Fully Convolutional Networks for physical and logical structure analysis of historical newspapers

Yann Soullard, Pierrick Tranouez, Stéphane Nicolas,
Clément Chatelain, Thierry Paquet

University of Rouen Normandy, INSA Rouen Normandy

June 6, 2019

Introduction

Historical newspapers analysis:

- Field identification in a page
  - ➥ title level 1, 2,..., text, advertisement, image, caption, ...
- Structure analysis
  - ➥ articles, sections, reading order.
- Information extraction
  - ➥ article classification, name entities retrieval,...

Introduction

Field identification in a page, it's difficult!

- Class imbalance
- Not always rectangular blocks
- Various scales
- Various fonts and printed styles
- Confusing classes

Introduction

Field identification in a page, it's difficult!

- Class imbalance
- Not always rectangular blocks
- Various scales
- Various fonts and printed styles
- Confusing classes

LE TEMPS DU 20 AOUT 1937

54e ANNÉE — N° 19.509

LA GUERRE D'ESPAGNE
et ses répercussions internationales

LE PORTUGAL ROMPT LES RELATIONS DIPLOMATIQUES AVEC LA TCHÉCOSLOVAQUIE CELLE-CI REFUSANT DE LIVRER UNE COMMANDE D'ARMES

Cette décision laisse-t-on entendre dans certains milieux serait motivée par les pressions politiques exercées en coulisse par Moscou

Encore un navire gouvernemental torpillé par un sous-marin au large de Ténédos face aux Dardanelles

AUDACIEUX ACTE de banditisme en Belgique

DÉGUISÉS EN DOUANIERS DEUX MALFAITEURS ARRÊTENT UN AUTOMOBILISTE LE LIGOTENT LUI DÉROBENT 200.000 FRANCS ET L'ABANDONNENT SUR LA ROUTE

Introduction

Field identification in a page, it's difficult!

- Class imbalance
- Not always rectangular blocks
- Various scales
- Various fonts and printed styles
- Confusing classes

Vendredi 20 Août 1937

Les vomissements sont fréquents chez l'hépatique

Que ce soit le matin au réveil, que ce soit après l'ingestion de mets qu'il tolère mal, l'hépatique est sujet aux migraines et aux nausées, suivies parfois de vomissements plus ou moins amers ou acides. Ces inconvénients disparaissent très vite, ainsi que tous les malaises d'origine hépatique (vertiges, démangeaisons, douleurs au côté droit, constipation ou diarrhée) lorsqu'on prend de l'Hépascol François. Lisez ce qu'écrit à ce sujet Mme B..., sage-femme à Toulouse :

« Une jeune femme de mes clientes vient d'être merveilleusement soulagée par votre Hépascol François. Elle souffrait d'une violente douleur au côté droit suivie de vomissements. Après deux flacons, elle va tout à fait bien. Son teint est plus clair et son caractère plus gai ». (Lettre F. 459).

L'Hépascol François est le remède type de toutes les affections du foie. Il est facile à prendre (20 à 30 gouttes aux deux principaux repas, dans le premier verre de boisson). Son action se fait sentir dès la première semaine. Toutes pharmacies : 10 fr. 60 le flacon; 16 fr. 95 le double flacon.

La présence d'une tribu de nomades provoque la démission d'une municipalité

State of the art

Newspaper analysis

- HOG features in input of a random forest classifier, rectangular bounding boxes [Lang et al., 2018]
- Contour classification and morphological operations [Vasilopoulos and Kavallieratou, 2017]

Related domains

- Localization of handwritten fields [Lemaitre et al., 2018]
  - ➥ rule-based system
- dhSegment [Oliveira et al., 2018]
  - ➥ Fully Convolutional Network (from a ResNet-50)
- Tools and software (PIVAJ [Tranouez et al., 2015], DMOS [Couasnon and Lemaitre, 2017])

State of the art - PIVAJ

History of PIVAJ

- An offline article analyser and an online viewer and transcription editor
- From a research project to a licensed software suite
- Currently used by National Libraries of Finland and Wales

State of the art - PIVAJ

History of PIVAJ

- An offline article analyser and an online viewer and transcription editor
- From a research project to a licensed software suite
- Currently used by National Libraries of Finland and Wales

MARDI 3 JANVIER 1815.

JOURNAL DES DÉBATS
POLITIQUES ET LITTÉRAIRES.

NOUVELLES ETRANGERES.

ESPAGNE.

Madrid, 13 décembre.

Le cordon de troupes destiné à empêcher la communication de Cadix avec l'intérieur, et à prévenir ainsi toute contagion, a été levé le 6 décembre.

Le gouvernement espagnol a appris aujourd'hui la défaite des révoltés du Mexique et la soumission volontaire des habitans de Buenos-Ayres. L'expédition qui devoit partir pour cette dernière colonie, a, dit-on, changé de destination, et ira au Mexique; dont Venegas seroit de nouveau vice-roi.

ITALIE.

Naples, 14 décembre.

La générosité avec laquelle le roi Joachim a permis, par son décret du 3 de ce mois, à tous ceux qui souscriroient quelques adresses de dévouement, de se décorer de la médaille d'honneur avec un ruban amaranthe, fait qu'il n'y a pas jusqu'à des écoliers des collèges et des clercs de séminaires, qui, pour avoir cette décoration, ne se conforment aux vues de cet industrieux monarque. Au reste, nous savons que ses agens les y forcent même avec des menaces, sans épargner les simples maisons d'éducation dans les trois légations occupées par nos troupes. Comme il promet à ses soldats la conquête totale de l'Italie, et que par ce moyen il exalte au dernier point leur courage, les provinces craignent; et dans leur inquiétude, il en est qui envoient aussi des adresses. De ce nombre est même une colonie albanaise qui occupe une des îles de l'Adriatique.

Milan, 19 décembre.

Le public est impatient de connoître le résultat des arrestations faites récemment à Milan et en d'autres villes d'Italie.

La régence vient de créer une commission extraordinaire à laquelle pourront s'adresser tous ceux qui ont des crédits sur le gouvernement français, pour fournitures faites, ou autres motifs énoncés dans l'article 19 du traité de paix conclu à Paris le 30 mai dernier; et la régence leur fait connoître que S. M. l'Empereur d'Autriche a établi, M. le comte de Bombelle, son commissaire impérial à Paris pour faire payer ces crédits.

Les officiers de la ci-devant armée du royaume d'Italie, qui ont été licenciés parce qu'ils appartenoient à l'une ou à l'autre des trois légations de Bologne, Ferrare et la Romagne, ont envoyé en commun, à S. Exc. le maréchal Bellegarde, une adresse de remercîment pour la pension de retraite qu'il leur a fait obtenir de S. M. l'Empereur d'Autriche.

La nuit dernière est arrivé en poste de Naples ici, un Français, général au service du roi Joachim (M. de Vaugiyon); il est venu, dit-on, pour démentir les bruits pourtant assez accrédités, que nos mécontens et nos conspirateurs avoient de secrètes intelligences avec l'armée napolitaine d'Ancône.

GRANDE-BRETAGNE.

Londres, 29 décembre.

Fonds publics. — Omnium, 1 1/4 disc. consol. pour janv. 67 1/4 1/8.

Il y a des gens qui semblent croire que M. Madison refusera de ratifier le traité signé par ses commissaires. Ils appellent à ce sujet la conduite de M. Jefferson, qui refusa sa ratification au traité signé le 31 décembre 1806 entre lord Holland et lord Auckland d'une part, et MM. Pinckney et Monroë de l'autre; et ces mêmes personnes se plaignent en même temps de ce que le dernier traité n'est pas honorable pour l'Angleterre. S'il en est ainsi, on peut être certain qu'il sera ratifié. Quoi qu'il en soit, les préparatifs de guerre ne se ralentissent pas. (*The Courrier.*)

FRANCE.

PARIS, 2 *janvier*.

Une note que l'on nous a remise, appuyée de détails qui en garantissent l'authenticité, mais que nous supprimons, d'après l'invitation qui nous en a été faite, pour ne pas humilier sans fruit les personnes qui seroient dans le cas de s'y reconnoître, nous autorise à désavouer toutes les quêtes qui se sont faites, se font ou pourroient se faire encore au nom de la sœur Marthe; elle n'a jamais chargé personne de solliciter les dons de la charité pour les malheureux auxquels elle se dévoue. Après avoir nourri, dans un temps, jusqu'à 8000 prisonniers par jour, sans le secours des intendans qui, depuis qu'elle est à Paris, se sont placés entr'elle et les bienfaiteurs des pauvres, elle est, aujourd'hui moins que jamais, dans la nécessité d'y avoir recours.

Ce qui pourroit, par la suite, lui manquer pour continuer ses œuvres de charité, elle le demandera directement à la famille royale, près de laquelle on est toujours sûr d'obtenir un accès facile, au nom de la religion et de l'humanité.

On a de fortes raisons de croire que des dons adressés par des souverains étrangers ont été interceptés. La crainte du scandale et le peu d'espoir que l'on a de recouvrer ces dons, auroient engagé à taire cette circonstance, s'il n'étoit nécessaire de prémunir contre de pareilles infidélités, les âmes généreuses qui, surtout en Suisse et en Angleterre, se sont associées au dévouement de la sœur Marthe.

C'est une consolation, en signalant d'aussi déplorables abus, d'avoir à citer parmi les noms que la reconnoissance des malheureux proclame, un de ceux que la gloire a le plus illustrés, celui de lord Wellington, qui a fait remettre à la sœur Marthe une somme pour l'aider à racheter un fonds qu'elle avoit aliéné à l'époque des plus nombreux

FEUILLETON DU JOURNAL DES DÉBATS.

Mardi 3 Janvier 1815.

ACADÉMIE ROYALE DE MUSIQUE.

Castor et Pollux, opéra en cinq actes, musique de M. Candeille.

THÉATRE FRANÇAIS.

L'Intrigue épistolaire, ou le Médecin malgré lui.

Demain, *Phèdre*, Mlle Duchesnois reparoîtra par le rôle de Phèdre.

THÉATRE DE L'OPÉRA-COMIQUE.

Le Dot. Paul et Virginie.

VARIÉTÉS.

Charlemagne, ou l'Eglise délivrée, poëme épique en vingt-quatre chants; par Lucien Buonaparte, membre de l'Institut de France, etc. etc.

(IIe et dernier article.)

Avant de faire justice des chants et des vers du poëte, je m'empresse de rendre justice aux principes et aux sentimens de l'auteur du poëme. Ces sentimens sont nobles et élevés, ces principes toujours conformes à la morale, à la justice, à l'honneur; ils sont religieux, édifians même, et l'on est souvent tenté, en lisant les vers de *Charlemagne*, de dire qu'ils sont bien plus d'un bon chrétien que d'un bon poëte. Toutes les vertus douces, sociales et généreuses y sont célébrées; l'amour de l'humanité y respire; les qualités odieuses, ou même les vices brillans qui asservissent les peuples, et font le malheur des hommes, y sont détestés et flétris; l'ambition surtout, l'insatiable soif des conquêtes, l'abus de la force, la tyrannie, la mauvaise foi dans les transactions, la perfidie dans la violation des traités y sont voués au mépris et à l'exécration des hommes, sinon avec le talent d'un poëte, du moins avec l'expression et la chaleur d'une âme indignée; quelquefois même, mais trop rarement, la vérité, l'ange ou la noblesse de la pensée, soutenant assez bien la strophe, lui donnent quelqu'éclat, et de même que, selon le législateur du Parnasse, *le vers se sent toujours des bassesses du cœur*, il se sent quelquefois, dans le poëme de M. Lucien Buonaparte, de l'élévation des sentimens. Il est certain que, si ce poëme eût été passablement conçu, et passablement exécuté; si enfin, réduit à de plus justes termes, et à des dimensions moins effrayantes, il eût été lisible, ces principes humains et généreux, ces vives et honorables déclamations donneroient à ce poëme un attrait assez piquant, celui de la satire dirigée contre un homme odieux, et de la satire faite par un frère. Disons de plus, à l'honneur de ce frère, que ce n'étoit pas seulement dans son poëme et dans ses vers, mais plus encore par sa conduite et par une résistance ferme, courageuse et constante qu'il protestoit contre la politique cruelle et les volontés oppressives du tyran de la France, et cette conduite honorable attestoit que l'auteur ne donne point ici de vaines paroles et ne prend pas un vain masque, quand même un accent de vérité et de conviction intérieure ne déposeroit pas en faveur de la franchise de ses discours.

J'ai depuis long-temps remarqué qu'on flattoit très-peu les poëtes en vantant leurs sentimens: ce sont leurs vers qu'il faudroit vanter. Je me suis même brouillé avec plusieurs, pour avoir dit ou insinué qu'ils étoient de fort honnêtes gens. Mais c'est ici un cas tout particulier, et je suis persuadé que la justice que je rends aux bons sentimens et aux excellens principes qui sont exprimés dans le poëme de M. Lucien Buonaparte, et qui sont sans doute dans son âme, ne lui déplaira pas. Elle attestera de plus mon impartialité, et lui prouvera que je n'aurois pas moins loué ses vers que ses principes, si sa poésie m'eût paru valoir sa morale. Mais je dois dire, au contraire, que ses vers sont tellement mauvais, que s'ils ne doivent pas faire oublier ses bons

State of the art - PIVAJ

PIVAJ offline

1. Pixel level classification (Text, Image, Separators)
2. Connected component, bounding boxes
3. Bottom up structural grid and reading order
4. Text level classification with Random Forests
5. Section and articles assembly through regular expressions

The goal of this current work is twofold:

- Investigate deep learning methods for physical and logical structure analysis,
- Gain in generalization capabilities (being less data-dependent).

FCN

Fully convolutional networks (FCN) were first introduced in semantic segmentation [Long et al. 2014].

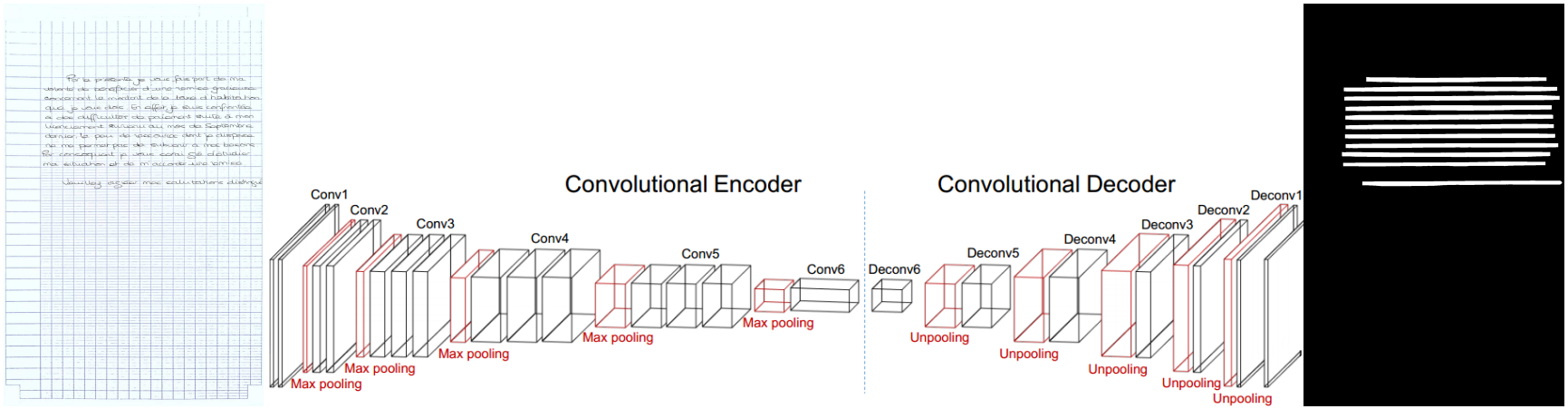
FCN are convolutional networks without a dense layer.

Removing dense layers has many advantages:

- the number of parameters is highly reduced,
- they can deal with variable input size,
- spatial information is kept inside the network: we can produce a pixel labeling.

FCN

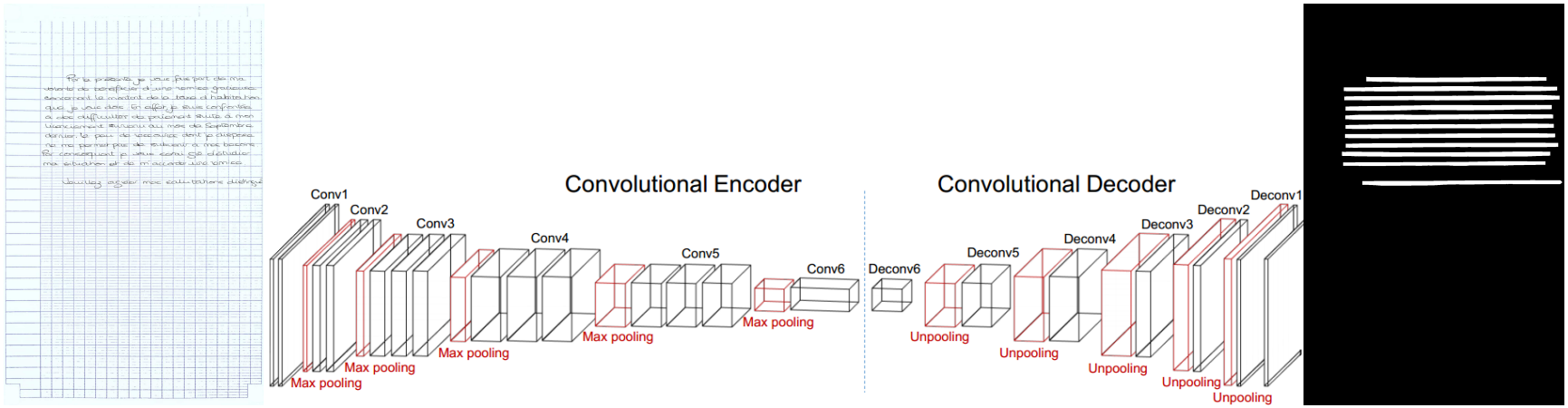
Pixel labeling ➠ How to get an output with the same dimension as the input image?



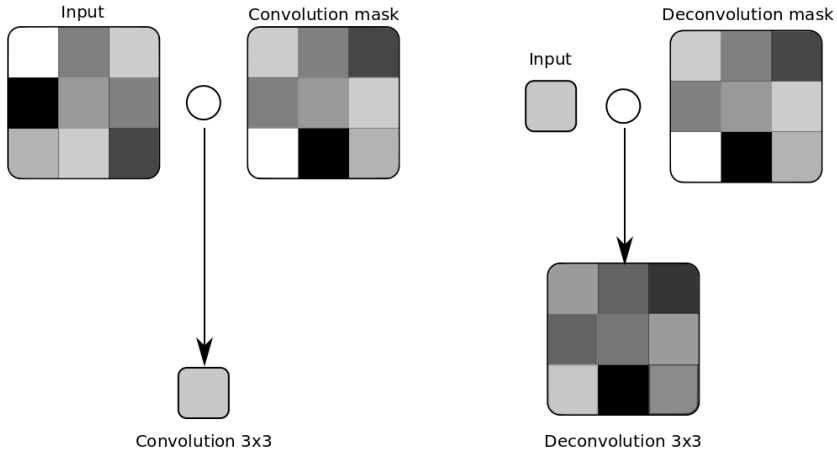
- Transposed convolutions,
- Unpooling layers,
- Dilated convolutions.

FCN

Pixel labeling ➠ How to get an output with the same dimension as the input image?

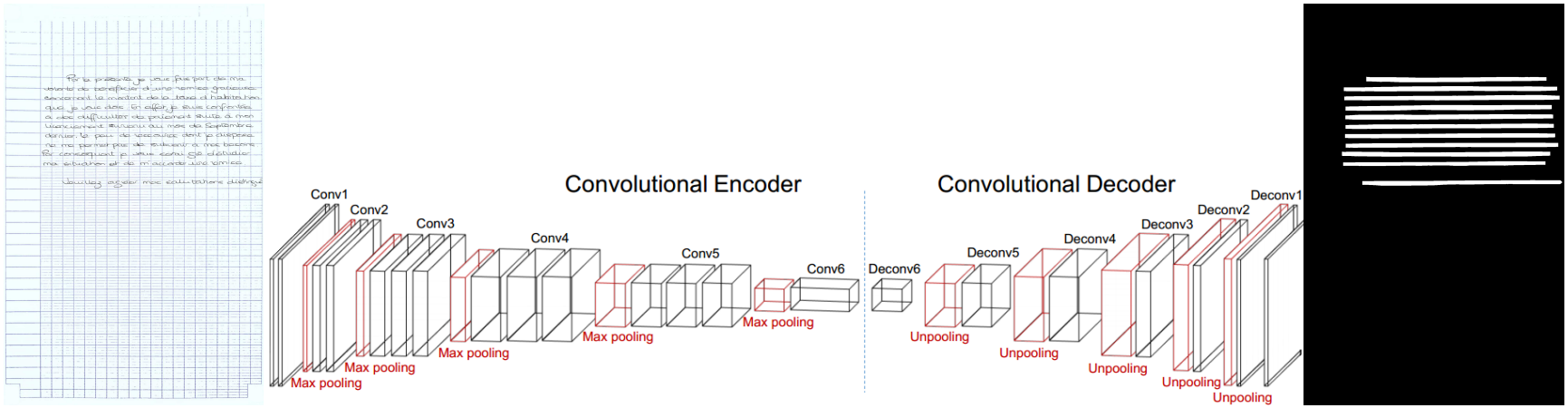


- Transposed convolutions,

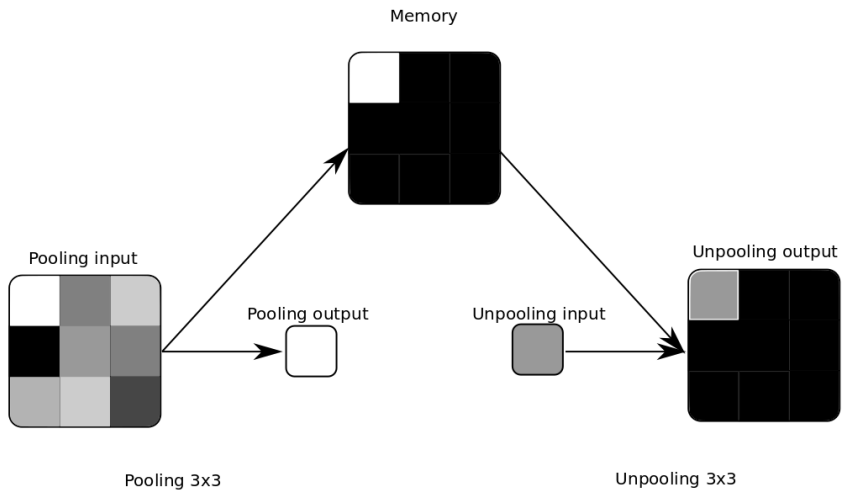


FCN

Pixel labeling ➠ How to get an output with the same dimension as the input image?

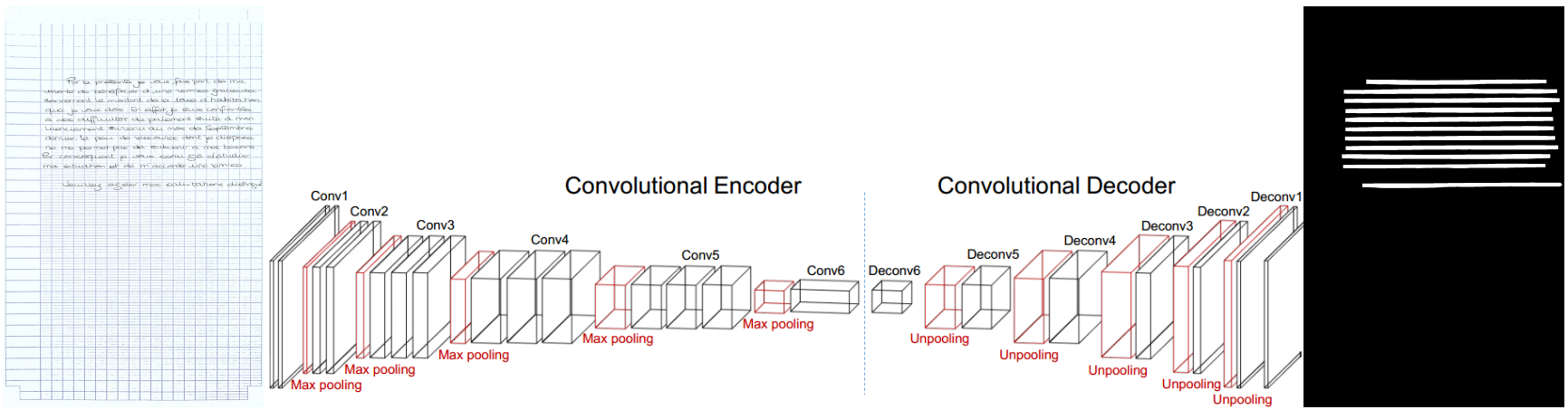


- Unpooling layers

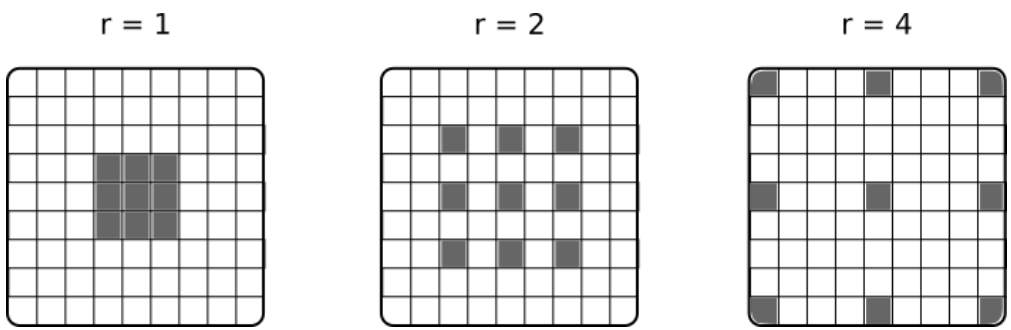


FCN

Pixel labeling ➠ How to get an output with the same dimension as the input image?



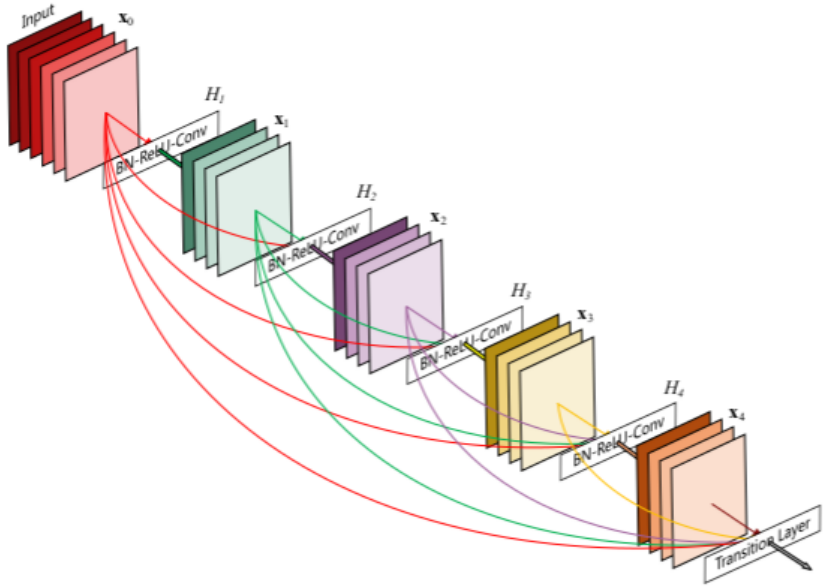
- Dilated convolutions.



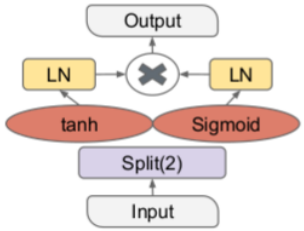
Improvements of convolutional networks

Strengthen the capacities of convolutional networks

- Dense Nets
  - ➥ Generalize a residual architecture



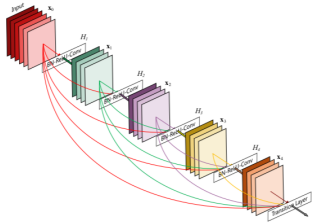
- Gated Nets
  - ➥ Auto-attention mechanism



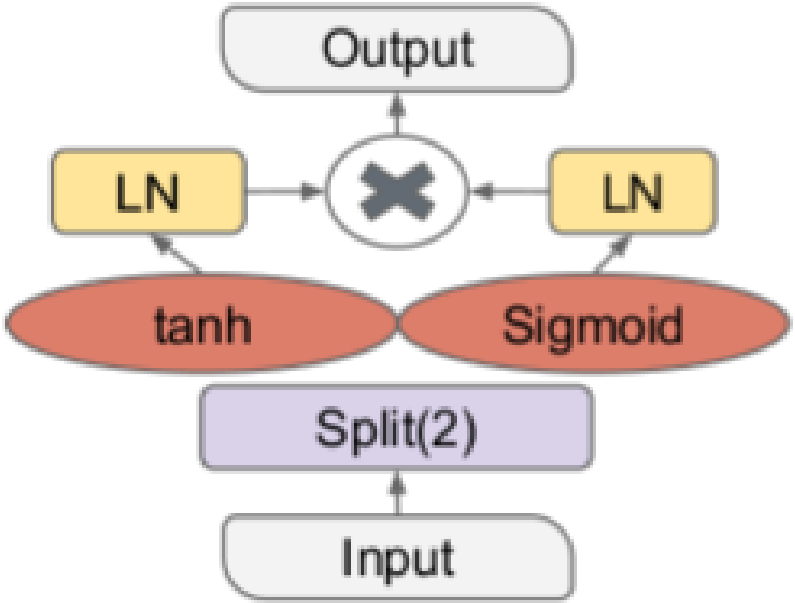
Improvements of convolutional networks

Strengthen the capacities of convolutional networks

- Dense Nets
  - ➥ Generalize a residual architecture



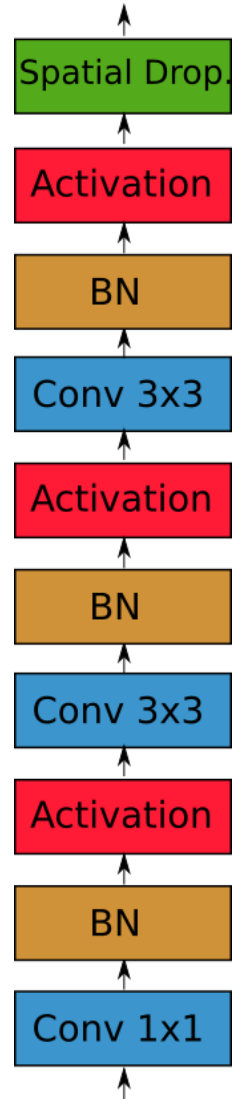
- Gated Nets
  - ➥ Auto-attention mechanism



Proposed architectures

We experiment 3 network architectures:

- **Dilated FCN 22**: FCN with 21 dilated convolutions (7 convolutional blocks with dilation rates 1-2-4-8-4-2-1)
- **Dense Net 22**: the same architecture than the Dilated FCN 22 with a dense architecture between blocs.
- **Gated Net 22**: the same architecture than the Dense Net with a gate at the end of each bloc.



One block

Data sets

Historical newspapers:

- **Le Matin**
  French daily newspaper from 1883 to 1944.
  801 examples in training, 9 classes (6 in the experiments)

- **GBNLA**
  ICDAR competition on historical newspapers.
  66 examples in training from 6 newspapers, 4 classes

- **Luxembourg Historical Newspapers**
  980 examples in training from 1877, 11 classes

Input image size are reduced to at least 1200 x 750 by keeping the ratio.

Results

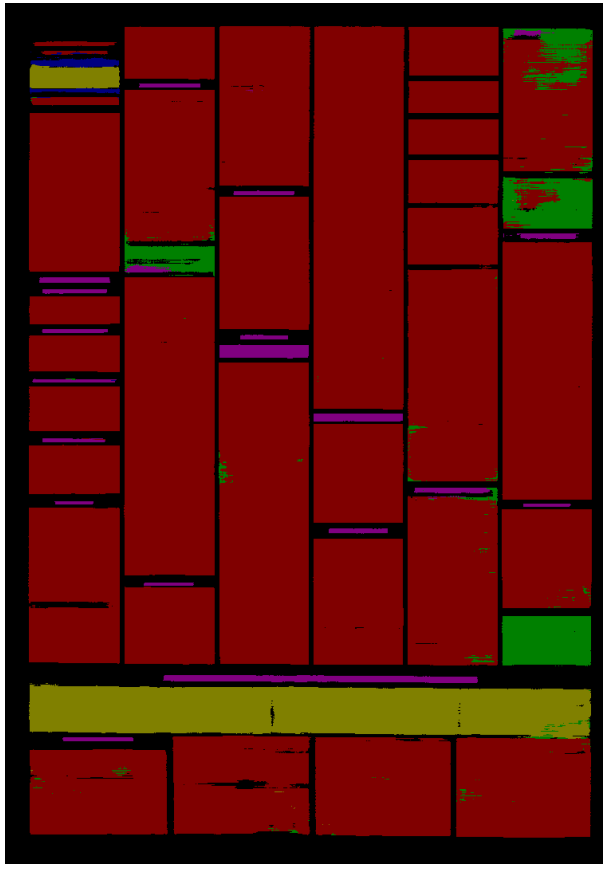
Accuracy (%) on each newspaper, training from scratch.

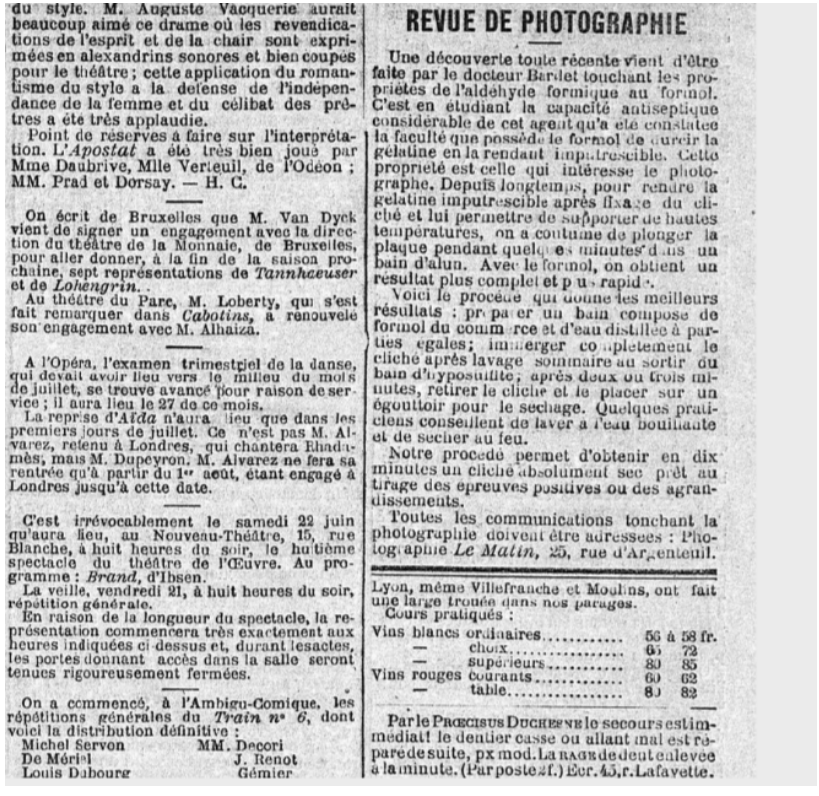
	Le Matin	GBNLA	Luxembourg
Dilated FCN 22	85.73	87.17	81.55
Dense Net 22	86.31	89.06	81.66
Gated Net 22	85.19	86.57	82.23

- Dense architecture improves performance,
- Having half the feature maps for the gates can be beneficial,
- Gated networks are prone to overfitting.

Additional observations (1/3)

Doubt is generally legitimate





du style. M. Auguste Vacquerie aurait beaucoup aimé ce drame où les revendications de l'esprit et de la chair sont exprimées en alexandrins sonores et bien coupés pour le théâtre ; cette application du romantisme du style à la défense de l'indépendance de la femme et du célibat des prêtres a été très applaudie.

Point de réserves à faire sur l'interprétation. L'*Apostat* a été très bien joué par Mme Daubrive, Mlle Verlieuil, de l'Odéon ; MM. Prad et Dorsay. — H. C.

On écrit de Bruxelles que M. Van Dyck vient de signer un engagement avec la direction du théâtre de la Monnaie, de Bruxelles, pour aller donner, à la fin de la saison prochaine, sept représentations de *Tannhæuser* et de *Lohengrin*.

Au théâtre du Parc, M. Loberty, qui s'est fait remarquer dans *Cabotins*, a renouvelé son engagement avec M. Alhaiza.

A l'Opéra, l'examen trimestriel de la danse, qui devait avoir lieu vers le milieu du mois de juillet, se trouve avancé pour raison de service ; il aura lieu le 27 de ce mois.

La reprise d'*Aïda* n'aura lieu que dans les premiers jours de juillet. Ce n'est pas M. Alvarez, retenu à Londres, qui chantera Rhadamès, mais M. Dupeyron. M. Alvarez ne fera sa rentrée qu'à partir du 1er août, étant engagé à Londres jusqu'à cette date.

C'est irrévocablement le samedi 22 juin qu'aura lieu, au Nouveau-Théâtre, 15, rue Blanche, à huit heures du soir, le huitième spectacle du théâtre de l'Œuvre. Au programme : *Brand*, d'Ibsen.

La veille, vendredi 21, à huit heures du soir, répétition générale.

En raison de la longueur du spectacle, la représentation commencera très exactement aux heures indiquées ci-dessus et, durant les actes, les portes donnant accès dans la salle seront tenues rigoureusement fermées.

On a commencé, à l'Ambigu-Comique, les répétitions générales du *Train n° 6*, dont voici la distribution définitive :

Michel Servon	MM. Decori
De Mériel	J. Renot
Louis Dubourg	Gémier

REVUE DE PHOTOGRAPHIE

Une découverte toute récente vient d'être faite par le docteur Bardet touchant les propriétés de l'aldéhyde formique au formol. C'est en étudiant la capacité antiseptique considérable de cet agent qu'a été constatée la faculté que possède le formol de durcir la gélatine en la rendant imputrescible. Cette propriété est celle qui intéresse le photographe. Depuis longtemps, pour rendre la gélatine imputrescible après fixage du cliché et lui permettre de supporter de hautes températures, on a coutume de plonger la plaque pendant quelques minutes dans un bain d'alun. Avec le formol, on obtient un résultat plus complet et plus rapide.

Voici le procédé qui donne les meilleurs résultats : préparer un bain composé de formol du commerce et d'eau distillée à parties égales; immerger complètement le cliché après lavage sommaire au sortir du bain d'hyposulfite; après deux ou trois minutes, retirer le cliché et le placer sur un égouttoir pour le séchage. Quelques praticiens conseillent de laver à l'eau bouillante et de sécher au feu.

Notre procédé permet d'obtenir en dix minutes un cliché absolument sec prêt au tirage des épreuves positives ou des agrandissements.

Toutes les communications touchant la photographie doivent être adressées : Photographie *Le Matin*, 25, rue d'Argenteuil.

Lyon, même Villefranche et Moulins, ont fait une large trouée dans nos parages.

Cours pratiqués :

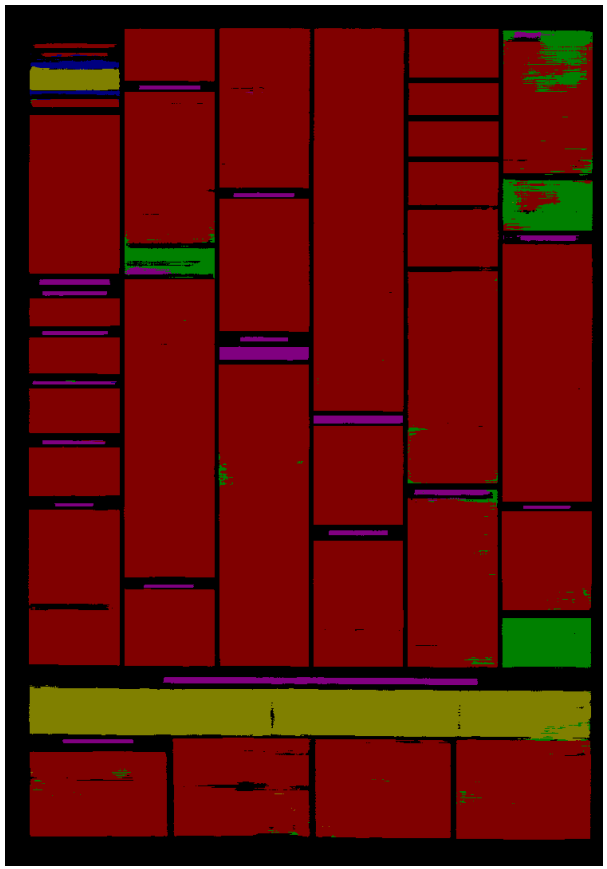
Vins blancs ordinaires	56 à 58 fr.	
— choix	65	72
— supérieurs	80	85
Vins rouges courants	60	62
— table	80	82

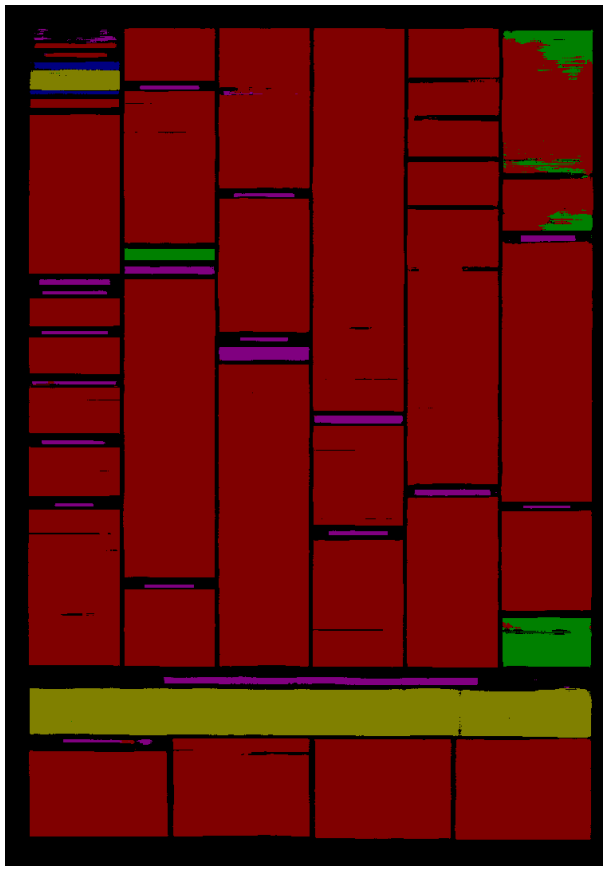
Le Matin newspaper.

- This is observed in case of confusing observations,
- There is sometimes labelling errors and clumsy labelling.

Additional observations (2/3)

Fine-tuning on a specific time period is important



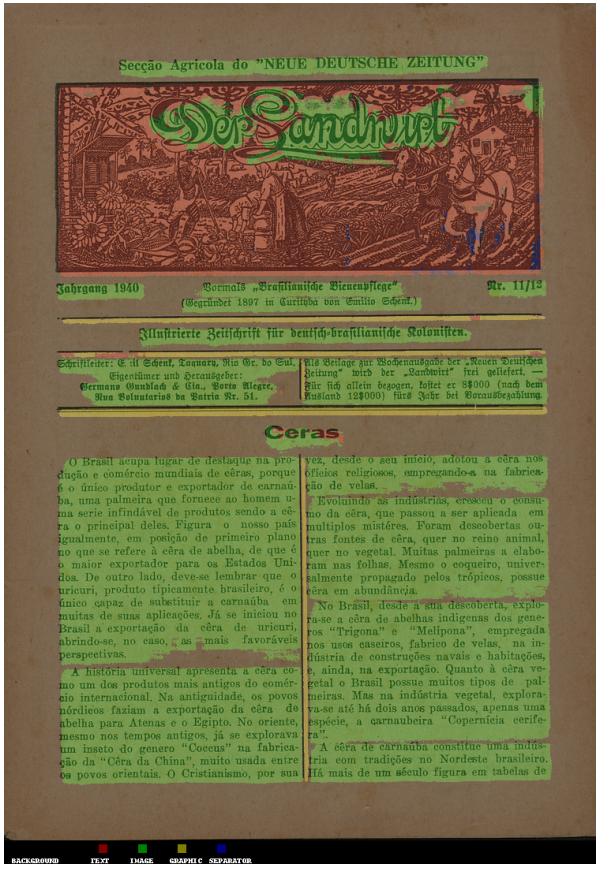


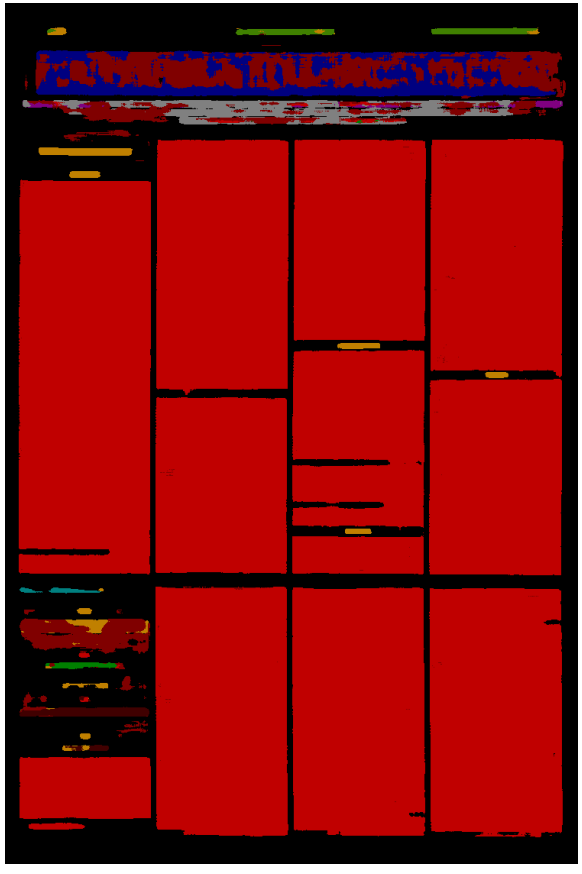
Le Matin newspaper.

- Predicted blocs are more precise and more confident,
- The model tends to overfit (help to label confusing classes).

Additional observations (3/3)

Class imbalance must be taken into account





GBNLA (left) and *Luxembourg* (right) newspapers.

- Text parts are well identified,
- Infrequent classes (e.g. levels of title) are difficult to recognize (even if there is highly similar pages in training).

Conclusion

We investigate **Fully Convolutional Networks for analyzing historical newspapers**

- FCN can manage input images of variable size,
- FCN provide pixel labeling,
- Recent improvements (e.g. Gate, denseNet) strengthen the model capacities with auto-attention and multiscale analysis.

Regarding **future works**

- Increase the receptive fields ➠ more context in prediction,
- Hierarchical analysis ➠ class imbalance,
- Joint multi-scale analysis (dense-like architecture) with attention mechanism ➠ recognize blocks of various scales,
- Extract polygon boxes ➠ further analysis (removing regions without interest, reading order...).

Thanks for your attention!