Classification multi-label à grande dimension pour la détection de concepts médicaux

Mamitiana Ignace Randrianarivony [1] Nomena Ny Hoavy [1] Josiane Mothe [2]
(1) MISA, Université d'Antananarivo, Madagascar
(2) IRIT, UMR5505, CNRS & Univ. Toulouse, France
mamitianaignace@gmail.com, nomena.ny-hoavy@irit.fr,
josiane.mothe@irit.fr

RÉSUMÉ

Dans ce papier, nous présentons une méthode pour associer de façon automatique des concepts à des images. Nous nous focalisons plus particulièrement sur des images médicales à annoter avec des concepts UMLS. Nous avons développé deux modèles de transfert d'apprentissage à partir des réseaux CNN VGG19 et ResNet50 . Nous avons utilisé des modèles avec des techniques simples et que nous avons optimisés pour l'apprentissage. Les résultats que nous avons obtenus en utilisant les données de la tâche ImageCLEF 2017 sont encourageants et comparables à ceux des autres participants.

ABSTRACT

Large multi-label classification for medical concepts detection

In this paper, we present a method to automatically associate concepts with images. We focus on medical images to annotate with UMLS concepts. Our work is based on transfer learning in a deep neuron network. We used two different models (VGG19 and ResNet50) with simple techniques that we optimized for training. The results we obtained using the ImageCLEF 2017 task data collection are encouraging and comparable to other participants.

MOTS-CLÉS : Recherche d'information, annotation d'images, indexation d'images, réseaux de neurones profonds, CNN, transfert d'apprentissage, ImageCLEF.

KEYWORDS: Information retrieval, Image annotation, Image indexation, CNN, transfer learning, ImageCLEF.

1 Introduction

Ce papier s'inscrit dans le domaine de recherche d'information d'images médicales. Nous nous sommes intéressés particulièrement à l'annotation automatique d'images médicales par des concepts médicaux. Les concepts médicaux utilisés (meta thesaurus UMLS [1]) sont des termes issus du domaine de la biomédecine. Ils ont été proposés par la "National Library of Medicine" dans le but de développer des systèmes informatiques capables de s'appuyer sur le vocabulaire spécialisé utilisé dans le domaine (Bodenreider, 2004).

Dans ce travail, notre objectif est d'extraire les concepts qui sont en relation avec le contenu d'une

1. https://www.nlm.nih.gov/research/umls/

image. Les évaluations s'appuient sur le lab proposé dans le forum d'évaluation ImageCLEF 2017 [2] sous le nom "ImageCLEF captions". Cette tâche a comme perspective de concevoir des systèmes de génération automatique de descriptions pour les images médicales (Ionescu *et al.*, 2017; Eickhoff *et al.*, 2017).

Pour répondre à cette tâche, nous avons assimilé le problème d'association de concepts à un problème de classification multi-label (classification à plusieurs sorties) ; les concepts étant les classes auxquelles l'image est associée.

Pour la classification multi-label, notre choix s'est porté sur l'apprentissage profond avec un réseau de neurones profond (Convolutional Neural Network - CNN) car depuis quelques années ces réseaux ont montré de bonnes performances et sont devenus l'état de l'art dans le domaine de la vision par ordinateur, surtout pour la classification d'images (Simonyan & Zisserman, 2014; Sharif Razavian *et al.*, 2014; Russakovsky *et al.*, 2015). Les CNN sont composés d'une succession de couches de produit de convolution des images. Ces produits de convolution servent à extraire petit à petit les informations caractéristiques sur l'image traitée.

Plus spécifiquement, dans notre travail nous avons utilisé un transfert d'apprentissage. Le transfert d'apprentissage est utilisé en apprentissage profond et a pour principe de migrer les connaissances d'un modèle déjà appris (les paramètres) vers un nouveau problème. L'architecture et les paramètres du CNN sont ainsi *transférés* (Yosinski *et al.*, 2014).

Dans le premier modèle développé, nous avons utilisé le principe d'encodeur-décodeur qui est une stratégie utilisée pour générer automatiquement les légendes des images (Vinyals *et al.*, 2015). Dans notre cas, cette stratégie a été appliqué pour la classification multi-label des images. Plus spécifiquement, à partir d'un CNN déjà entraîné, nous avons extrait un vecteur de caractéristiques (encodage). Après cet encodage, le vecteur de caractéristiques est décodé par un autre réseau de neurones afin d'obtenir les concepts.

Le deuxième modèle que nous avons utilisé fait une réglage fin d'un CNN existant. Dans ce cas, nous avons repris un CNN entraîné et nous l'avons adapté à notre tâche en faisant un ré-apprentissage de ses paramètres à partir de la collection visée.

L'évaluation des modèles que nous proposons ici a été réalisée sur les données de la tâche ImageCLEF 2017 (Ionescu *et al.*, 2017) et prolonge les travaux présentés dans (Mothe *et al.*, 2017).

Le reste de l'article est organisé comme suit : dans la section 2, nous présentons une description détaillée du problème ainsi que les données qui servent de base à l'évaluation. Nous présentons ensuite dans la section 3 les modèles nous avons développés. Les résultats que nous avons obtenus sont présentés dans la section 4 tandis que la section 5 décrit les travaux similaires présentés dans la campagne ImageCLEF.

2 Détail de la tâche

Le but de la détection automatique de concepts médicaux est d'identifier les concepts pertinents à associer à une image. Cette tâche consiste dans un premier temps à analyser le contenu de l'image représentée par une matrice de pixels et ensuite de prédire les concepts les plus pertinents contenus dans cette image. Notre travail porte sur l'analyse d'un des modèles pré-entrainés sur des images

2. http://www.imageclef.org/2017

naturelles pour les adapter aux images médicales de la collection visée. A partir des données d'apprentissage nous avons construit deux modèle pour classer les images dans des classes prédéfinies (concepts). Le premier modèle utilise un CNN pré-entrainé pour extraire le vecteur caractéristique de chaque image. Un réseau interconnecté est ensuite entraîné pour classer chaque vecteur dans les concepts prédéfinis. Le second modèle est obtenu par réglage fin d'un CNN pré-entraîné en effectuant une réapprentissage du modèle sur les images médicales de la collection.

2.1 Étapes du processus

Dans l'acquisition des données, nous avons d'abord redimensionné tous les images en taille 256*256 pour les uniformiser.

Puis, ces images entrent dans un CNN encodeur.Parceque la succession de différentes couches de convolution et de mise en commun dans le CNN permet d'extraire les caractéristiques pertinentes des images. Ce traitement nous a permis de réduire le bruit dans l'image traitée et de conserver les informations pertinentes pour la classification. Un vecteur nommé vecteur caractéristique est obtenu pour représenter l'image. Une fois les caractéristiques obtenues, les autres parties de nos modèles ont été construites de sorte qu'elles prédisent les concepts pertinents présents dans l'image.

Comme dans notre cas les images doivent être associées à un ou plusieurs concepts, nous avons adapté les modèles pour qu'ils puissent fournir en sortie, non pas une classe, mais plusieurs. Le nombre de sorties de la dernière couche des deux modèles correspond donc au nombre de concepts candidats.

Pour l'apprentissage de la 2ème partie du premier modèle et le réglage fin du second, nous avons effectué une augmentation de données en appliquant des déformations géométriques aux images : symétrie horizontale et verticale. Ceci permet aux modèles de mieux distinguer les caractéristiques invariantes.

Les concepts ont été représentés par un vecteur sac de mots à valeur binaire selon l'index du concept.

Ce vecteur est utilisé dans la dernière couche, le but est de prédire ce vecteur (car prédire les concepts est équivalent à prédire ce vecteur).

2.2 Collection

La collection de données que nous avons utilisée a été proposée par le Lab ImageCLEF en collaboration avec U.S. National Library of Medicine (NLM) (Ionescu *et al.*, 2017; Eickhoff *et al.*, 2017).

La collection contient au total 184 614 images. A chacune des images sont associés des concepts. La collection est séparée en 3 sous-ensembles :

- les données d'entraînement (training set) contiennent 164614 images avec au total 20463 concepts distincts ;
- les données de validation (validation set) contiennent 10000 images et 7070 concepts dont 309 concepts ne sont pas inclus dans les concepts de l'ensemble d'entraînement ;
- les données de tests (test set) sont composées de 10000 images ; les concepts sont à prédire.

Nous avons analysé les ensembles fournis. Nous avons trouvé que 3,9% des images de l'ensemble d'apprentissage et 3,79% des images d'ensemble de validation ne sont associées à aucun concept.

FIGURE 1 – Exemple d'une image composée. Cette image est composée de 4 sous images.

Nous avons approximativement 10-20 % des images qui sont composées (voir par exemple la figure 1) ou ne sont pas des images médicales. Cela créé du bruit dans l'ensemble de données et rend la tâche plus difficile, sans la rendre forcément plus réaliste. Nous avons toutefois conservé des images qui sont dans la collection de référence pour permettre des comparaisons simples avec les autres études utilisant les mêmes données.

3 Modèle CNN et transfert d'apprentissage

Comme nous l'avons mentionné précédemment, le premier modèle que nous avons développé est un transfert d'apprentissage avec Oxford VGG19 comme modèle de CNN de base (Simonyan & Zisserman, 2014) [3]

Nous avons choisi ce modèle car il a déjà fait ces preuves avec 7,3% d'erreur seulement pour la classification d'images sur 1000 catégories du challenge ImageNet Large Scale Visual Recognition (ILVRC 2014) (Russakovsky *et al.*, 2015)). Nous avons retenu le modèle appris sur le challenge ILVRC.

Le CNN VGG19 contient en total 19 couches, avec une succession de couches de convolution et de mise en commun ; il est facile à implémenter sur **caffe**, et se finit avec une couche de sortie de 1000 classes. Le VGG19 est fait pour classifier des images dans 1000 classes mono-label. Nous avons donc adapté ce réseau au problème de classification multi-label que nous souhaitons ici résoudre (nous avons 20463 labels). La création de ce modèle se fait en 2 étapes.

La première étape , la **phase d'encodage**, est d'extraire les vecteurs de représentation des variables pour chaque image. Cette extraction sera réalisée quelle que soit la provenance de l'image (ensemble d'apprentissage, de validation ou de test). Cette réduction de dimension permet de gagner beaucoup de temps et d'espace sur le disque. Ce vecteur de représentation se trouve à l'avant dernière couche du VGG19 (fully connected FC7 dans la Figure) ; la couche correspondante dans le réseau est composée de 4096 neurones. Ce vecteur résume les caractéristiques de l'image encodée, mais il est difficile d'interpréter ce vecteur comme le soulignent Sharif *et al.* (Sharif Razavian *et al.*, 2014). De plus, d'après Bengio *et al.*, entraîner un nouveau modèle à partir de ces vecteurs offre des grandes potentialités pour des problèmes reconnaissance ou de classification (Bengio *et al.*, 2013; Coates

3. Le modèle VGG19 appris sur ILSRVC-2014 est disponible à `http://www.robots.ox.ac.uk/\%7Evgg/research/very_deep/`.

et al., 2010).

La deuxième étape consiste à utiliser ces vecteurs précédemment extraits dans un nouveau réseau de neurones (**phase de décodage**) où nous traitons le problème de la classification multi-label proprement dit. L'architecture de ce second réseau est la suivante :

— Couche d'entrée : c'est cette couche qui reçoit les vecteurs de représentation de taille 4096, sortie du précédent réseau (étape de codage) ;
— Couche cachée : cette couche de taille 20463 que nous avons nommée **multi-label layer** permet de traiter l'aspect multi-label du problème ;
— Couche d'erreur (sigmoidCrossEntropy loss) : il s'agit d'une couche visant à optimiser le modèle. Avec la fonction d'activation sigmoïde sur la couche de sortie, le réseau de neurones modélise la probabilité d'un label l_j (voir équation 1).

Les résultats n'étaient optimums. Nous pensons que soit les vecteurs caractéristiques 4 096 n'apportent pas assez d'information pour le décodage, soit que le réseau VGG19 n'est pas adapté pour des images médicales. Nous avons développé un autre modèle à partir de Résidual Network50 (ResNet50) qui est aussi un CNN issu la collection ImageNet [4] avec seulement 3. 6% d'erreur (He *et al.*, 2016). De plus ResNet50 est aussi utilisé en imagerie médicale (Islam *et al.*, 2017).

Ce second modèle est très simple et ne nécessite qu'une seule étape. Dans la définition de l'architecture du ResNet50, nous avons modifié la définition de la couche FC-1000 en **multi-label layer** comme précédemment et la couche d'erreur softmax loss en sigmoid entropy loss.

Nous avons initialisé ce nouveau réseau de neurones convolutif grâce aux poids de l'ancien réseau ResNet50 (c'est là qu'intervient le transfert d'apprentissage). Nous avons alors réalisé un ré-apprentissage du modèle complet pour mettre à jour les poids initiaux afin qu'ils soient plus compatibles avec la collection d'images médicales visée.

Les deux modèles que nous avons développés ont été entraîné pour optimiser la fonction d'erreur sigmoide cross entropy (cf équation 1) ; en effet, la probabilité de chaque label ou concept étant considéré comme indépendant, cette fonction est adaptée (Tsoumakas *et al.*, 2009).

$$E = -\sum_{n=1}^{N} p(c_i)\log(\hat{p}(c_i)) + q(c_i)\log(\hat{q}(c_i)) \tag{1}$$

$p(c_i)$: 1 si le labels appartient à l'image 0 sinon. $q(c_i)) = 1 - p(c_i)$
$\hat{p}(c_i)$:c_i ième élément de $[1/(1+\exp(-s))] \in [0,1]$ avec s la valeur interne du neurone.

Nous avons réalisé l'apprentissage du modèle avec VGG19 avec les hyper-paramètres suivants : la taille du traitement par lot est de 256 images, le pas d'apprentissage est fixé à 0,0001, nous avons gardé la valeur par défaut du momentum (0,9).

Le modèle ResNet50 utilise les paramètres suivants : la taille du lot est de 50 images, le pas d'apprentissage est diminué selon l'équation 2 à chaque tranche de 7 époques [5].

Cette réduction du pas d'apprentissage est requis pour un réglage fin (Karpathy, 2016; Yosinski *et al.*, 2014).

$$lr = init\_lr * (0.1^{\frac{curent\_epoch}{decay\_epoch}}) \tag{2}$$

4. `http://www.image-net.org/`
5. Une époque est terminée lorsque l'ensemble des données est traité dans l'algorithme d'apprentissage.

lr : pas d'apprentissage
$inti\_lr$: pas d'apprentissage initial = 0,001
$decay\_epoch$: tranche d'époques pour la réduction du pas d'apprentissage = 7
$curent\_epoch$: numéro de l'époque courante

4 Evaluation et résultats

Pour évaluer la performance des modèles d'extraction de concepts des images, les organisateurs de la tâche ImageCLEF 2017 ont proposé la mesure **F1-score** (Eickhoff *et al.*, 2017) . F1-score représente la moyenne harmonique entre la précision et le rappel. La précision calcule le pourcentage des labels prédits qui sont pertinents, et le rappel calcule le pourcentage des labels pertinents qui sont prédits pour une image donnée.

$$Rappel_i = \frac{\mid \widehat{y_i} \cap y_i \mid}{\mid y_i \mid} \tag{3}$$

$$Precision_i = \frac{\mid \widehat{y_i} \cap y_i \mid}{\mid \widehat{y_i} \mid} \tag{4}$$

$$F1\_score_i = \frac{2 * Rappel_i * Precision_i}{Rappel_i + Precision_i} \tag{5}$$

$$HamLoss = \frac{1}{N} \sum_{i=1}^{N} \frac{xor(\widehat{y_i}, y_i)}{L}. \tag{6}$$

i : une image
N : nombre d'images
L : nombre de labels
y_i labels réels
$\widehat{y_i}$ labels prédits

Nous avons complété l'évaluation avec la distance de **Hamming**.

La distance de Hamming (Hamming loss) mesure le taux de labels qui sont mal classés dans une classification multi-label. C'est à dire les labels (concepts) réels qui ne sont pas prédits (vrais négatifs) et les non-labels réels qui sont pourtant prédits (faux positifs).

Le tableau 1 présente les différentes évaluations sur les différentes mesures pour la détection de concepts. Nous pouvons constater dans ce tableau que notre second modèle est plus performant que notre premier modèle. L'amélioration du score F1 est surtout dû à l'augmentation du rappel.

5 Travaux reliés

Dans le cadre de la tâche ImageCLEF 2017, plusieurs méthodes ont été proposées pour identifier les concepts des images. Ce sont des méthodes à base d'apprentissage profond qui ont été essentiellement

Nom du mesure	modèle	valeur
F1-Score (moyenne)	VGG19	0.047
	ResNet50	0.067
precision (moyenne)	VGG19	0.040
	ResNet50	0.049
rappel (moyenne)	VGG19	0.07
	ResNet50	0.014
Hamming loss	VGG19	0.0002
	ResNet50	0.0002

TABLE 1 – Evaluation sur les données de validation - comparaison des deux modèles

	F1-Score
(Abacha *et al.*, 2017) *	0.1718
(Valavanis & Stathopoulos, 2017) *	0.1436
(Hasan *et al.*, 2017) *	0.1208
(Lyndon *et al.*, 2017) *	0.0958
ResNet50	**0.0663**
VGG19	**0.0462**
(Stefan *et al.*, 2017)*	0.0028

TABLE 2 – Les meilleurs F1-scores des équipes évalué par ImageCLEF sur les données de test , dans la catégorie sans ressource externe. Dans (Eickhoff *et al.*, 2017) , il y a en total 20 soumissions des équipes, mais nous avons gardé dans ce tableau le meilleur score de chaque équipe

proposées pour cette tâche. Pour représenter les informations visuelles des images, l'utilisation des CNN a été fréquente (Dimitris & Ergina, 2017; Lyndon *et al.*, 2017; Abacha *et al.*, 2017; Stefan *et al.*, 2017). Les CNN ont produit des modèles robustes.

D'autres travaux utilisent des méthodes plus conventionnelles d'extraction d'information des images comme SIFT , "bag of color" (Abacha *et al.*, 2017; Valavanis & Stathopoulos, 2017). Ils appliquent des méthodes de recherche de similarité pour identifier les images candidates et extraire les concepts à partir de ces images candidates.

Hassan et al. (Hasan *et al.*, 2017) a développé un modèle qui s'appuie sur la notion d'attention visuelle qui présente une amélioration considérable par rapport au simple transfert d'apprentissage de CNN. Dans leur modèle les auteurs ont combiné un CNN et un RNN [6]. Le RNN sert à générer les concepts un par un à partir du CNN.

Le tableau présente les meilleurs F1-scores des équipes d'ImageCLEF pour la résolution de la détection de concepts sans à avoir eu recours à des ressources et données externes.

Les modèles proposés par des participants à ImageCLEF sont indiqués par une *. Les deux modèles que nous avons proposés dans cet article sont en gras dans le tableau. Nous ne disposons malheureusement pas de toutes les données pour comparer tous les algorithmes selon les différentes mesures comme présentés dans le tableau 1.

6. RNN : réseau de neurones récurent, qui est destiné aux données à sortie variable.

6 Conclusions et perspectives

Nous avons présenté dans ce papier notre méthode d'extraction des concepts adéquats pour représenter des images médicales. Nous avons travaillé avec une collection de 184614 images au total.

Les difficultés rencontrées sur cette tache ont été de quatre ordres :

— Nous ne disposions pas de machines forcément très adaptées aux algorithmes CNN ;

— Les images de la collection de données sont très variées et diversifiées (des images non médicales, des graphiques statistiques, des images composées) ;

— Le grand nombre de concepts (plus de 20000) à prédire est difficile à prendre en main ;

— Les techniques existantes s'intéressent à la classification de quelques classes seulement (1000 pour ImageNet) et pour une classification mono-label. Il a donc été nécessaire d'adapter les algorithmes existants pour une classification multi-label.

Les modèles que nous avons créés sont basés sur des transferts d'apprentissage Le premier modèle utilise VGG19 comme encodeur, et le décodeur pour avoir la classification multi-label est un réseau de neurones simple. En effet, pour adapter le modèle à la tâche visée nous avons inséré un réseau de neurones interconnectés à deux couches comme classifieur qui est facile à entraîner et nécessite donc moins de ressources. Il s'agit donc d'un modèle optimisé en matière de calcul et d'espace car nous étions limités en ressources.

Le second modèle est un réglage fin à partir de res Net50, ce dernier est plus lourd par rapport au précédent. Nous avons remarqué que le modèle que nous avons développé à tendance à prédire les concepts qui sont les plus fréquents (ou dominants) dans la collection d'entraînement. cet aspect mérite d'être corrigé dans le futur.

Pour les futurs travaux, nous visons également des améliorations en pré-classant les concepts en catégories avec une méthode de classification. Cela permettra de réduire le nombre de concepts à prédire. Nous envisageons également de traiter spécialement le cas des images composées avec des modèles CNN basés sur des méthodes de segmentation pour les détecter (par exemple les modèles développés dans (Ren *et al.*, 2015). Une fois que la segmentation des images en régions aura été réalisée, il sera alors possible d'utiliser les méthodes décrites dans ce papier pour extraire les concepts par région.

Remerciements

Experiments presented in this paper were carried out using the Grid'5000 testbed, supported by a scientific interest group hosted by Inria and including CNRS, RENATER and several Universities as well as other organizations (see https ://www.grid5000.fr).

Références

ABACHA A. B., DE HERRERA A. G. S., GAYEN S., DEMNER-FUSHMAN D. & ANTANI S. (2017). Nlm at imageclef 2017 caption task.

BENGIO Y., COURVILLE A. & VINCENT P. (2013). Representation learning : A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, **35**(8), 1798–1828.

BODENREIDER O. (2004). The unified medical language system (umls) : integrating biomedical terminology. *Nucleic acids research*, **32**, D267–D270.

COATES A., LEE H. & NG A. Y. (2010). An analysis of single-layer networks in unsupervised feature learning. *Ann Arbor*, **1001**(48109), 2.

DIMITRIS K. & ERGINA K. (2017). Concept detection on medical images using deep residual learning network.

EICKHOFF C., SCHWALL I., GARCÍA SECO DE HERRERA A. & MÜLLER H. (2017). Overview of ImageCLEFcaption 2017 - image caption prediction and concept detection for biomedical images.

HASAN S. A., LING Y., LIU J., SREENIVASAN R., ANAND S., ARORA T. R., DATLA V., LEE K., QADIR A., SWISHER C. *et al.* (2017). Prna at imageclef 2017 caption prediction and concept detection tasks.

HE K., ZHANG X., REN S. & SUN J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, p. 770–778.

IONESCU B., MÜLLER H., VILLEGAS M., ARENAS H., BOATO G., DANG-NGUYEN D.-T., DICENTE CID Y., EICKHOFF C., GARCIA SECO DE HERRERA A., GURRIN C., ISLAM B., KOVALEV V., LIAUCHUK V., MOTHE J., PIRAS L., RIEGLER M. & SCHWALL I. (2017). Overview of ImageCLEF 2017 : Information extraction from images. In *CLEF 2017 Proceedings*, volume 10456 of *Lecture Notes in Computer Science*, Dublin, Ireland : Springer.

ISLAM M. T., AOWAL M. A., MINHAZ A. T. & ASHRAF K. (2017). Abnormality detection and localization in chest x-rays using deep convolutional neural networks. *arXiv preprint arXiv :1705.09850*.

KARPATHY A. (2016). Cs231n convolutional neural networks for visual recognition. *Neural networks*, **1**.

LYNDON D., KUMAR A. & KIM J. (2017). Neural captioning for the imageclef 2017 medical image challenges.

MOTHE J., NY HOAVY N. & RANDRIANARIVONY M. I. (2017). IRIT & MISA at Image CLEF 2017 - Multi label classification. In *International Conference of the CLEF Association*, volume 1866 of *ISSN 1613-0073*, http ://CEUR-WS.org : CEUR Workshop Proceedings.

REN S., HE K., GIRSHICK R. & SUN J. (2015). Faster r-cnn : Towards real-time object detection with region proposal networks. In C. CORTES, N. D. LAWRENCE, D. D. LEE, M. SUGIYAMA & R. GARNETT, Eds., *Advances in Neural Information Processing Systems 28*, p. 91–99. Curran Associates, Inc.

RUSSAKOVSKY O., DENG J., SU H., KRAUSE J., SATHEESH S., MA S., HUANG Z., KARPATHY A., KHOSLA A., BERNSTEIN M. *et al.* (2015). Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, **115**(3), 211–252.

SHARIF RAZAVIAN A., AZIZPOUR H., SULLIVAN J. & CARLSSON S. (2014). Cnn features off-the-shelf : an astounding baseline for recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, p. 806–813.

SIMONYAN K. & ZISSERMAN A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv :1409.1556*.

STEFAN L.-D., IONESCU B. & MÜLLER H. (2017). Generating captions for medical images with a deep learning multi-hypothesis approach : Medgift–upb participation in the imageclef 2017 caption task.

TSOUMAKAS G., KATAKIS I. & VLAHAVAS I. (2009). Mining multi-label data. In *Data mining and knowledge discovery handbook*, p. 667–685. Springer.

VALAVANIS L. & STATHOPOULOS S. (2017). Ipl at imageclef 2017 concept detection task.

VINYALS O., TOSHEV A., BENGIO S. & ERHAN D. (2015). Show and tell : A neural image caption generator. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, p. 3156–3164.

YOSINSKI J., CLUNE J., BENGIO Y. & LIPSON H. (2014). How transferable are features in deep neural networks ? In *Advances in neural information processing systems*, p. 3320–3328.