE-B

CUTIE-B est construit avec un seul réseau principal mais utilise la convolution atrous pour capturer les connexions longue distance. Pour la convolution atrous, la carte de caractéristiques d'entrée est m, le filtre est w et la sortie est n. Pour chaque position i, la convolution atrous est appliquée sur la carte de caractéristiques d'entrée m comme suit

$$n[i] = X\ k\ m[i + r - k]w[k] \quad (5)$$

où r est le taux atrous qui indique le pas d'échantillonnage du signal d'entrée, qui est mis en œuvre en convoluant la caractéristique d'entrée avec des filtres suréchantillonnés en insérant r - 1 zéros entre deux valeurs de filtre consécutives le long de chaque dimension spatiale. La convolution standard est un cas particulier de la convolution atrous avec r = 1 [3].

4. Expériences

La méthode proposée est évaluée dans le cadre du défi de lecture robuste ICDAR 2019 sur le jeu de données SROIE et également sur un jeu de données créé par nous-mêmes avec 3 types d'images de documents numérisés, qui contiennent 8 classes d'informations clés et une classe de type "don't care". Pour chaque classe d'information clé spécifique, plusieurs jetons peuvent être inclus. La performance globale est appelée précision moyenne stricte (AP) et mesurée en termes de précision par classe sur les 9 classes, où une classe est considérée comme correcte uniquement si chaque token de la classe est correct. Pour réaliser une analyse plus approfondie de la performance de la méthode proposée, nous proposons d'utiliser un critère supplémentaire, la précision moyenne souple (softAP), où la prédiction d'une classe d'information clé est déterminée comme correcte si les vérités de base positives sont correctement prédites même si certains faux positifs sont inclus dans la prédiction finale. La précision douce est importante car elle indique la capacité du modèle à extraire des informations clés correctes tout en tolérant l'inclusion de certains faux positifs. En fait, des post-traitements peuvent être utilisés pour éliminer les faux positifs. Par conséquent, l'analyse conjointe de la PA et de la softAP permet de mieux comprendre la performance du modèle. Nous comparons la performance de la méthode proposée avec deux méthodes de pointe : CloudScan [9] et BERT pour NER [7]. Pour comparaison, le modèle Cloud Scan pour SROIE est formé à partir de zéro, mais avec plusieurs caractéristiques conçues par des experts, comme décrit dans [9]. Le modèle BERT pour SROIE est transformé à l'aide du modèle de base publié par Google qui est pré-entraîné sur un énorme ensemble de données contenant 3 300 millions de mots [7, 1]. Pour fournir une comparaison équitable, 4 484 échantillons, dont environ 1 500 provenant de taxis, de repas et de divertissements (ME) et d'hôtels, ont été utilisés pour l'apprentissage et 1 125 pour les tests, comme indiqué dans le tableau 1. Un seul modèle est formé sur le jeu de données pour ces trois types de reçus de documents, soit pour CloudScan, BERT ou CUTIE. Nous utilisons un taux d'apprentissage de 1e-3 avec l'optimiseur Adam et la stratégie d'apprentissage step decay. Le taux d'apprentissage est réduit à 1e - 4 et 1e - 5 aux 15 000e et 30 000e étapes, respectivement. L'apprentissage se termine au bout de 40 000 étapes avec une taille de lot de 32. Nous formons notre modèle sur le GPU Tesla V100 où 11 à 19 Go de mémoire sont utilisés en fonction de la configuration du cadre du modèle et de la taille de

l'ensemble de données. La normalisation des instances et l'extraction des négatifs durs sont utilisées pour faciliter l'entraînement. L'abandon est appliqué avec une probabilité de maintien de 0,9. Notre modèle est formé de bout en bout sans préformation par morceaux d'un quelconque composant. La taille d'incorporation par défaut est de 128, la forme d'augmentation cible est de 64 pour les lignes et les colonnes. L'ensemble de données est divisé en un ensemble d'entraînement et un ensemble de test avec un ratio de 75 : 25. CUTIE ne fait appel à aucun prétraitement ni post-traitement.

Jeu de données

L'ensemble de données SROIE de l'ICDAR 2019 (tâche 3) contient 1000 images de reçus numérisés entiers. Chaque image de reçu contient environ quatre champs de texte clés, tels que le nom des marchandises, le prix unitaire, la date et le coût total. Le texte annoté dans le jeu de données est principalement composé de chiffres et de caractères anglais.

Le jeu de données est divisé en un ensemble de formation/validation (trainval) et un ensemble de test (test). L'ensemble d'entraînement est composé de 627 images de reçus qui sont mises à la disposition des participants avec leurs annotations. Comme l'annotation de test n'est pas encore disponible, nous effectuons un lavage sur l'ensemble de formation et sélectionnons 517 échantillons, où les 55 échantillons filtrés étaient mal étiquetés. Ensuite, nous avons divisé l'ensemble de formation dans le rapport 75 : 25, où 75% sont utilisés comme données de formation et 25% sont utilisés pour la validation. Le jeu de données auto-construit contient 4 484 documents de reçus espagnols numérisés et annotés, y compris des reçus de taxi, des reçus de repas et de divertissement (ME) et des reçus d'hôtel, avec 9 classes d'informations clés différentes. Nous générons les textes et les cadres correspondants à l'aide de l'API OCR de Google. Chaque texte et son cadre est étiqueté manuellement comme l'une des 9 classes différentes : 'DontCare', 'VendorName', 'VendorTaxID', 'InvoiceDate', 'InvoiceNumber', 'ExpenseAmount', 'BaseAmount', 'TaxAmount', et 'TaxRate'. Nous utilisons ensuite le tokenizer introduit dans la section 3.1 pour segmenter les textes en unités minimales de tokens, où les boîtes de délimitation du texte sont également segmentées en conséquence. L'ensemble de données utilisé dans ce travail est beaucoup plus difficile que les images de documents scannés soignées, puisque différentes mises en page de reçus ont été capturées dans différents scénarios avec des téléphones mobiles. Des exemples d'images de documents scannés dans notre jeu de données sont illustrés à la figure 2. Notez que les rectangles colorés ne servent qu'à des fins de référence visuelle et que les données étiquetées sont au niveau des jetons plutôt qu'au niveau des lignes comme indiqué dans la figure.

Performance globale

Nous présentons les résultats de notre méthode en termes de PA et les comparons avec d'autres méthodes de pointe dans le tableau 2. Nous présentons également les résultats de la softAP pour CUTIE-A et CUTIE-B dans le tableau 2, où la softAP de CUTIE-A et CUTIE-B dépasse largement leur AP. Des exemples de résultats d'inférence sont illustrés à la figure 3. Notre grand réseau CUTIE-A obtient 90,8 % de PA et 97,2 % de softAP sur les reçus de taxi, 77,7 % de PA et 91,4 % de softAP sur les reçus de repas et de divertissement, et 69,5 % de PA et 87,8 % de softAP sur les reçus d'hôtel. Par rapport à CloudScan, CUTIE-A et CUTIEB sont plus performants dans chaque cas de test. De plus, par rapport à BERT pour NER, qui est pré-entraîné sur un jeu de données de 3 300 millions de mots et appris par transfert sur notre jeu de données, notre grand réseau CUTIE-A améliore la PA de 2,7 % sur les reçus de taxi mais est moins précis sur les autres types de documents en utilisant seulement 1/2 paramètre ; notre petit réseau CUTIE-B améliore la PA de 5. 9 % sur les reçus de taxi, de 1,4 % sur les reçus de repas et de divertissement et de 2,9 % sur les reçus d'hôtel, dépassant ainsi les autres méthodes dans tous les cas de test, mais avec une complexité bien moindre et une taille de modèle plus petite, avec

seulement 1/9 des paramètres, et sans nécessiter un énorme ensemble de données pour le pré-entraînement du modèle. Nous prouverons dans la section 4.3.2 que CUTIE-B est également capable d'atteindre des performances de pointe avec seulement 1/10 des paramètres de BERT. Bien que CUTIE-B soit plus petite en capacité, elle surpasse CUTIE-A dans plusieurs critères d'évaluation. Cela est dû au fait que CUTIE-B élargit le champ de vision en employant la convolution atrous plutôt que les processus de pooling ou de striding, le modèle CUTIE-B a un plus grand champ de vision et une meilleure compréhension de la relation spatiale relative des tokens puisqu'aucune restriction n'est appliquée sur les formes des cartes de caractéristiques. Il est intéressant de noter que la différence entre AP et softAP conduit à des résultats intéressants. L'un de ces résultats est que CUTIE est capable d'extraire les textes intéressés mais implique parfois des textes qui ne sont pas dans la vérité terrain. Une autre constatation est que les reçus d'hôtel sont très différents des reçus de repas et des reçus de taxi, où les informations clés apparaissent plusieurs fois dans différentes zones du reçu, alors que les étiqueteurs humains ont tendance à n'étiqueter qu'une seule de leurs apparitions. Nous approfondissons ce point dans la partie suivante de cette section en analysant certains cas de résultats d'inférence. La figure 4 présente des exemples typiques de reçus présentant un score AP faible mais un score softAP élevé. La plupart des faux positifs se produisent dans la classe " VenderName ", où les noms ont tendance à varier considérablement, ce qui rend l'inférence du modèle difficile. Cependant, il n'est pas difficile de constater que ces faux positifs peuvent être facilement évités en ajoutant un post-processeur basé sur un dictionnaire à l'extracteur d'informations clés. Un cas rare de faux positif, le troisième reçu de la première rangée, est une lettre 'L' reconnue à tort comme '1' par le moteur OCR employé. Bien que la lettre soit distincte des autres chiffres dans l'image numérisée du reçu, elle est mal interprétée comme faisant partie de la classe "BaseAmount" en raison de sa proximité spatiale avec les chiffres et du fait qu'elle est elle-même un chiffre. On peut également voir sur le premier reçu de la deuxième ligne que plusieurs chiffres éloignés dans l'espace ont été prédits à tort comme faisant partie de la classe "BaseAmount". Bien que ces cas soient rares dans notre ensemble de tests, ils suggèrent que l'incorporation d'informations au niveau de l'image peut améliorer la précision de l'inférence, malgré les caractéristiques sémantiques et spatiales déjà impliquées. En outre, nous constatons que certains cas erronés sont en fait corrects, car les vérités de base ont été mal étiquetées par les étiqueteurs humains. Comme l'illustre la figure 5, le "VendorName" apparaît deux fois dans l'image scannée mais n'est étiqueté qu'une fois dans le premier et le second reçu de la première ligne, alors que le modèle interprète correctement les deux occurrences comme étant le "VendorName" dans le premier reçu et interprète correctement la seconde occurrence dans le second reçu. En outre, l'étiquette "TaxRate" est négligée dans le troisième reçu de la première ligne et l'étiquette "VendorName" est erronée dans le premier reçu de la deuxième ligne. Le troisième reçu de la deuxième ligne et les premier et deuxième reçus de la troisième ligne sont tous négligés avec l'étiquetage "TaxRate", et le troisième reçu de la troisième ligne est négligé avec l'étiquetage "BaseAmount" alors que le modèle entraîné a correctement déduit la bonne classe. Il n'est pas difficile de constater que le modèle entraîné produit même de meilleurs résultats que l'étiqueteur humain sur ces reçus, ce qui prouve encore l'efficacité de la méthode proposée. Nous évaluons également le modèle CUTIE-B sur l'ensemble d'entraînement ICDAR 2019 SROIE et listons les résultats dans le tableau 5. Le résultat au niveau de la classe indique que le résultat inféré est parfaitement identique à la vérité de base, tandis que le résultat au niveau de la classe molle indique que le résultat inféré incorpore la vérité de base mais a également inclus certains tokens de texte non liés.

Le résultat au niveau des jetons indique la précision de la classification des jetons de chaque classe malgré la classe "DontCare".

Études sur l'ablation

Bien que nous ayons démontré des résultats empiriques extrêmement solides, les résultats présentés sont obtenus par la combinaison de chaque aspect du cadre CUTIE. Dans cette section, nous effectuons des études d'ablation sur un certain nombre de facettes de CUTIE afin d'avoir une meilleure compréhension de leur importance relative. CUTIE-B est utilisé comme modèle par défaut avec une augmentation de la grille, la taille d'intégration étant de 128. Dans cette section, nous utilisons toutes les données de ME, dont 75 % sont utilisées pour l'entraînement du modèle et les 25 % restants pour le test.

Effet de l'augmentation de la grille sur la compréhension des informations spatiales

L'une de nos principales affirmations est que la haute performance de CUTIE est obtenue par l'analyse conjointe des informations sémantiques et spatiales avec le cadre proposé. La capacité d'analyse spatiale hautement efficace est permise par le cadre CUTIE, contrairement aux méthodes précédentes basées sur les NER. Le processus d'augmentation de la grille améliore encore cette capacité. Pour étayer cette affirmation, nous évaluons CUTIE avec ou sans le processus d'augmentation de la grille afin de tester les performances du modèle en termes de diversité spatiale. Les résultats sont présentés dans le tableau 6. Nous pouvons constater que l'ajout du processus d'augmentation de la grille dans CUTIE améliore considérablement les performances. Ces résultats démontrent que CUTIE peut grandement bénéficier de l'amélioration de la diversité des données dans la distribution spatiale. Pour cette raison, une analyse plus approfondie des techniques d'augmentation de la grille peut améliorer les performances de CUTIE, par exemple en déplaçant aléatoirement certains textes vers le haut, le bas, la gauche ou la droite de plusieurs pixels pendant le processus de mappage de la position de la grille.

Impact de la capacité d'information sémantique

Ensuite, nous évaluons l'impact de l'information sémantique de CUTIE en comparant l'évaluation avec différentes tailles d'incorporation de mots. Comme le montre le tableau 3, les performances de CUTIE suivent une courbe en forme de cloche lorsque la taille d'incorporation augmente. La meilleure performance est obtenue par CUTIE-B avec 85,0 % en AP et 92,9 % en softAP en utilisant une taille d'incorporation de 64. Une découverte intéressante est que le modèle CUTIE obtient de bonnes performances même avec une capacité d'information sémantique limitée. Nous en déduisons que, pour le problème SROIE avec 9 classes d'informations clés, une petite quantité de jetons clés contribue majoritairement à l'inférence du modèle et que le modèle peut obtenir de bonnes performances en accordant une attention particulière à ces jetons clés. Le résultat indique également qu'une taille d'incorporation trop importante diminue également les performances du modèle.

Impact du nombre d'échantillons d'entraînement

Pour évaluer l'impact de l'utilisation de différents nombres d'échantillons d'entraînement sur les performances du modèle, nous entraînons CUTIE avec 3 %, 12 %, 21 %, 30 %, 39 %, 48 %, 57 %, 66 % et 75 % de notre ensemble de données et nous présentons les résultats dans le tableau 4. Il n'est pas difficile de constater, d'après les résultats, qu'un plus grand nombre de données d'entraînement entraîne de meilleures performances, que ce soit en termes de PA ou de softAP. CUTIE-B obtient le meilleur PA 85,0% avec 66% de l'ensemble de données comme échantillons d'entraînement et le meilleur softAP 91,5% avec 75% de l'ensemble de données comme échantillons d'entraînement. Il est intéressant de noter que CUTIE-B est déjà capable d'atteindre 79,6 % de PA et 88,7 % de softAP avec seulement 21 % de l'ensemble de données comme échantillons d'entraînement, ce qui prouve l'efficacité de la méthode proposée, capable d'obtenir de bons résultats avec une quantité limitée de données d'entraînement.

Autres expériences

Nous présentons ici un autre ensemble de données comprenant 613 échantillons d'images où un jeton de texte est étiqueté pour indiquer s'il est contenu ou non dans une région de tableau et à quelle colonne de tableau il appartient. Pour tester la performance de la méthode proposée sur différentes tâches, deux modèles CUTIE-B sont entraînés, l'un pour localiser la région de la table dans les images de documents et l'autre pour identifier la colonne spécifique à laquelle le jeton appartient. 463 échantillons sont utilisés pour la formation et 150 échantillons pour le test. Pour le test de localisation de la région de la table, les résultats expérimentaux montrent que 86 % des tables sont parfaitement localisées, la plupart des erreurs étant dues à l'incorporation incorrecte de jetons qui n'appartiennent pas à la véritable région de la table, comme l'illustre la figure 6. Pour le test d'identification de colonne, 94,8% des tokens sont parfaitement inférés, comme l'illustre la Figure 7.

Discussion

L'extraction automatique de mots ou d'informations intéressants à partir d'images de documents numérisés présente un grand intérêt pour divers services et applications. Cet article propose CUTIE pour résoudre ce problème sans nécessiter de pré-formation ou de post-traitement. Les résultats expérimentaux vérifient l'efficacité de la méthode proposée. Contrairement aux méthodes précédentes, la méthode proposée est facile à entraîner et nécessite une quantité beaucoup plus faible de données d'entraînement tout en atteignant des performances de pointe avec une vitesse de traitement allant jusqu'à 50 échantillons par seconde. Le gain de performance est principalement obtenu en explorant trois facteurs clés : la relation spatiale entre les textes, les informations sémantiques des textes et le mécanisme de mappage positionnel de la grille. Un résultat intéressant est que le modèle CUTIE formé prédit correctement certaines informations clés qui sont négligées par l'étiqueteur humain, ce qui prouve l'efficacité de la méthode proposée. Il convient également de mentionner que, comme le montrent les résultats expérimentaux, l'incorporation d'un meilleur module de traitement des caractéristiques sémantiques ou l'utilisation de caractéristiques au niveau de l'image pourraient améliorer les performances du modèle.

Structure du modèle

Nous présentons le modèle CUTIE-B dans le tableau 7. Les jetons sont d'abord intégrés dans des caractéristiques à 128 dimensions. Ensuite, 4 opérations de convolution consécutives sont effectuées dans le bloc conv avec un pas de 1 et 4 convolutions atrous consécutives sont effectuées dans le bloc conv atrous avec un pas de 1 et un taux de 2. Après le bloc de convolution atrous, un module ASPP est utilisé pour fusionner les caractéristiques multi-résolution. La caractéristique de bas niveau mais de haute résolution de la première sortie du bloc de convolution est également ajoutée au modèle dans la couche de raccourci avec une opération de concaténation et une convolution 1 × 1. Enfin, la sortie d'inférence est obtenue par une convolution 1 × 1.