CUTIE : Apprendre à comprendre les documents avec un extracteur d'informations textuelles universel convolutif

Xiaohui Zhao, Endi Niu, Zhuo Wu et Xiaoguang Wang
New IT Accenture
xh.zhao@outlook.com henryniu14@gmail.com zhuo.wu@accenture.com danny.x.wang@accenture.com

Résumé

L'extraction d'informations clés à partir de documents, tels que des reçus ou des factures, et la préservation des textes intéressés sous forme de données structurées est cruciale dans les processus de bureautique à forte intensité documentaire dans des domaines tels que la comptabilité, la finance et la fiscalité, sans toutefois s'y limiter. Pour éviter de concevoir des règles expertes pour chaque type de document spécifique, certains travaux publiés tentent de résoudre le problème en apprenant un modèle pour explorer le contexte sémantique dans les séquences de texte sur la base de la méthode de reconnaissance des entités nommées (NER) dans le domaine du NLP. Dans cet article, nous proposons d'exploiter les informations efficaces provenant à la fois de la signification sémantique et de la distribution spatiale des textes dans les documents. Plus précisément, le modèle que nous proposons, Convolutional Universal Text Information Extractor (CUTIE), applique des réseaux neuronaux convolutionnels à des textes en grille où les textes sont intégrés en tant que caractéristiques avec des connotations sémantiques. Nous étudions en outre l'effet de l'utilisation de différentes structures de réseaux neuronaux convolutifs et proposons une structure rapide et portable. Nous démontrons l'efficacité de la méthode proposée sur un ensemble de données comprenant jusqu'à 4 484 *reçus étiquetés, sans aucun pré-entraînement ni post-traitement, en obtenant des performances de pointe qui sont bien meilleures que les méthodes basées sur le NER en termes de vitesse et de précision. Les résultats expérimentaux démontrent également que le modèle CUTIE proposé est capable d'atteindre de bonnes performances avec une quantité beaucoup plus faible de données d'entraînement.*

1. Introduction

La mise en œuvre de l'OCR des reçus numérisés et de l'extraction d'informations (SROIE) est très utile pour les services et les applications tels que l'archivage efficace, le contrôle de la conformité et l'indexation rapide dans les processus de rationalisation de la bureautique à forte intensité documentaire dans des domaines tels que la comptabilité, la finance et la fiscalité, sans toutefois s'y limiter. Deux tâches spécifiques sont impliquées dans le SROIE : l'OCR des reçus et l'extraction des informations clés. Dans ce travail, nous nous concentrons sur l'OCR des reçus et l'extraction des informations clés.

L'extraction d'informations clés est une deuxième tâche qui est rare dans les recherches publiées. En fait, l'extraction d'informations clés est confrontée à de grands défis, où les différents types de structures de documents et le grand nombre de mots clés potentiellement intéressants introduisent de grandes difficultés. Bien que la méthode à base de règles couramment utilisée puisse être complétée par des règles d'experts soigneusement conçues, elle ne peut fonctionner que sur certains types de documents spécifiques et nécessite un effort non négligeable pour s'adapter à de nouveaux types de documents. Il est donc souhaitable de disposer d'une méthode d'extraction d'informations clés basée sur l'apprentissage qui nécessite peu de ressources humaines et qui utilise uniquement la technique d'apprentissage profond sans concevoir de règle d'expert pour un type de document spécifique.

CloudScan est un système d'analyse de factures basé sur l'apprentissage [9]. Afin de ne pas dépendre des modèles de présentation des factures, CloudScan forme un modèle qui pourrait être généralisé à des présentations vocales inédites, où un modèle est formé en utilisant la mémoire à long terme (LSTM) ou la régression logistique (LR) avec des règles conçues par des experts comme caractéristiques d'apprentissage. Cela s'avère extrêmement similaire à la résolution d'une tâche de reconnaissance d'entités nommées (NER) ou de remplissage d'emplacements. Pour cette raison, plusieurs modèles peuvent être utilisés, par exemple le modèle BERT (Bi-directional En- coder Representations from Transformers) est une méthode de pointe récemment proposée et qui a remporté un grand succès dans un large éventail de tâches de NLP, y compris la NER [7]. Toutefois, les modèles NER n'ont pas été conçus à l'origine pour résoudre le problème de l'extraction d'informations clés dans la SROIE. Pour utiliser les modèles NER, les mots du document original sont alignés comme un long paragraphe sur la base d'une règle basée sur les lignes. En fait, les documents, tels que les reçus et les factures, présentent différents styles de mise en page qui ont été conçus pour différents scénarios ou pour différentes entités de l'entreprise. L'ordre ou la distance mot à mot des textes dans le long paragraphe aligné à la ligne a tendance à varier considérablement en raison des variations de mise en page, ce qui est difficile à traiter avec les méthodes orientées vers le langage naturel. Des exemples typiques de documents avec des mises en page différentes sont illustrés à la figure 2.

Dans ce travail, nous tentons d'impliquer les informations spatiales dans le processus d'extraction des informations clés.

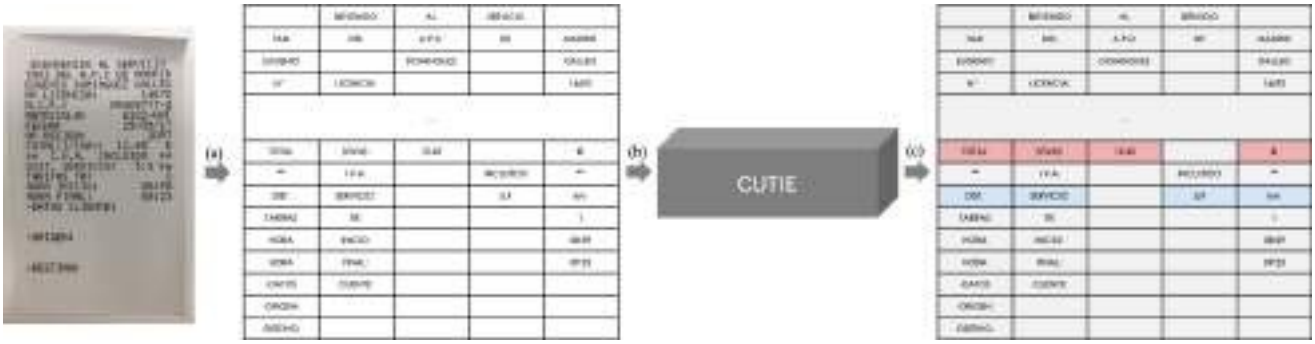


Figure 1. Cadre de la méthode proposée : (a) cartographie positionnelle de l'image du document numérisé en une grille préservant la relation spatiale relative du texte, (b) introduction de la grille générée dans le CNN pour l'extraction des informations clés, (c) cartographie inversée des informations clés extraites pour la référence visuelle.

Nous proposons d'aborder ce problème en utilisant la structure de réseau basée sur le CNN et d'impliquer les caractéristiques sémantiques d'une manière soigneusement conçue. En particulier, le modèle que nous proposons, appelé Convolutional Universal Text Information Extractor (CUTIE), s'attaque au problème de l'extraction d'informations clés en appliquant un modèle d'apprentissage profond convolutionnel sur les textes quadrillés, comme illustré à la figure 1. Les textes quadrillés sont formés à l'aide de la méthode proposée de mappage positionnel de la grille, où la grille est générée selon le principe de la préservation de la relation spatiale relative des textes dans l'image originale du document numérisé. L'information sémantique riche est encodée à partir des textes quadrillés au tout début du réseau neuronal convolutionnel avec une couche d'incorporation de mots. CUTIE permet d'examiner simultanément les informations sémantiques et les informations spatiales des textes dans l'image du document scanné et peut atteindre un nouveau résultat de pointe pour l'extraction d'informations clés, qui surpasse le modèle BERT, mais sans nécessiter de préformation sur un énorme ensemble de données textuelles [7, 12].

2. Travaux connexes

Plusieurs systèmes d'analyse de factures basés sur des règles ont été proposés dans [10, 6, 8]. Intellix de DocuWare exige qu'un modèle soit annoté avec les champs pertinents [10]. Pour cette raison, une collection de modèles doit être construite. SmartFix utilise des règles de configuration spécifiquement conçues pour chaque modèle [6]. Esser et al. utilise un ensemble de données de positions d'informations clés fixes pour chaque modèle [8]. Il n'est pas difficile de constater que les méthodes basées sur des règles s'appuient fortement sur les règles de configuration prédéfinies pour extraire des informations à partir de modèles de factures spécifiques.

des post-traitements sont ajoutés pour améliorer encore les résultats de l'extraction. Toutefois, la méthode d'extraction des caractéristiques basée sur les lignes ne peut pas atteindre ses meilleures performances lorsque les textes des documents ne sont pas parfaitement alignés. En outre, le modèle LSTM bidirectionnel de CloudScan, basé sur la classification RNN, a une capacité limitée à apprendre la relation entre des mots éloignés.

Bidirectional Encoder Representations from Transformers (BERT) est un modèle récemment proposé qui est pré-entraîné sur un énorme ensemble de données et peut être affiné pour une tâche spécifique, y compris la reconnaissance des entités nommées (NER), qui surpasse la plupart des résultats de l'état de l'art dans plusieurs tâches NLP [7]. Étant donné que les méthodes précédentes basées sur l'apprentissage traitent le problème d'extraction d'informations clés comme un problème de reconnaissance d'entités nommées, l'application de BERT peut obtenir un meilleur résultat que la bi-LSTM dans CloudScan.

3. Méthodes

Dans cette section, nous présentons la méthode proposée pour créer des données de grille pour l'apprentissage du modèle. Nous présentons ensuite les architectures de réseau qui capturent l'information à longue distance et évitent la perte d'information dans les réseaux neuronaux convolutionnels qui ont des processus de striding ou de pooling.

3.1. Cartographie de la position de la grille

Pour générer des données de grille d'entrée pour le réseau neu- ral convolutionnel, les images de documents scannés sont traitées par un moteur OCR afin d'acquérir les textes et leur valeur absolue.
/ positions relatives. Si l'image du document numérisé est de forme (w, h), la boîte de délimitation minimale autour du i-ième texte intéressé$_i$ est b_i qui est limitée par deux coordonnées cor- ner, où la coordonnée du coin supérieur

*gauche dans la boîte de dé*limitation minimale est b qui est limitée par deux coordonnées

CloudScan est un travail qui tente d'extraire des informations clés.

tion avec des modèles basés sur l'apprentissage [9].

Premièrement, les N-grammes fea-

Les structures sont formées en connectant des règles conçues par des experts, cal-

Les résultats calculés sur les textes de chaque ligne de document. Ensuite, les caractéristiques sont utilisées pour former un RNN ou un régresseur logistique.

La classification basée sur la vision pour l'extraction d'informations clés. Il n' est pas difficile de constater que l'implication de prétraitements qui combinent

cor- ner, où la coordonnée du coin supérieur gauche dans la boîte de dé*limitation minimale est b* .

le document numérisé soit $_{top}^{i}$) et en bas à droite

(x^i

de la boîte englobante soit , $^{i\,fond}$). Pour éviter l'af-

(x^i y

des boîtes englobantes superposées et révèlent les effets réels des $_{droit}$

boîtes englobantes superposées et révèlent les effets réels des boîtes englobantes superposées.

position relative entre les textes, nous calculons le point central

(c^i , c^i) des boîtes englobantes comme position de référence. Il

x y

La transformation des textes en entités significatives sera bénéfique pour le processus de cartographie positionnelle de la grille. Cependant, ce n'est pas l'objectif principal de cet article et nous le laissons aux recherches futures. Dans cet article, les mots du texte sont transformés en jetons à l'aide d'un algorithme gourmand de la plus longue correspondance en utilisant un dictionnaire prédéfini [1].

Soit le processus de mappage positionnel de la grille G et la taille de la grille tar- get (c_g m , r_g m). Pour générer les données de la grille, l'objectif de G est de mapper les textes de l'image du document scanné original à la grille cible, de sorte que la grille mappée préserve la relation spatiale originale entre les textes tout en étant plus appropriée pour être utilisée en tant qu'entrée pour le processus convolutionnel de mappage.
réseau neuronal. La position des textes dans la grille est calculée comme suit

$$c_x^i = c_{g\,m} \frac{xgauche + \frac{(xdroite-xgauche)}{2}}{w} \quad (1)$$

$$r_y^i = r_{g\,m} \frac{y + \frac{(ybottom-ytop)}{2}}{h} \quad (2)$$

Pour les textes tokenisés, la boîte de délimitation est divisée horizontalement en plusieurs boîtes et leurs positions de référence en ligne et en colonne sont calculées en utilisant les mêmes critères que ceux de l'Equ. 1 et Equ. 2, séparément. En outre, afin d'améliorer la capacité de CUTIE à mieux gérer les documents avec différentes mises en page, nous augmentons les données de la grille pour obtenir des formes avec différentes lignes et colonnes en échantillonnant au hasard une distribution gaussienne avec la probabilité suivante

$$p_c(k) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(k-c_g)^2}{2\sigma^2}} \quad (3)$$

$$p_r(k) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(k-r_g)^2}{2\sigma^2}} \quad (4)$$

où c_g t est le centre moyen de la taille de la grille d'augmentation cible, r_g t est le centre moyen de la taille de la grille d'augmentation cible et σ est l'écart type. Si deux jetons occupent la même cellule de la grille, les jetons de la même ligne sont déplacés pour les placer. Ceci est acceptable puisque le modèle apprend la relation spatiale relative entre les jetons et que les réseaux neuronaux convolutifs gèrent bien les translations.

3.2. Modèle CUTIE

En faisant correspondre la sortie de CUTIE avec les données de la grille étiquetée, le modèle apprend à générer l'étiquette pour chaque texte dans l'entrée de la grille en

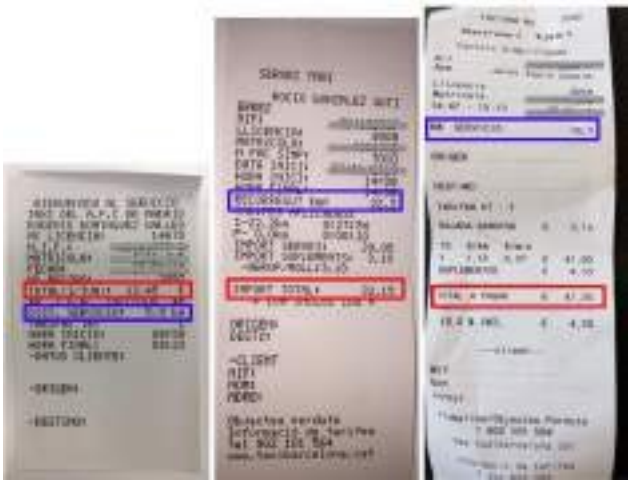


Figure 2. Exemple d'images de reçus de taxi scannés. Nous fournissons deux rectangles de couleur pour aider les lecteurs à trouver les informations clés sur la distance du trajet et le montant total avec le bleu et le rouge, re-
de manière sélective. Notez les différents types d'agencements spatiaux et de textes de formation clés dans ces images de réception.

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 provenant de différentes
Les échelles deviennent proéminentes dans les premiers réseaux en utilisant des données d'entrée à échelles multiples pour rassembler des caractéristiques multi-échelles.
Les derniers réseaux rétrécissent les cartes de caractéristiques pour agrandir les cartes réceptives.
et réduire les charges de calcul, puis capturer des détails plus fins en récupérant progressivement les informations spatiales. explorant les caractéristiques spatiales et sémantiques. C'est pourquoi 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 dispersés (jetons de texte), contrairement aux images parsemées de pixels. Les textes clés cartographiés en fonction de la position de 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 contextuel multi-échelle est bénéfique pour le réseau.

à partir des caractéristiques des couches inférieures. Cependant, la résolution spatiale est réduite dans le processus de codage et le processus de décodage n'exploite que les caractéristiques à haute résolution mais de bas niveau pour couvrir à nouveau la résolution spatiale, le processus de codage par stries consécutives décimant les informations détaillées [11]. En outre, le processus de codage et de décodage applique des restrictions de forme au processus d'augmentation de la forme de la grille, tel qu'il a été présenté à la section 3.1.

En revanche, le champ de vision des filtres peut également 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 aqueuse [2, 3, 4, 5]. Pour capturer les connexions à longue distance et éviter la perte potentielle d'informations 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 différents types de structures de modèle et nous n'en détaillons que deux ici pour éviter un article fastidieux. Plus précisément, le modèle CUTIE-A proposé est un réseau neuronal con- volutionnel à haute capacité qui fusionne les caractéristiques multi-résolution sans perdre les caractéristiques haute résolution, le modèle CUTIE-B proposé est un réseau convolu- tionnel avec convolu- tion atruse et un réseau convolu- tionnel à haute capacité.

Tableau 1. Statistiques sur le nombre d'images de documents de réception étiquetés et de classes d'informations clés dans l'ensemble de données.

	Ensemble de formation	Ensemble de tests	#Classes
ME	1109	375	9
Taxi	1125	375	6
Hôtel	1125	375	9

tion pour agrandir le champ de vision et le module Atrous Spatial Pyra- mid Pooling (ASPP) pour capturer des contextes multi-échelles.

CUTIE-A et CUITE-B effectuent tous deux un processus d'encodage du sens sémantique avec une couche d'intégration de mots au tout début. L'exclusion est appliquée à la couche d'intégration afin d'améliorer la capacité de généralisation de CUTIE. La fonction de perte d'entropie croisée est appliquée pour comparer la grille de classe de jetons prédite et la grille de vérité de terrain.

3.2.1 CUTIE-A

CUTIE-A évite la perte d'informations dans le processus de codage tout en tirant parti des codeurs en combinant les résultats du codage aux représentations haute résolution maintenues tout au long du processus convolutif. Comme pour le HRNet proposé dans [11], un réseau à haute résolution sans striding est utilisé comme réseau de base et plusieurs sous-réseaux de haute à basse résolution sont progressivement ajoutés et connectés au réseau principal de base. Au cours du processus de connexion du réseau principal et des sous-réseaux, les caractéristiques à plusieurs échelles sont fusionnées pour générer des représentations riches.

3.2.2 CUTIE-B

CUTIE-B est construit avec un seul réseau dorsal mais utilise la convolution atrous pour capturer les connexions à 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 à la carte de caractéristiques d'entrée *m sous la* forme suivante

$$n[i] = \sum_{k} m[i + r - k]w[k] \qquad (5)$$

où r est le taux d'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 ze- ros entre deux valeurs de filtre consécutives le long de chaque dimension spatiale. La convolution standard est un cas particulier de convolution atrous avec $r = 1$ [3].

4. Expériences

La méthode proposée est évaluée sur le défi de lecture de buste ICDAR 2019 sur l'ensemble de données SROIE et également sur un ensemble de données auto-construit avec 3 types d'images de documents numérisés,

qui contiennent 8 classes d'informations clés et 1 classe d'indifférence. Pour chaque classe d'informations clés spécifique, plusieurs to- kens peuvent être inclus. La performance globale est appelée précision moyenne stricte (AP) et mesurée en termes de précision par classe sur les 9 classes, une classe n'étant considérée comme correcte que si tous les jetons de la classe sont corrects. Pour approfondir l'analyse des performances de la méthode proposée, nous proposons d'utiliser un autre critère, la précision moyenne souple (softAP), où la prédiction d'une classe d'information clé est considérée comme correcte si les vérités de terrain positives sont correctement prédites, même si certains faux positifs sont inclus dans la prédiction finale. Le SoftAP est important car il indique la capacité du modèle à extraire des informations clés correctes tout en tolérant l'incorporation de certains faux positifs. En fait, des post-traitements peuvent être utilisés pour éliminer les faux positifs. Par conséquent, l'analyse conjointe de l'AP et du softAP permet de mieux comprendre les performances du modèle.

Nous comparons les performances de la méthode proposée avec deux méthodes de pointe, CloudScan [9] et BERT pour NER [7]. À titre de comparaison, le modèle Cloud Scan pour SROIE est formé à partir de zéro mais avec plusieurs caractéristiques dé- signées par des experts comme décrit dans [9]. Le modèle BERT pour SROIE est transformé avec le modèle de base publié par Google qui est pré- entraîné sur un énorme ensemble de données avec 3 300 millions de mots [7, 1]. Pour une comparaison équitable, 4 484 échantillons, dont environ 1 *500* pour les taxis, les repas et divertissements (ME) et les hôtels, dont 3 375 échantillons sont utilisés pour la formation et 1 125 pour les tests, comme indiqué dans le tableau 1. Un seul modèle est formé sur l'ensemble 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 par décroissance. Le taux d'apprentissage est réduit à *1e* 4 et *1e5* aux 15 000e et 30 000e étapes, respectivement. L'apprentissage est terminé après 40 000 étapes avec une taille de lot de 32. Nous entraînons 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 0,9. Notre modèle est entraîné de bout en bout sans pré-entraînement par morceaux d'aucun composant. La taille d'incorporation des défauts 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 de formation et un ensemble de test dans un rapport de 75 : 25. Aucun prétraitement CUTIE ne fait l'objet d'aucun post-traitement.

4.1. Ensemble de données

L'ensemble de données SROIE de l'ICDAR 2019 (tâche 3) contient 1000 images de reçus numérisés. 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 l'ensemble de données se compose principalement de chiffres et de caractères anglais.

Tableau 2. Comparaison des performances sur différents types de documents. (AP/softAP)

Méthode	#Params	Taxi	ME	Hôtel
CloudScan[9]	-	82 / -	64 / -	60 / -
BERT pour NER[7]	110M	88.1 / -	80.1 / -	71.7 / -
CUTIE-A	67M	90.8 / 97.2	77.7 / **91.4**	69.5 / **87.8**
CUTIE-B	14M	**94.0** / **97.3**	**81.5** / 89.7	**74.6** / 87.0

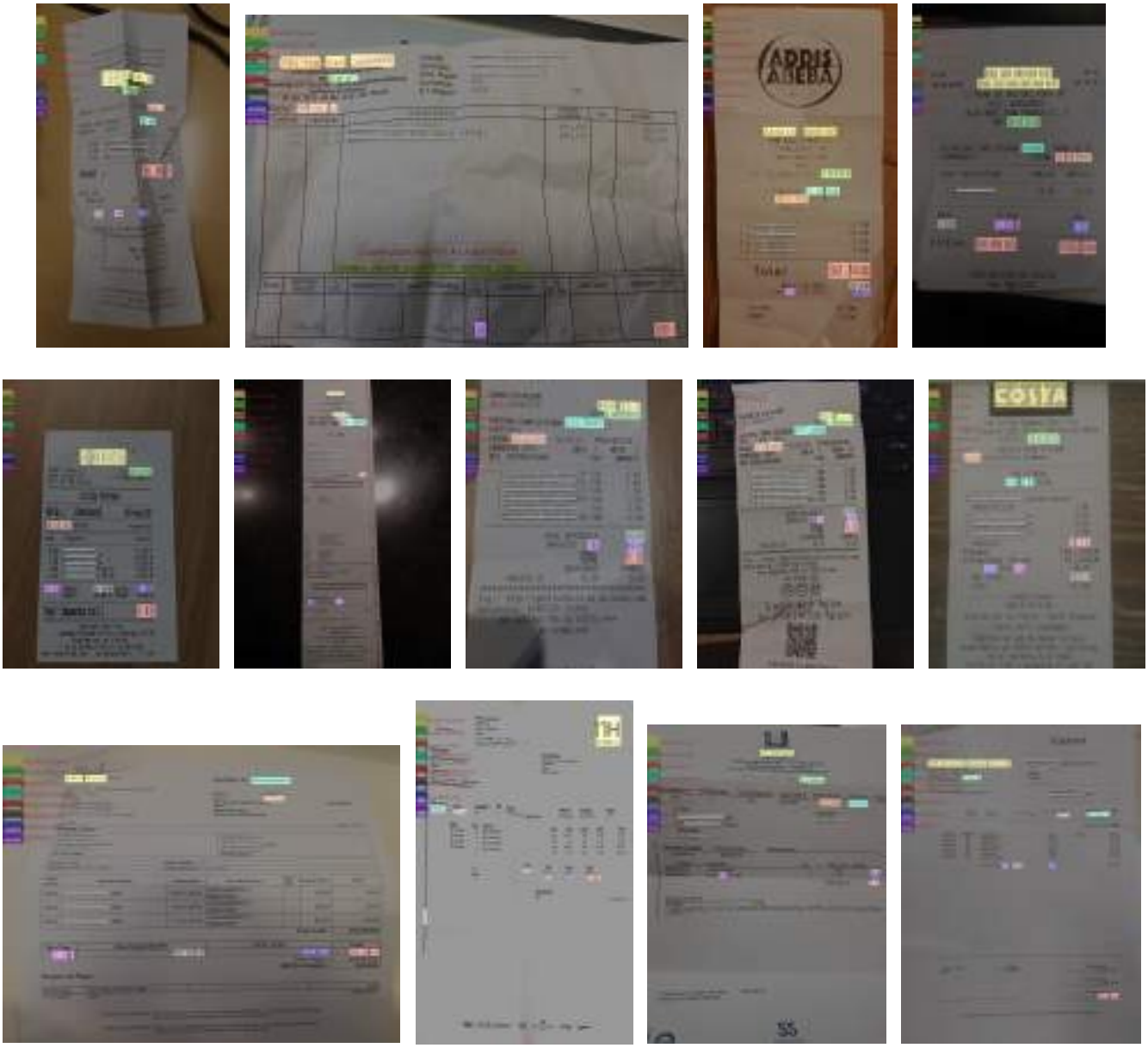


Figure 3. Exemple de résultats de l'inférence CUITE. La légende des couleurs dans le coin supérieur gauche indique les classes d'informations clés. Chaque couleur indique une classe d'information clé, où les rectangles remplis sont les vérités de base tandis que les rectangles délimités sont les résultats de l'inférence. Le résultat est parfaitement correct si les rectangles remplis se chevauchent avec les rectangles délimités. Dans ces figures, nous masquons certaines informations privées à l'aide de rectangles gris remplis. (zoomer pour vérifier les détails)

L'ensemble de données est divisé en un ensemble de formation/validation (trainval) et un ensemble de test (test). L'ensemble d'entraînement se compose 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 validée, nous effectuons une lessive sur l'ensemble de formation et sélectionnons 517 échantillons, dont les 55 échantillons filtrés ont été mal étiquetés. Nous avons ensuite divisé l'ensemble de formation dans un rapport de 75 : 25, où 75 % sont utilisés comme données de formation et 25 % sont utilisés pour la validation.

L'ensemble de données auto-construit contient 4 484 documents de reçus espagnols scannés et annotés, y compris des reçus de taxi, des reçus de repas et de divertissements (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", "Ven- dorName", "VendorTaxID", "InvoiceDate", "InvoiceNum- ber", "ExpenseAmount", "BaseAmount", "TaxAmount", et "TaxRate". Nous utilisons ensuite le tokenizer présenté à la section 3.1 pour segmenter les textes en unités minimales de tokens, où les zones de délimitation du texte sont également segmentées en conséquence.

L'ensemble de données de ce travail est beaucoup plus difficile que les images de documents scannés, puisque diverses dispositions de reçus ont été capturées dans différents scénarios avec des téléphones mobiles. La figure 2 illustre des exemples d'images de documents scannés dans notre ensemble de données. Notez que les angles rectangulaires colorés servent uniquement de référence visuelle et que les données étiquetées réelles sont au niveau du jeton plutôt qu'au niveau de la ligne, comme indiqué dans la figure.

4.2. Performance globale

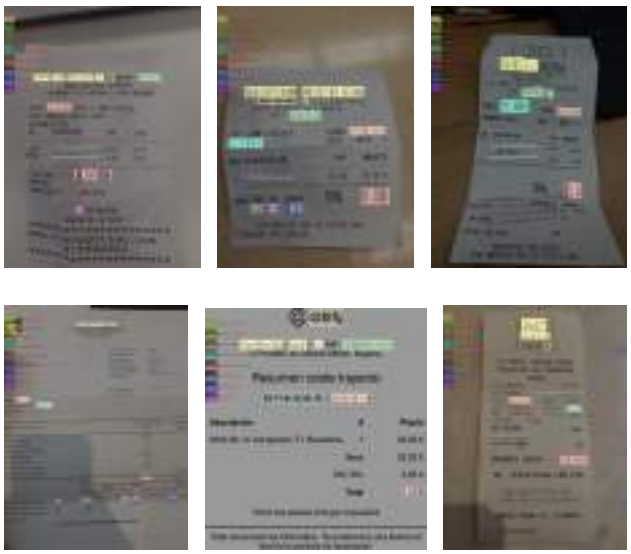
Nous présentons les résultats de notre méthode en termes de PA et les comparons avec d'autres méthodes de pointe dans le tableau 2. Nous fournissons également les résultats de la PA souple pour CUTIE-A et CUTIE-B dans le tableau 2, où la PA souple de CUTIE-A et CUTIE-B dépasse largement leur PA. Des 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 69,5 % de PA et 87,8 % de softAP sur les reçus d'hôtel. Par rapport à CloudScan, CUTIE-A et CUTIE-B sont plus performants dans tous les cas de test. De plus, comparé à BERT pour NER, qui est pré-entraîné sur un ensemble de données avec 3 300 millions de mots et qui a appris le transfert sur notre ensemble de données, notre grand réseau CUTIE-A améliore le PA de 2,7% sur les reçus de taxi mais est moins précis sur d'autres types de documents tout en utilisant seulement 1/2 paramètres ; notre petit réseau CUTIE-B im- prouve le PA de 5.9 % sur les reçus de taxi, 1,4 % sur les reçus de repas et 2,9 % sur les reçus d'hôtel, surpassant les autres méthodes dans tous les cas de test, mais avec beaucoup moins de complexité et une taille de modèle plus petite avec seulement *1/9* paramètres et sans nécessiter un énorme ensemble de données pour le préapprentissage du modèle. Nous prouverons en outre 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 petit en capacité, il surpasse CUTIE-A dans plusieurs critères d'évaluation. En effet, CUTIE-B élargit le champ de vision en employant la convolution atrous plutôt que les processus de pooling ou striding, le modèle CUTIE-B a un champ de vision plus large et une meilleure compréhension de la relation spatiale relative des jetons puisqu'aucune restriction n'est appliquée sur les formes des cartes de caractéristiques.

Il convient de noter que la différence entre AP et sof- tAP conduit à des résultats intéressants. L'une d'entre elles est que CUTIE est capable d'extraire les textes intéressés, mais qu'il implique parfois des textes qui ne figurent pas dans la vérité de 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 parties du reçu, alors que les étiqueteurs humains ont tendance à n'étiqueter qu'une seule de leurs apparitions. Nous approfondissons ce point dans la suite de cette section en analysant quelques cas de résultats d'inférence.

La figure 4 présente des exemples typiques de reçus ayant un score AP faible mais un score sof- tAP élevé. La plupart des cas de faux positifs se produisent dans la classe "VenderName", où les noms ont tendance à varier considérablement et entraînent des difficultés dans l'inférence du modèle. Toutefois, 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 ligne, est une lettre "L" reconnue à tort comme "1" par le moteur OCR utilisé. Bien que la lettre soit distincte des autres chiffres dans l'image numérisée du reçu, elle est interprétée à tort comme faisant partie de la classe "BaseAmount" en raison de sa proximité avec les chiffres et du fait qu'elle est elle-même un chiffre. On peut également voir dans 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 qu'il s'agisse de cas rares dans notre ensemble de tests, cela suggère que l'incorporation d'informations au niveau de l'image peut encore améliorer la précision de l'inférence malgré les caractéristiques sémantiques et spatiales déjà impliquées.

En outre, nous constatons que certains cas erronés sont en réalité corrects car les vérités de base ont été mal étiquetées par les étiqueteurs humains. Comme l'illustre la figure 5, le "Ven- dorName" apparaît deux fois dans l'image numérisée mais n'est étiqueté qu'une seule fois dans le premier et le deuxième reçu de la première ligne, alors que le modèle interprète correctement les deux occurrences comme le "VendorName" dans le premier reçu et interprète correctement la deuxième occurrence dans le deuxième reçu. En outre, le "TaxRate" n'est pas étiqueté dans le troisième reçu de la première ligne et le "VendorName" est étiqueté à tort 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 "VendorName".



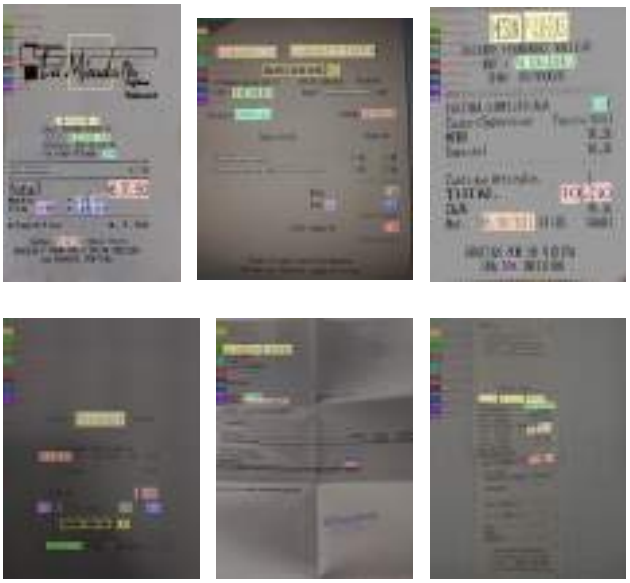


Figure 4. Exemples de faux positifs des résultats de prédiction de CUITE. La légende des couleurs dans le coin supérieur gauche indique les classes d'informations clés. Chaque couleur indique une classe d'informations clés, où les rectangles remplis sont les vérités de base tandis que les rectangles délimités sont les résultats de l'inférence. Les faux positifs sont les résultats pour lesquels les rectangles délimités ne sont pas recouverts par les rectangles remplis. Dans ces figures, nous masquons certaines informations privées par des rectangles gris remplis. (zoomer pour vérifier les détails)

'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 une fois de plus l'efficacité de la méthode proposée.

Nous évaluons également le modèle CUTIE-B sur l'ensemble de formation ICDAR 2019 SROIE et énumérons 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 souple indique que le résultat inféré incorpore la vérité de base, mais inclut également certains tokens de texte sans rapport. Le résultat au niveau des jetons indique la précision de la classification des jetons de chaque classe en dépit de la classe "Ne pas s'en préoccuper".

4.3. Études d'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 réalisons des études d'ablation sur un certain nombre de facettes de CUTIE afin de mieux comprendre leur importance relative. CUTIE-B est utilisé comme modèle par défaut avec une augmentation de la grille et une taille d'intégration de 128. Dans cette section, nous utilisons toutes les données de ME, dont 75 % sont utilisées pour former le modèle et les 25 % restants pour le tester.

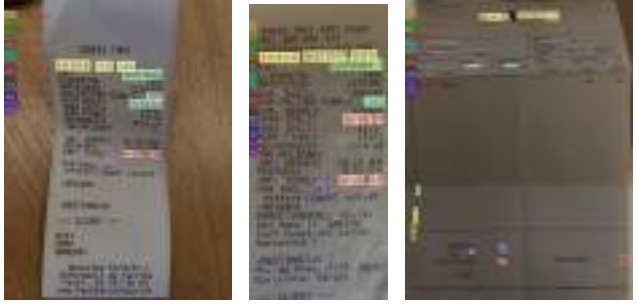


Figure 5. Exemples de faux étiquetage des résultats de prédiction de CUITE, où l'erreur est en fait causée par l'étiquetage erroné de l'étiqueteur humain. La légende des couleurs dans le coin supérieur gauche indique les classes d'informations clés. Chaque couleur indique une classe d'informations clés, où les rectangles remplis sont les vérités de base tandis que les rectangles délimités seulement sont les résultats de l'inférence. Dans ces figures, nous masquons certaines informations privées à l'aide de rectangles gris remplis. (zoomer pour vérifier les détails)

4.3.1 Effet de l'augmentation de la grille sur la compréhension de l'information spatiale

L'un de nos principaux arguments est que les performances élevées de CUTIE sont obtenues grâce à l'analyse conjointe des informations sémantiques et spatiales dans le cadre proposé. La capacité d'analyse spatiale très efficace est rendue possible par le cadre CUTIE, contrairement aux méthodes précédentes basées sur le NER. Le processus d'augmentation de la grille renforce encore cette capacité. Pour apporter plus de preuves à cette affirmation, nous évaluons CUTIE avec ou sans le processus d'augmentation de la grille pour tester la performance 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 de manière significative les performances du modèle en termes de diversité spatiale.

Tableau 3. Évaluation des performances de CUTIE sur ME avec différentes tailles d'intégration.

Taille d'intégration	1	2	4	8	16	32	64	128	256	512
#Params	10.6M	10.6M	10.7M	10.7M	10.9M	11.3M	12.1M	13.6M	16.6M	22.7M
AP	79.1	82.8	83.1	82.8	82.9	83.2	83.8	84.3	**84.6**	84.4
softAP	87.3	90.3	90.4	92.0	90.6	90.7	92.3	**92.4**	91.9	91.9

Tableau 4. Évaluation des performances de CUTIE-B sur ME avec différents nombres d'échantillons d'entraînement.

Pourcentage (%)	3	12	21	30	39	48	57	66	75
AP	56.2	76.4	79.6	80.8	82.6	81.4	83.4	**85.0**	84.3
softAP	76.0	86.5	88.7	90.3	90.5	90.7	90.7	91.1	**91.5**

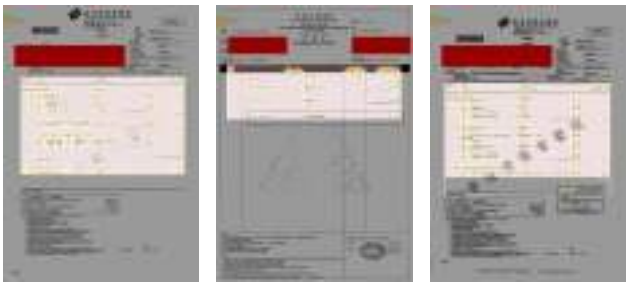
les performances. Ces résultats démontrent que CUTIE peut grandement bénéficier de l'amélioration de la diversité des données dans la distribution spatiale. C'est pourquoi une analyse plus approfondie des techniques d'augmentation de la grille peut améliorer les performances de CUTIE, par *exemple* en déplaçant de manière aléatoire certains textes vers le haut, vers le bas, vers la gauche ou vers la droite de plusieurs pixels au cours du processus de mappage positionnel de la grille.

4.3.2 Impact de la capacité d'information sémantique

Ensuite, nous évaluons l'impact des informations sémantiques de CUTIE en comparant l'évaluation avec différentes tailles d'encastrement de mots. Comme le montre le tableau 3, les performances de CUTIE suivent une courbe en forme de cloche au fur et à mesure que la taille de l'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. Il est intéressant de constater que le modèle CUTIE obtient de bonnes performances même avec une capacité d'information sémantique limitée. Nous en déduisons que la raison en est que, pour le problème SROIE avec 9 classes d'informations clés, il y a un petit nombre de jetons clés qui fournissent la contribution majoritaire 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 montre également qu'une taille d'intégration trop importante réduit également les performances du modèle.

4.3.3 Impact du nombre d'échantillons de formation

Pour évaluer l'impact de l'utilisation d'un nombre différent d'échantillons d'entraînement sur les performances du modèle, nous entraînons CUTIE avec 3 %, 12 %, 21 %, 30 %, 39 %, 48 %, 57 %, 66 % et 75 % de nos échantillons d'entraînement.



et les résultats sont présentés dans le tableau 4. Il n'est pas difficile de constater, d'après les résultats, qu'un plus grand nombre de données d'entraînement permet d'obtenir de meilleures performances, que ce soit en termes d'AP ou de softAP. CUTIE-B obtient le PA le plus élevé (85,0 %) avec 66 % de l'ensemble de données comme échantillons d'entraînement et le softAP le plus élevé (91,5 %) avec 75 % de l'ensemble de données comme échantillons d'entraînement. Il convient de noter que CUTIE-B est déjà capable d'atteindre 79,6% AP et 88,7% softAP avec seulement 21% de l'ensemble de données comme échantillons de formation, ce qui prouve encore l'efficacité de la méthode proposée qui est capable d'atteindre un niveau de performance élevé.

Figure 6. Localisation de la région de la table avec CUTIE-B. Les tokens de la table sont indiqués par des cadres jaunes et les régions de la table par des cadres roses. Les rectangles remplis représentent les vérités de base, tandis que les rectangles uniquement délimités représentent les résultats de l'inférence. Dans ces figures, nous masquons certaines informations privées à l'aide de rectangles rouges remplis. (zoomer pour vérifier les détails)

d'obtenir de bons résultats avec une quantité limitée de données d'apprentissage.

4.3.4 Autres expériences

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

Tableau 5. Évaluation des performances de CUTIE-B sur la tâche 3 de l'ICDAR 2019 SROIE.

	niveau de	niveau de la classe (soft)
niveau du jeton CUTIE-B	92.7%	94.2%
		86.7%

Tableau 6. Évaluation des performances de CUTIE sur ME avec ou sans le processus d'augmentation de la grille.

	sans augmentation	w augmentation
AP	83.3	84.3
softAP	93.7	92.4

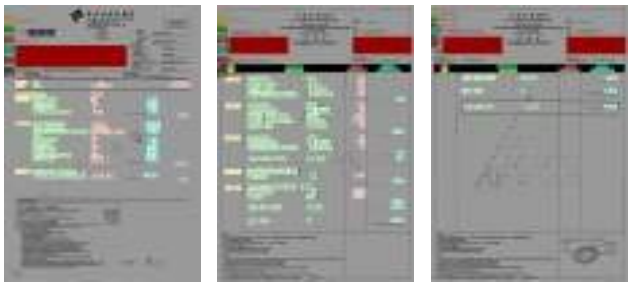


Figure 7. Identification des jetons de colonne avec CUTIE-B. La légende des couleurs dans le coin supérieur gauche indique l'identifiant de la colonne, tandis que les rectangles remplis représentent les vérités de base et les rectangles délimités les résultats de l'inférence. Dans ces figures, nous masquons certaines informations privées à l'aide de rectangles rouges remplis. (zoomer pour vérifier les détails)

5. Discussion

L'extraction automatique de mots ou d'informations intéressantes 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-entraînement 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 beaucoup moins 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, l'information sémantique des textes et le mécanisme de mappage positionnel de la grille. Il est intéressant de noter que le modèle CUTIE entraîné prédit correctement certaines informations clés que l'étiqueteur humain néglige d'étiqueter, 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 encore améliorer les performances du modèle, ce que nous laissons à la recherche future.

Références

[1]BERT. https://github.com/google-research/bert.

[2]L. Chen, G. Papandreou, I. Kokkinos, K. Murphy et A. L. Yuille. Semantic image segmentation with deep convolu- tional nets and fully connected crfs. *CoRR*, abs/1412.7062, 2014.

[3]L. Chen, G. Papandreou, I. Kokkinos, K. Murphy et A. L. Yuille. Deeplab : Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *CoRR*, abs/1606.00915, 2016.

[4]L. Chen, G. Papandreou, F. Schroff et H. Adam. Re-thinking atrous convolution for semantic image segmenta- *tion. CoRR*, abs/1706.05587, 2017.

[5]L. Chen, Y. Zhu, G. Papandreou, F. Schroff et H. Adam. Encoder-decoder with atrous separable convolution for se- mantic image segmentation. *CoRR*, abs/1802.02611, 2018.

[6]A. Dengel et B. Klein. smartfix : A requirements-driven system for document analysis and understanding. volume 2423, pages 433-444, 08 2002.

[7]J. Devlin, M. Chang, K. Lee et K. Toutanova. BERT : pré-entraînement de transformateurs bidirectionnels profonds pour l'un- derstanding du langage. *CoRR*, abs/1810.04805, 2018.

[8]D. Esser, D. Schuster, K. Muthmann et A. Schill. Auto- matic indexing of scanned documents - a layout-based ap- proach. volume 8297, 01 2012.

[9]R. B. Palm, O. Winther, et F.Laws. Cloudscan - A configuration-free invoice analysis system using recurrent neural networks. *CoRR*, abs/1708.07403, 2017.

[10]D. Schuster, K. Muthmann, D. Esser, A. Schill, M. Berger, C. Weidling, K. Aliyev et A. Hofmeier. Intellix - end-user trained information extraction for document archiving. 08 2013.

[11]K. Sun, B. Xiao, D. Liu et J. Wang. Deep high-resolution representation learning for human pose estimation, 2019.

[12]A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser et I. Polosukhin. Attention is all you need. *CoRR*, abs/1706.03762, 2017.

6. ANNEXE

6.1. Structure du modèle

Le modèle CUTIE-B est présenté 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 une foulée de 1 et 4 convolutions atrous consécutives sont effectuées dans le bloc conv avec une foulée de 1 et un taux de 1.

2. Après le bloc de convolution atrous, un module ASPP est utilisé pour fusionner les caractéristiques multirésolution. La caractéristique de bas niveau mais de haute résolution de la première sortie du bloc de convolution est également ajoutée au modèle dans le raccourci.

couche avec une opération de concaténation et une convolu- tion 1 1 . Enfin, la sortie de l'inférence est obtenue par une convolution 1 1 .

Tableau 7. Structure du modèle CUTIE-B proposé.

nom de la couche	opérations	dimension de l'entrée	dimension de la sortie	commentaires
couche d'intégration	-	20000	128	
bloc de conv	[3 × 5]×4	256	256	stride=1
bloc de conv	[3 × 5]×4	256	256	stride=1,
atrous module	[3 × 5]×*3*, regroupement global, concat, 1 × 1	256	256	rate=2
ASPP couche de raccourcissement	concat, 1 × 1	256	64	stride=1,
couche de sortie	1 × 1	64	9	*rate={4*,8,*16}*