Image Captioning with Semantic Attention

Janvier 2019 - Jérémie BOUSQUET

Mars 2016
IEEE Conference on Computer Vision and Pattern Recognition (CVPR)

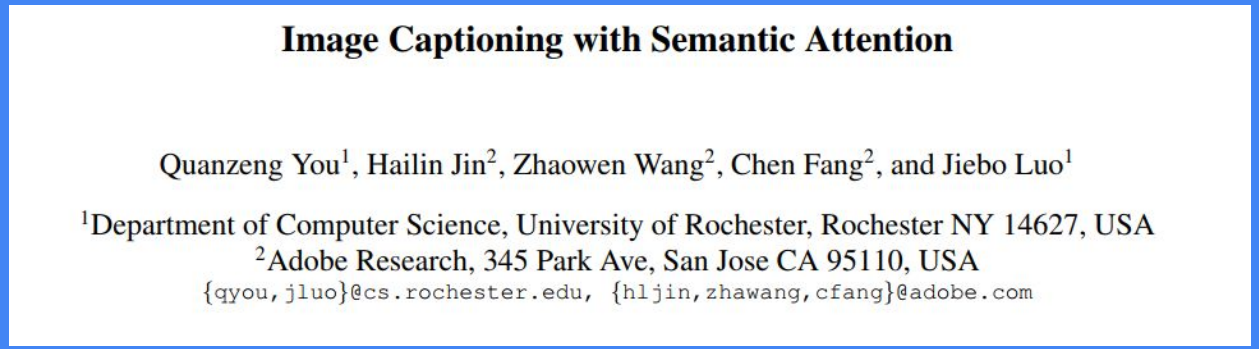


Image Captioning with Semantic Attention

Quanzeng You[1], Hailin Jin[2], Zhaowen Wang[2], Chen Fang[2], and Jiebo Luo[1]

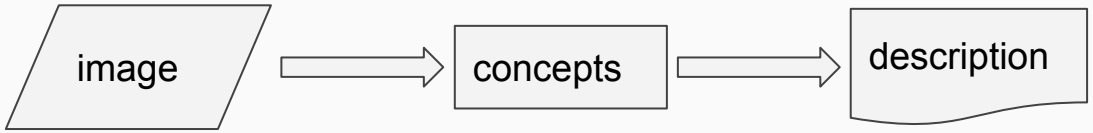
[1]Department of Computer Science, University of Rochester, Rochester NY 14627, USA
[2]Adobe Research, 345 Park Ave, San Jose CA 95110, USA
{qyou,jluo}@cs.rochester.edu, {hljin,zhawang,cfang}@adobe.com

Image captioning

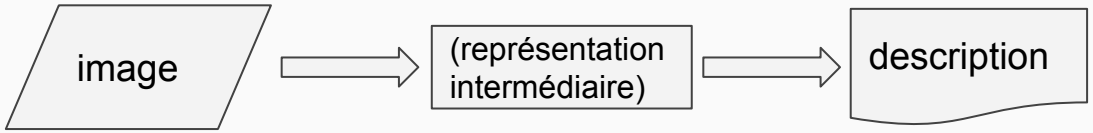
“a group of ducks swimming in a pond”

Image captioning - Deux approches

- Extraire mots et concepts de l'image, et utiliser un modèle de langage (bottom-up)

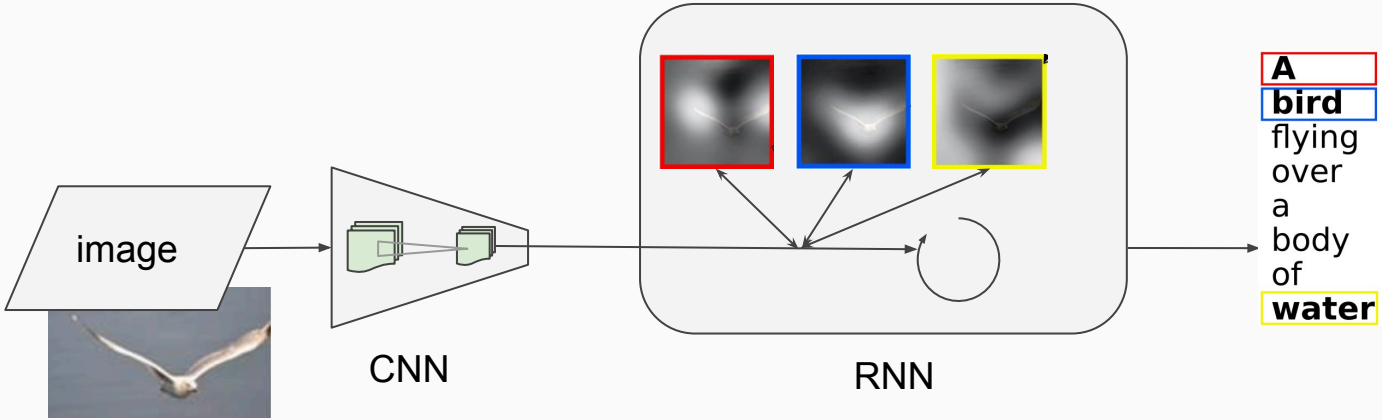


- **Traduire** une image sous forme textuelle (top-down)



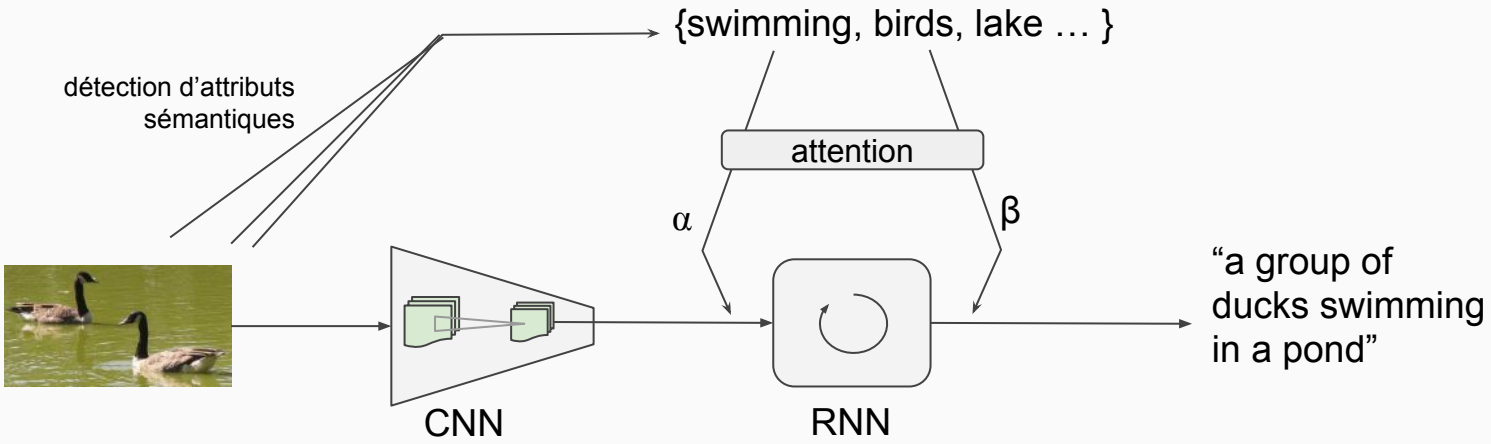
Travaux connexes - Show, Attend and Tell (2016)

- “Show, attend and tell: Neural image caption generation with visual attention”
  (Kelvin Xu, Aaron Courville, Kyuinghyun Cho - Montreal, Jimmy Lei Ba, Ryan Kiros, RSalakhu, Zemel - Toronto)



Mécanisme d’attention sur des parties de l’image, s’appuyant sur les filtres de la dernière couche de convolution du CNN (VGGNet / 14x14x512).

Architecture proposée



Détection des attributs

Méthode non paramétrique

- Méthode des plus proches voisins appliquée aux embeddings d'image (k-NN)
- Sélection des termes par TF (Term Frequency)

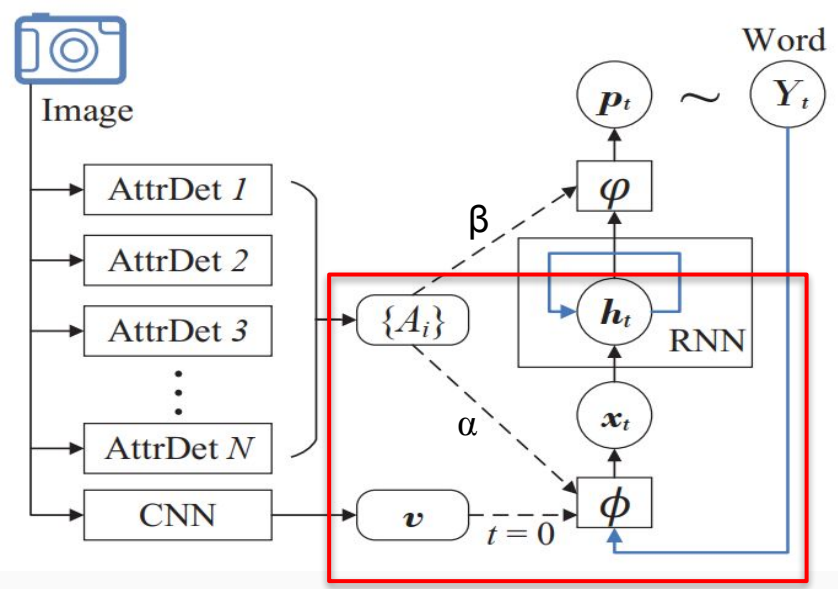
Détection des attributs

Deux méthodes paramétriques

- Utilisation de deux approches fondées sur des CNN
- Classifieur multi-label
  - Prédiction de plusieurs labels et estimation de probabilités associées (top-k ranking)
  - SuperVision / Toronto
- Fully Convolutional Network et patchs d'image
  - Extraction de régions d'intérêt
  - Prédiction d'attribut associé à chaque région

Dans les deux cas un score attribut vs image est calculé, et les attributs de meilleur score sont conservés.

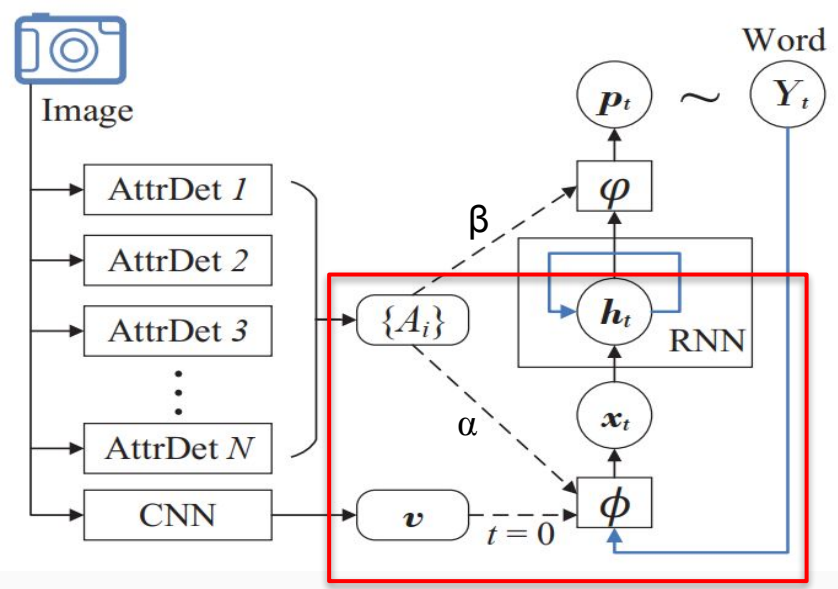
Attention en entrée du RNN



Chaque mot (légendes, attributs) fait partie d'un dictionnaire et est encodé "one-hot" (vecteur de dimension = taille du vocabulaire).

Des plongements (embeddings) sont utilisés pour réduire le nombre de paramètres à apprendre.

Attention en entrée du RNN



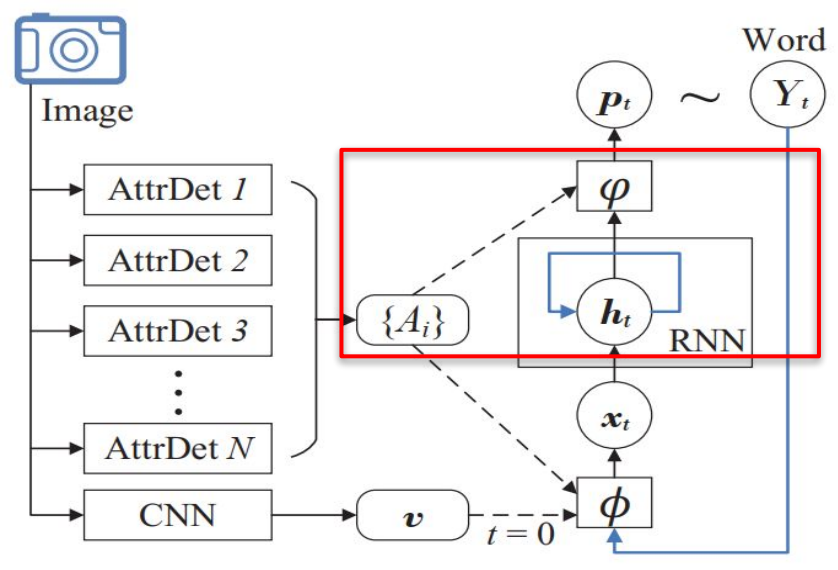
t0 : “initialisation” du RNN (embedding d’image v)

Aux t suivants, en entrée du RNN:

- le mot généré à t-1
- les attributs {Ai}

Chaque paramètre α_i^t pondère l’importance de l’attribut Ai à l’instant t (apprentissage d’une matrice α)

Attention en sortie du RNN



Même principe d'attention qu'en entrée mais on utilise:

- l'état du RNN h_t
- les attributs {Ai}
- les coefficients pour l'attention sont β_i^t

Expériences et résultats - apprentissage

- Données
  - Deux datasets utilisés :
    - Flickr30k (31 783 images)
    - MS-COCO (123 287 images)
  - 5 légendes par image annotées collaborativement
    - Amazon Mechanical Turk (AMT)
  - Le découpage (train/val/test) est celui utilisé pour le projet NeuralTalk de A.Karpathy (Stanford)
  - Les architectures proposées sont apprises sur chaque dataset séparément

Expériences et résultats - apprentissage

- RNN
  - LSTM constitué de deux couches (entrée et cachée) de taille 512
  - fonction d'activation : tanh
  - Plongements Glove de dimension 300
- CNN
  - Dernière couche de convolution de GoogleNet, de dimension 1024 (-> v)
- Paramètres d'apprentissage
  - RMSProp
  - batchs de 256

Apprentissage - paramètres pour l’attention

Termes de régularisation permettant d’ajuster le mécanisme d’attention

$$g(\boldsymbol{\alpha}) = \|\boldsymbol{\alpha}\|_{1,p} + \|\boldsymbol{\alpha}^T\|_{q,1}$$
$$= [\sum_i [\sum_t \alpha_t^i]^p]^{1/p} + \sum_t [\sum_i (\alpha_t^i)^q]^{1/q}$$

pénalise l’attention sur un attribut spécifique pour générer tous les Y_t (tous les mots d’une légende).

pénalise l’attention sur “trop” d’attributs différents à un instant t

Expériences et résultats - scores

Des métriques initialement conçues pour les traductions, peuvent être utilisées pour le captioning (légende générée vs légendes de référence).

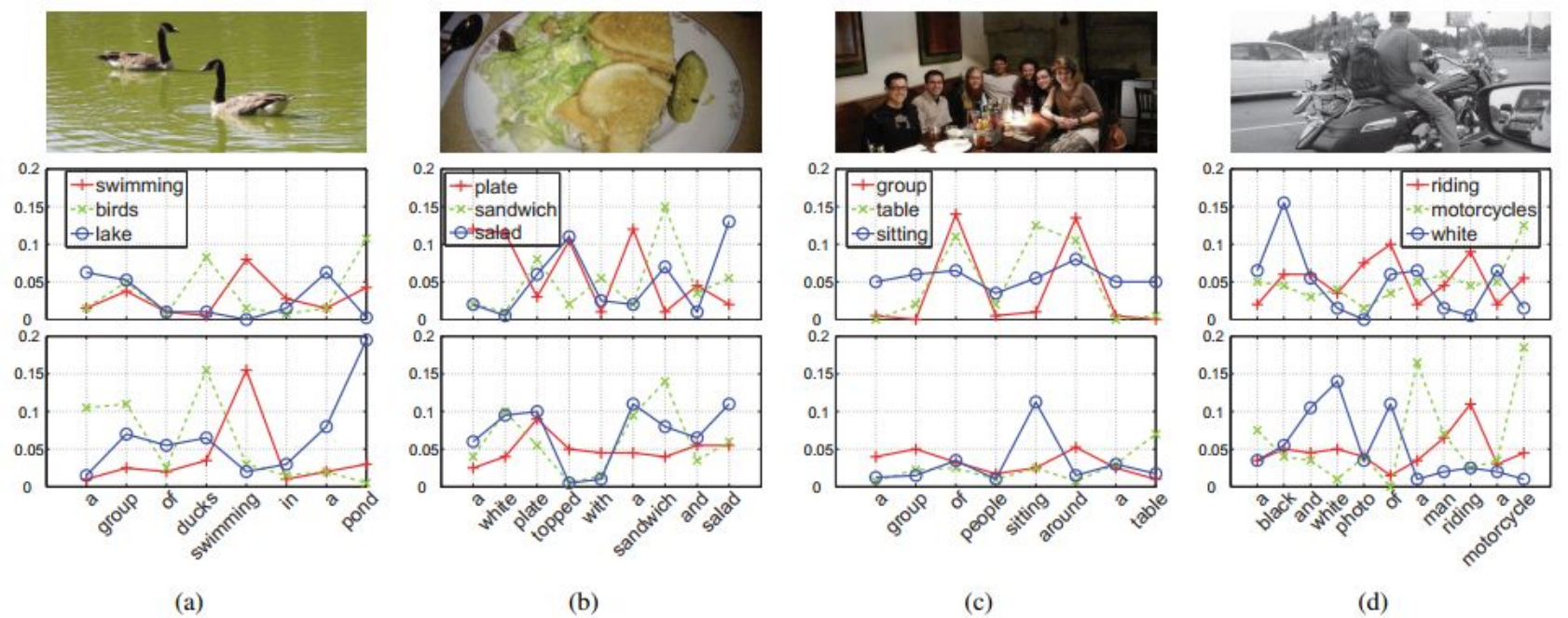
BLEU	BeLingual Evaluation Understudy (BLEU-n, n-grams)	Entre 0 (moins bon) et 1 (meilleur)
METEOR	Metric for Evaluation of Translation with Explicit ORdering	

Résultats - comparatif

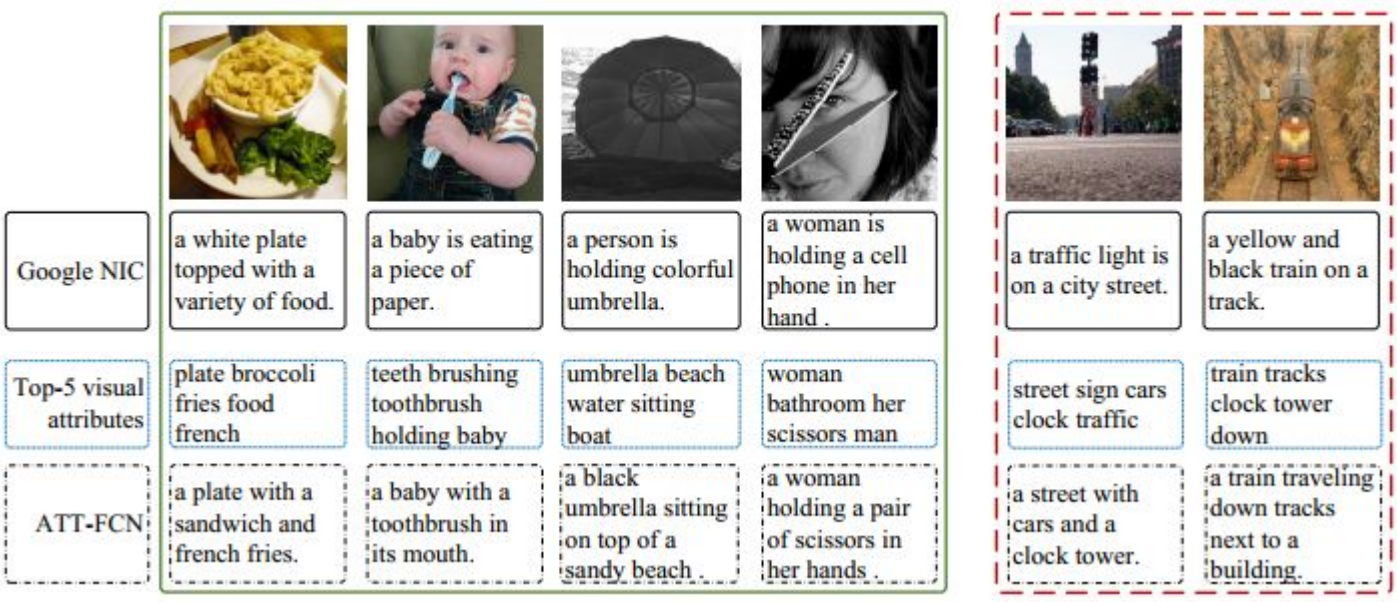
	Flickr30k					MS-COCO				
Model	B-1	B-2	B-3	B-4	METEOR	B-1	B-2	B-3	B-4	METEOR
Google NIC [35]	0.663	0.423	0.277	0.183	–	0.666	0.451	0.304	0.203	–
Toronto [37]	**0.669**	0.439	0.296	0.199	0.185	**0.718**	0.504	0.357	0.250	0.230
Ours-ATT-k-NN	0.618	0.428	0.290	0.195	0.172	0.676	0.505	0.375	0.281	0.227
Ours-ATT-RK	0.617	0.424	0.286	0.193	0.177	0.679	0.506	0.375	0.282	0.231
Ours-ATT-FCN	0.647	**0.460**	**0.324**	**0.230**	**0.189**	0.709	**0.537**	**0.402**	**0.304**	**0.243**

[35] : Show and Tell
[37] : Show, Attend and Tell

Exemples


swimming
birds
lake
a group of ducks swimming in a pond
(a)
plate
sandwich
salad
a white plate topped with a sandwich and salad
(b)
group
table
sitting
a group of people sitting around a table
(c)
riding
motorcycles
white
a black and white photo of a man riding a motorcycle
(d)

Exemples



Google NIC	a white plate topped with a variety of food.	a baby is eating a piece of paper.	a person is holding colorful umbrella.	a woman is holding a cell phone in her hand .	a traffic light is on a city street.	a yellow and black train on a track.
Top-5 visual attributes	plate broccoli fries food french	teeth brushing toothbrush holding baby	umbrella beach water sitting boat	woman bathroom her scissors man	street sign cars clock traffic	train tracks clock tower down
ATT-FCN	a plate with a sandwich and french fries.	a baby with a toothbrush in its mouth.	a black umbrella sitting on top of a sandy beach .	a woman holding a pair of scissors in her hands .	a street with cars and a clock tower.	a train traveling down tracks next to a building.

Conclusion

Un modèle pour la génération automatique de légende

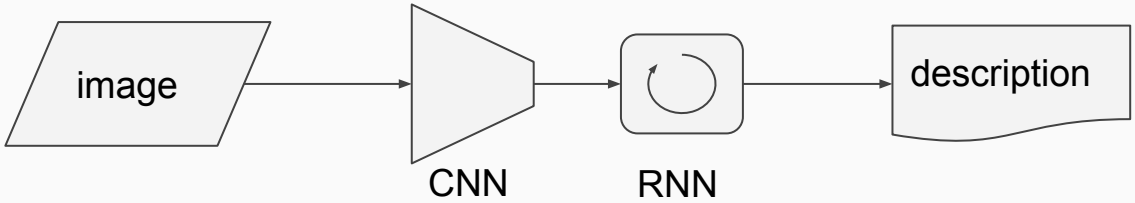
- extraction d’embeddings d’image
- extraction d’attributs sémantiques
- ménacisme d’attenton sur ces attributs

Partant de “Show, Attend and Tell”, les améliorations proposées permettent d’obtenir de meilleurs résultats.

Merci de votre attention

Travaux connexes - Show and Tell (2016)

- “Show and tell: a neural image caption generator” (Oriol Vinyals, Alexander Toshev, Samy Bengio, Dumitriu Erhan - Google)



Traduction d’une image en texte, représentation intermédiaire générée par le CNN.

CNN “encoder” (dernière couche dense), RNN “decoder” (LSTM).

Détection des attributs

Méthodes paramétriques

“**Deep convolutional ranking for multilabel image annotation”** Y. Gong, Y. Jia, T. Leung, A. Toshev, and S. Ioffe. Proceedings of the International Conference on Learning Representations (ICLR), 2014.”

“**From Captions to Visual Concepts and Back**” (Hao Fang,Saurabh Gupta,Forrest Iandola,Rupesh K. Srivastava,Li Deng,Piotr Dollar,Jianfeng Gao,Xiaodong He,Margaret Mitchell,John C. Platt,C. Lawrence Zitnick,Geoffrey Zweig - Microsoft Research. 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)

Expériences et résultats - attention

Dataset	Model	B-1	B-2	B-3	B-4	METEOR	ROUGE-L	CIDEr
Flickr30k	Ours-GT-ATT	0.824	0.679	0.534	0.412	0.269	0.588	0.949
	Ours-GT-MAX	0.719	0.542	0.396	0.283	0.230	0.529	0.747
	Ours-GT-CON	0.708	0.534	0.388	0.276	0.222	0.516	0.685
MS-COCO	Ours-GT-ATT	0.910	0.786	0.654	0.534	0.341	0.667	1.685
	Ours-GT-MAX	0.790	0.635	0.494	0.379	0.279	0.580	1.161
	Ours-GT-CON	0.766	0.617	0.484	0.377	0.279	0.582	1.237

- Attributs “ground truth” (annotations du dataset = non détectés)
- Pour estimation du mécanisme d’attention proposé:
  - ATT: mécanisme d’attention décrit plus haut
  - CON: concaténation des attributs
  - MAX: maximum (par composante) des attributs