Amélioration de la robustesse des réseaux de neurones multimodaux par identification et désactivation des modalités endommagées

Robin CONDAT, Alexandrina ROGOZAN, Samia AINOUZ, Abdelaziz BENSRHAIR

Normandie Univ, INSA Rouen, UNIROUEN, UNIHAVRE, LITIS
Avenue de l'Université, 76801 Saint-Étienne-du-Rouvray Cedex, France
robin.condat@insa-rouen.fr, alexandrina.rogozan@insa-rouen.fr
samia.ainouz@insa-rouen.fr, abdelaziz.bensrhair@insa-rouen.fr

Résumé – Pour la compréhension de scènes routières, la robustesse des systèmes de perception multimodaux est nécessaire, afin de garantir une fiabilité en cas de dysfonctionnement de capteurs. Dans ce contexte, une modalité d'entrée endommagée peut perturber leur bon fonctionnement. Dans cet article, nous proposons Noise Augmentation, une technique d'augmentation de données qui produit des modalités inexploitables pendant l'entraînement d'un réseau de neurones multimodal, et ModAM, un réseau de neurones convolutionnel qui pré-traite les modalités d'entrée, afin d'identifier et désactiver celles qui sont endommagées. Les expériences montrent que la combinaison de nos 2 contributions rend nos détecteurs d'objets significativement plus robustes en conditions dégradées, sans diminuer leurs performances globales en conditions normales.

Abstract – For road scene understanding, robustness of multimodal perception systems is necessary to ensure reliability in case of sensor malfunction. In this context, a noisy input modality may disrupt their proper operation. In this paper, we propose Noise Augmentation, a data augmentation technique that produces unusable modalities during a multimodal neural network training, and ModAM, a convolutional neural network that preprocesses input modalities, in order to identify and deactivate the unusable ones. Experiments show that the combination of our 2 contributions makes our object detectors significantly more robust in degraded conditions, without decreasing their overall performances in normal conditions.

1 Introduction

La détection des usagers de la route joue un rôle décisif dans le domaine des ADAS (Advanced Driver-Assistance Systems). La moindre défaillance peut avoir des conséquences importantes, avec des accidents dans le pire des cas. De nombreux travaux ont été réalisés ces dernières années pour assurer une conduite sûre malgré la complexité de cette tâche. Parmi eux, on retrouve des réseaux de neurones convolutionnels (CNN) utilisant la multimodalité, profitant de sources provenant de plusieurs capteurs. Cette stratégie apporte une meilleure robustesse face aux conditions météorologiques adverses, au manque de visibilité de certains objets ou dans des environnements complexes. Cependant, l'utilisation de multiples capteurs entraîne également davantage de possibilités de défaillances.

La robustesse des réseaux de neurones est donc un enjeu clé pour garantir une perte minimale de performance en cas d'événements inattendus. A notre grande surprise, peu de travaux abordent ce sujet dans le contexte ADAS. Parmi ces contributions, on retrouve des datasets axés sur la robustesse en conditions météorologiques [1, 2], des architectures spécifiques de CNNs pour prendre en charge des données déteriorées [3, 4] ainsi que des méthodes d'augmentation de données pour simuler des dysfonctionnements de capteurs durant l'entraînement de réseaux de neurones [5, 6]. Dans notre contexte, nous pensons qu'il est nécessaire de générer des données bruitées pour qu'un CNN multimodal puisse apprendre à les gérer. Ces modalités endommagées peuvent cependant apporter de fausses informations pouvant perturber le CNN. Par conséquent, il est également important d'identifier ces modalités bruitées, afin de les traiter au mieux pour éviter d'éventuels désagréments.

Dans cet article, nous proposons Noise Augmentation, une technique d'augmentation de données qui produit des modalités inexploitables pendant l'apprentissage d'un CNN pour améliorer sa robustesse en cas de dysfonctionnement de capteur. Nous introduisons ensuite le modèle d'activation de modalité (ModAM), un CNN qui pré-traite les modalités d'entrée afin d'identifier et désactiver celles qui sont endommagées. Cette stratégie vise à éliminer les fausses informations des modalités bruyantes afin de ne pas perturber le réseau multimodal, tout en apprenant à gérer l'absence d'une ou plusieurs modalités pendant l'apprentissage. Nous évaluons la robustesse d'un CNN multimodal dans différentes conditions dégradées et analysons l'impact de nos deux contributions sur ses performances.

2 Méthodes proposées

2.1 Noise Augmentation

Noise Augmentation génère des modalités endomagées durant l'apprentissage d'un CNN multimodal en ajoutant un bruit fort sur une partie des données d'entrée, simulant ainsi des dys-

fonctionnements de capteurs. L'objectif est de permettre au réseau de neurones d'apprendre à faire face à des modalités inexploitables contenant uniquement des informations erronées, et de les ignorer, au lieu d'essayer d'extraire des informations qui sont largement erronées. Avec Noise Augmentation, chaque modalité d'entrée sera soit intacte, soit fortement bruitée et donc inexploitable. Lorsqu'un bruit est appliqué sur une modalité, notre méthode sélectionne de manière aléatoire un type de bruit parmi un ensemble de méthodes de bruitage préalablement défini. Ainsi, une même modalité inexploitable sera représentée de différentes manières, de sorte que le réseau n'apprenne pas à identifier un seul bruit spécifique.

L'injection de données bruitées durant l'apprentissage d'un CNN multimodal permet d'améliorer sa robustesse, mais au détriment de performances parfois réduites en conditions normales si ces modalités bruitées représentent une part trop importante des échantillons d'apprentissage. Afin d'éviter ce phénomène, Noise Augmentation définit pour chaque modalité une probabilité de bruitage (Noise Augmentation Rate ou NAR), comprise entre 0 et 100 %, correspondant au pourcentage de modalités d'entrée qui seront endommagées. Les probabilités de bruitage des données d'entrée sont indépendantes les unes des autres et peuvent donc être différentes pour chaque modalité en fonction de ces caractéristiques, ce qui nous apporte plus de modularité. De plus, plusieurs modalités d'un même échantillon peuvent être bruitées par des méthodes différentes. Néanmoins, nous avons verrouillé la possibilité que Noise Augmentation bruite toutes les modalités d'un même échantillon afin d'en avoir au moins une correcte durant l'apprentissage.

2.2 ModAM

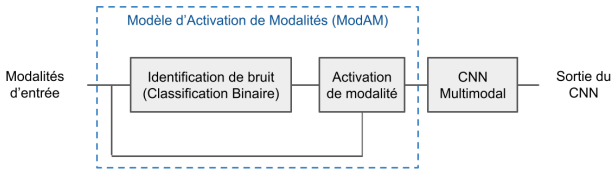

FIGURE 1 – Architecture de ModAM.

Le modèle d'activation de modalité (ModAM) que nous proposons intervient en amont d'un CNN multimodal pour analyser ses différentes modalités et éliminer celles qui sont inexploitables. Son architecture est illustrée dans la figure 1. L'identification du bruit par ModAM se fait grâce à un classifieur binaire, prenant en entrée chaque modalité et produisant en sortie des facteurs d'activation. Pour une modalité d'entrée, son facteur d'activation en sortie sera soit 1 (la modalité est correcte) soit 0 (la modalité est inexploitable). Ensuite, ModAM active les modalités d'entrée correctes et désactive les modalités endommagées en multipliant chaque canal de données d'entrée par son facteur d'activation. En supposant que le modèle fonctionne parfaitement, tous les canaux des modalités bruitées seront remplacés par une image nulle de même taille. Lorsqu'une modalité à plusieurs canaux obtient des facteurs d'activation différents pour ses canaux, ModAM fixe tous ces facteurs à 0 par défaut, privilégiant un faux négatif (modalité non bruitée désactivée) à un faux positif (modalité bruitée non filtrée). L'idée derrière ModAM est de simplifier toutes ces modalités endommagés par une même image noire ne contenant aucune information, plus facile à gérer pour le CNN multimodal.

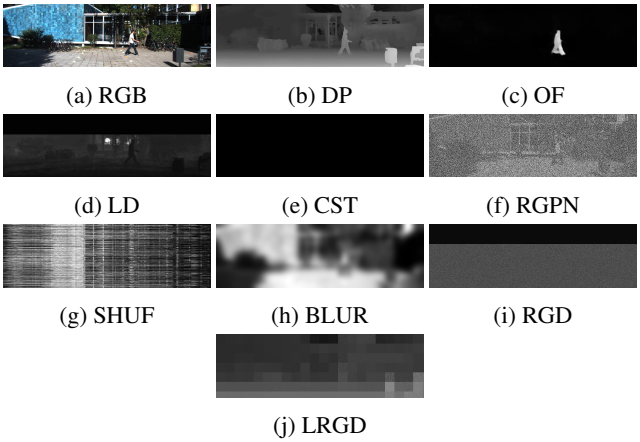

FIGURE 2 – Les modalités (a-d) et les méthodes de bruitages (e-j) appliquées sur différentes modalités d'entrée.

Le principal avantage de ModAM est sa modularité, puisqu'il est applicable à tout type de CNN multimodal, quel que soit le type de fusion appliqué. De plus, une modalité très bruitée est facilement distinguable d'une modalité non bruitée, ce qui signifie que le classifieur de ModAM peut être très léger, ralentissant légèrement le temps d'exécution.

3 Expérimentations

3.1 Dataset et modalités d'entrée

Pour nos expérimentations, nous utilisons KITTI 2D Object Detection Benchmark [7], qui se compose de 7481 images avec 39597 objets labelisés dans 8 classes différentes. Les images du dataset sont agrandies afin d'avoir une taille moyenne de $2000 \times 600 \times 3$ pixels. Les modalités d'entrée proviennent d'un système de capteurs comprenant 2 caméras RGB stéréo et un LiDAR. Par conséquent, nous sommes en mesure d'extraire quatre modalités : l'image couleur de la caméra gauche (RGB), la profondeur via stéréovision (DP), le flux optique à partir de deux images temporellement adjacentes (OF) et le nuage de points LiDAR dense projeté dans le plan de la caméra gauche (LD). Nous avons utilisé, pour leur extraction, GANet (Guided Aggregation Net) [8] pour la modalité DP, VCN (Volumetric Correspondence Network) [9] pour la modalité OF et NLSPN (Non-Local Spatial Propagation Network) [10] pour la modalité LD. Ces trois modalités, sous forme d'images à niveaux de gris, sont ensuite concaténées pour créer une image couleur appelée DOL. Un échantillon des différentes modalités extraites est présenté dans la figure 2.

Paramètres en entraînement		Conditions d'évaluation								
NAR	ModAM	Conditions Normales	Bruit sur cette modalité				Bruit sur toutes les autres modalités			
			RGB	DP	OF	LD	RGB	DP	OF	LD
0 %	N	92.4 ± 1.1	10.8 ± 3.7	71.1 ± 7.0	77.3 ± 8.3	63.7 ± 5.1	53.5 ± 13.0	0.0 ± 0.0	0.1 ± 0.1	0.0 ± 0.0
	Y	92.1 ± 0.6	15.8 ± 15.1	85.6 ± 2.2	90.6 ± 0.9	65.6 ± 4.6	60.5 ± 10.5	0.7 ± 0.7	0.2 ± 0.4	1.7 ± 2.0
5 %	N	92.3 ± 0.9	80.8 ± 1.3	91.6 ± 0.8	91.8 ± 0.9	91.4 ± 0.8	89.6 ± 0.9	33.6 ± 3.1	8.8 ± 0.8	44.8 ± 3.5
	Y	92.7 ± 1.0	82.2 ± 1.8	91.8 ± 0.9	92.3 ± 1.0	92.1 ± 0.9	89.5 ± 1.3	51.4 ± 2.4	12.8 ± 0.4	67.6 ± 2.0
25 %	N	92.6 ± 1.0	84.0 ± 1.8	92.3 ± 0.6	92.5 ± 0.9	92.0 ± 1.0	91.0 ± 0.8	61.8 ± 1.8	20.8 ± 0.9	68.0 ± 2.7
	Y	92.8 ± 0.7	84.7 ± 1.1	92.2 ± 0.6	92.6 ± 0.7	92.7 ± 0.6	90.8 ± 1.2	72.6 ± 2.0	25.7 ± 1.0	77.7 ± 2.0
50 %	N	92.3 ± 0.7	82.7 ± 2.0	91.9 ± 0.7	92.0 ± 0.8	91.9 ± 0.8	90.8 ± 0.9	68.3 ± 2.1	29.8 ± 2.1	72.8 ± 2.5
	Y	92.4 ± 0.5	83.8 ± 1.3	92.4 ± 0.6	92.6 ± 0.5	92.1 ± 0.5	91.0 ± 0.8	76.4 ± 1.5	36.0 ± 2.1	79.6 ± 1.4

TABLE 1 – Performances (mAP_{50} ± écart type) de Gated Fusion Double RetinaNet, en fonction de la probabilité de bruitage appliquée (NAR) et de la présence de ModAM (O pour Oui, N pour Non) en cas de bruit sur une ou plusieurs modalités.

3.2 Méthodes de bruitage utilisées

6 méthodes de génération de bruit, illustrées dans la figure 2, sont utilisées pour Noise Augmentation :

- CST : Cette méthode renvoie une image de la même taille remplie d'une seule valeur, entre 0 et 255 ;
- RGPN : Un bruit gaussien aléatoire est appliqué, avec une moyenne nulle et un grand écart type ;
- SHUF : La modalité de sortie contient les mêmes pixels que celle d'entrée mais mélangés de façon aléatoire ;
- BLUR : Un flou gaussien est appliqué sur la modalité avec un grand noyau ;
- RGD : La modalité d'entrée est remplacée par une distribution gaussienne aléatoire ayant la même moyenne et la même variance ;
- LRGD : RGD est appliqué localement sur la modalité via une grille carrée régulière.

3.3 Protocole expérimental

Pour nos expérimentations, nous utilisons GFD-Retina [11], un CNN multimodal prenant des images RGB et DOL en entrée. Afin d'analyser l'impact de nos contributions, nous avons varié nos réseaux selon 4 probabilités de bruitage associées à chaque modalité d'entrée : 0 % (aucun bruit), 5 %, 25 % et 50 %, et selon la présence ou non de ModAM en pré-traitement.

Les extracteurs de caractéristiques de GFD-Retina sont initialisés avec un ResNet50, et nous les entraînons via 2 RetinaNet [12] durant 100 itérations. Ensuite, nous transférons et bloquons leur poids vers GFD-Retina. Enfin, nous apprenons les couches supérieures de GFD-Retina, sur 100 itérations. Tous ces CNNs ont été entraînés avec un optimiseur Adam et un taux d'apprentissage de 10^{-5}. Nous avons utilisé une validation croisée à 5 dossiers pour confirmer les résultats.

Concernant ModAM, nous utilisons 5 AlexNet [13]. Nous modifions leur première couche de convolution pour qu'elle prenne en entrée des images à un canal, et leur dernière couche pour qu'elle ne produise qu'une seule valeur par modalité. Enfin, nous les entraînons durant 10 itérations avec un optimiseur Adam et un taux d'apprentissage de 10^{-6}.

4 Résultats et discussions

Nous évaluons nos CNNs sur leur dossier de validation en conditions normales et en conditions dégradées, en fonction de la probabilité de bruitage appliquée (NAR) et de la présence ou non de ModAM. Pour comparaison, nous utilisons la mean average precision avec un seuil d'intersection sur l'union de 50 % (mAP_{50}). Pour l'évaluation en conditions dégradées, nous avons créé plusieurs versions des sets de validation de nos réseaux avec une ou plusieurs modalités bruitées. Nous avons utilisé pour cela les mêmes méthodes que celles décrites en section 3.2. Pour chaque modalité d'entrée du CNN, nous évaluons le réseau sous 2 conditions : lorsque cette modalité est bruitée et lorsque toutes les autres modalités sauf celle-ci sont bruitées (ce qui est un cas extrême et très rare).

Le tableau 1 montre les performances de nos CNNs pour chaque condition. Tout d'abord, on constate que Noise Augmentation et ModAM n'affectent pas la mean average precision mAP_{50} de GFD-Retina en conditions normales. Cependant, il faudrait utiliser des métriques complémentaires pour pouvoir affirmer que nos deux contributions n'ont véritablement aucun impact dans cette configuration. Ensuite, l'utilisation de Noise Augmentation améliore considérablement la robustesse des CNNs. Plus le NAR augmente, plus les performances dans des conditions dégradées s'améliorent. Dans les cas où une seule modalité est bruitée, les métriques atteignent un plateau avec un NAR de 25 %. En revanche, nos CNNs avec un NAR de 50 % obtiennent les meilleurs résultats dans des conditions extrêmes où une seule modalité est correcte. Enfin, l'utilisation de ModAM donne de meilleurs résultats aux réseaux de neurones pour plusieurs conditions dégradées. Lorsque ce n'est pas le cas, les performances obtenues sont similaires. Cela s'explique par le fait qu'une modalité désactivée est plus simple à gérer, grâce à sa simplicité et son uniformité.

Nous avons également évalué nos CNNs en fonction du bruit appliqué aux images. Pour cela, nous avons fortement bruité 50 % de nos modalités de test avec une seule méthode de bruitage (CONST, RGPN, SHUF, BLUR, RGD et LRGD). Nous nous sommes assurés que chaque échantillon de test comporte au

Méthodes de bruitage	Sans ModAM	Avec ModAM
CST	77.88 ± 1.57	82.52 ± 1.84
RGPN	76.55 ± 1.77	82.52 ± 1.84
SHUF	78.47 ± 1.59	82.52 ± 1.84
BLUR	84.00 ± 1.67	82.52 ± 1.84
RGD	79.98 ± 1.36	82.52 ± 1.84
LRGD	81.46 ± 1.57	82.52 ± 1.84

TABLE 2 – Performances (mAP_{50} ± écart type) de Gated Fusion Double RetinaNet, selon la présence de ModAM et du bruit appliqué sur les modalités inexploitables.

moins une modalité bruitée et une modalité correcte.

Le tableau 2 montre les performances de GFD-Retina avec un NAR de 25 % pour chaque méthode de bruitage en fonction de la présence de ModAM. Nous constatons que sans ModAM, nos CNNs ont des performances qui varient en fonction du type de bruit appliqué. Avec ModAM, nos CNNs obtiennent les mêmes performances pour chaque méthode de bruit, puisque ce dernier désactive toutes les modalités bruitées, ainsi le jeu de données d'évaluation devient le même quel que soit le bruit. Hormis pour le bruit BLUR, ModAM améliore les résultats pour chaque méthode de bruit. La désactivation des modalités endommagées est donc une solution efficace pour améliorer la robustesse des CNN multimodaux. Cependant, compte tenu des résultats plus faibles de ModAM sur le bruit BLUR, des améliorations sont possibles sur son module de désactivation, possiblement en remplaçant les modalités bruitées par une matrice fixe plus adaptée à chaque modalité plutôt que par une image noire.

5 Conclusion

Cet article présente tout d'abord Noise Augmentation, une méthode d'augmentation de données pour améliorer la robustesse des CNNs multimodaux face aux possibles dysfonctionnements de capteurs, puis ModAM, un CNN de pré-traitement pour l'identification et la désactivation de modalités inexploitables. Les expériences montrent que la combinaison de nos deux contributions améliore significativement la robustesse de nos CNNs multimodaux en conditions dégradées, sans affecter leurs performances en conditions normales.

Cependant, nous avons étudié ici un cas plutôt binaire : soit notre modalité est parfaite, soit elle est totalement inexploitable. La réalité est toute autre, avec des modalités partiellement ou légèrement bruitées. Pour nos travaux futurs, nous souhaitons adapter ModAM pour prendre en charge des modalités bruitées plus complexes et partiellement exploitables. Nous allons ensuite nous assurer que ModAM généralise bien face à des bruits qu'il n'a jamais rencontré durant l'entraînement. Nous pensons également expérimenter nos contributions sur d'autres CNNs multimodaux utilisant d'autres modalités d'entrée pour diverses tâches de vision par ordinateur, afin de confirmer nos résultats obtenus avec d'autres modalités d'entrée et stratégies de fusion.

Références

[1] A. Pfeuffer, M. Schön, C. Ditzel, and K. Dietmayer, "The aduulm-dataset - a semantic segmentation dataset for sensor fusion," in *31th British Machine Vision Conference*, BMVA Press, 2020.

[2] R. Blin, S. Ainouz, S. Canu, and F. Meriaudeau, "The polarlitis dataset : Road scenes under fog," *IEEE Transactions on Intelligent Transportation Systems*, pp. 1–10, 2021.

[3] J. Kim, J. Koh, Y. Kim, J. Choi, Y. Hwang, and J. W. Choi, "Robust deep multi-modal learning based on gated information fusion network," in *Asian Conference on Computer Vision*, 2018.

[4] G.-H. Liu, A. Siravuru, S. Prabhakar, M. Veloso, and G. Kantor, "Learning end-to-end multimodal sensor policies for autonomous navigation," in *1st Annual Conference on Robot Learning*, 2017.

[5] S. de Blois, M. Garon, C. Gagné, and J.-F. Lalonde, "Input dropout for spatially aligned modalities," in *2020 IEEE International Conference on Image Processing (ICIP)*, 2020.

[6] R. Condat, A. Rogozan, and A. Bensrhair, "Random Signal Cut for Improving Multimodal CNN Robustness of 2D Road Object Detection," in *28th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN)*, 2020.

[7] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving ? the kitti vision benchmark suite," in *2012 IEEE Conference on Computer Vision and Pattern Recognition*, 2012.

[8] F. Zhang, V. Prisacariu, R. Yang, and P. H. Torr, "Ga-net : Guided aggregation net for end-to-end stereo matching," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2019.

[9] G. Yang and D. Ramanan, "Volumetric correspondence networks for optical flow," in *Advances in Neural Information Processing Systems*, 2019.

[10] J. Park, K. Joo, Z. Hu, C.-K. Liu, and I. S. Kweon, "Non-local spatial propagation network for depth completion," in *European Conference on Computer Vision (ECCV)*, 2020.

[11] R. Condat, A. Rogozan, and A. Bensrhair, "Gfd-retina : Gated fusion double retinanet for multimodal 2d road object detection," in *23rd International Conference on Intelligent Transportation Systems (ITSC)*, 2020.

[12] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *IEEE international conference on computer vision*, 2017.

[13] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in neural information processing systems*, vol. 25, 2012.