Missing Modalities Imputation via Cascaded Residual Autoencoder

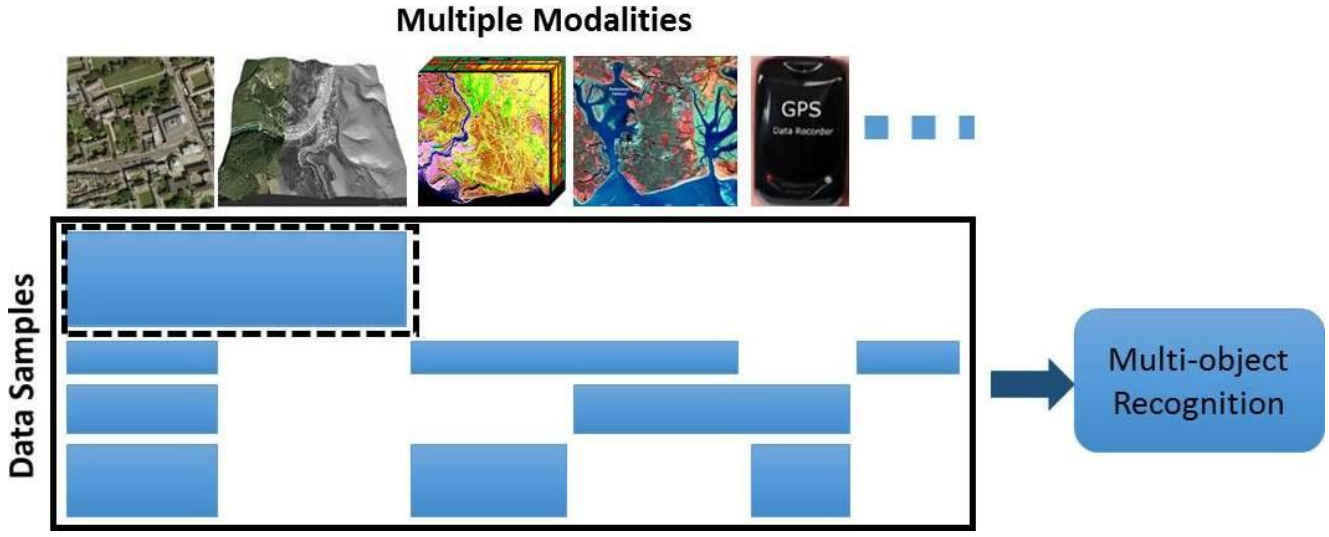
Luan Tran, Xiaoming Liu, Jiayu Zhou, Rong Jin

Department of Computer Science and Engineering, Michigan State University
Alibaba Group Holding Limited

Journal Club
Vendredi 10 Novembre 2017

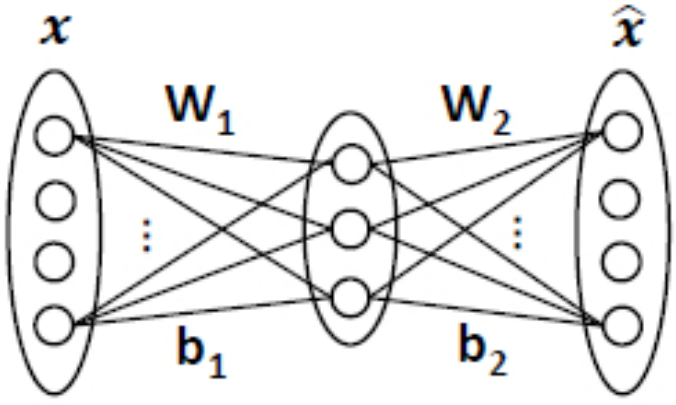
Introduction

- Analyse de données multimodales
  - images RGB
  - LiDAR
  - images hyperspectrales
  - ...
- Données incomplètes
  - appareils de mesure défectueux
  - données provenant de sources différentes



Autoencoder

- réseau de neurones avec une couche d'entrée comportant le même nombre de noeuds que la couche de sortie
- principe : réduire la dimension de la couche d'entrée, puis reconstruire une couche de sortie



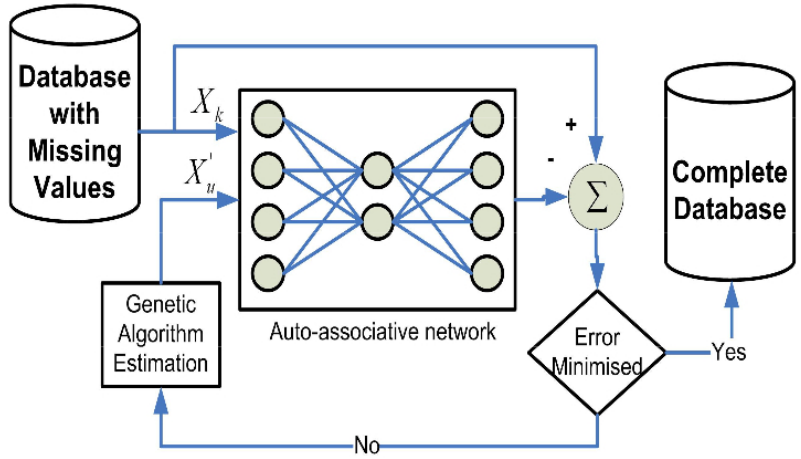
$$\begin{aligned} \mathbf{x} &= (x_1, \ldots, x_J) \\ \mathbf{z} &= f(\mathbf{W}^{(1)}\mathbf{x} + \mathbf{b}^{(1)}) \\ \hat{\mathbf{x}} &= f(\mathbf{W}^{(2)}\mathbf{z} + \mathbf{b}^{(2)}) \end{aligned}$$

$\mathbf{W}^{(1)} \in \mathbb{R}^{S \times J}$, $\mathbf{b}^{(1)} \in \mathbb{R}^{S}$, f fonction d'activation, $\mathbf{W}^{(2)} \in \mathbb{R}^{J \times S}$, $\mathbf{b}^{(2)} \in \mathbb{R}^{J}$

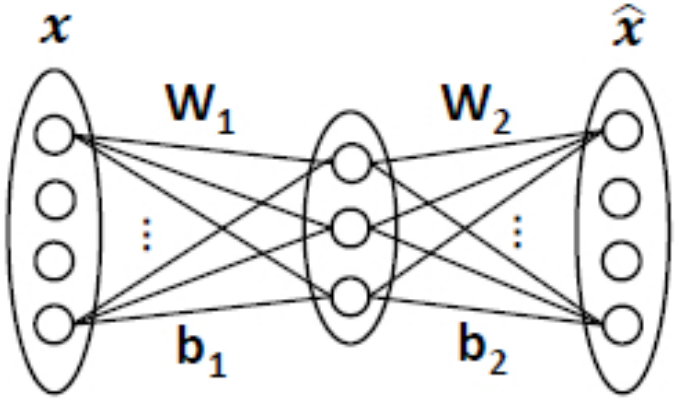
- généralisation à L couches : $\mathbf{a}^{(\ell+1)} = f(\mathbf{W}^{(\ell)}\mathbf{a}^{(\ell)} + \mathbf{b}^{(\ell)})$, $\hat{\mathbf{x}} = \mathbf{a}^{L}$
- optimisation : rechercher les matrices $\mathbf{W}$ et $\mathbf{b}$ tel que $\hat{\mathbf{x}}$ soit "proche" de $\mathbf{x}$

Imputation via AE

1. AE basé sur algorithmes génétiques [Abdella and Marwala, 2005]



2. AE débruité [Vincent et al., 2008]

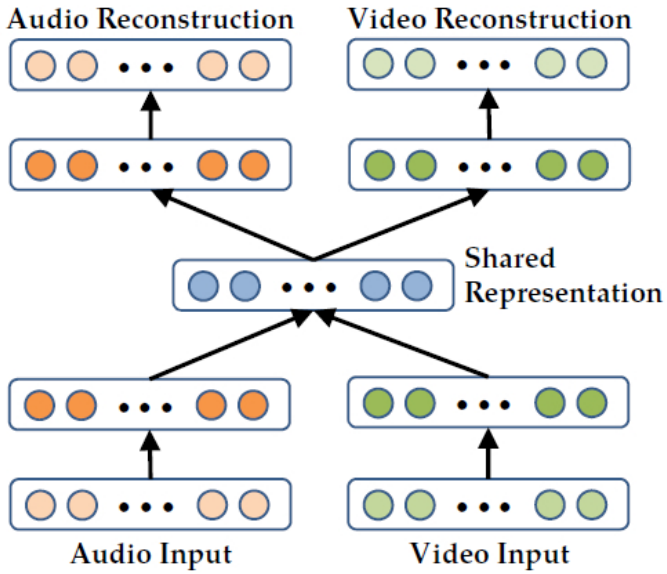


$$\begin{aligned} \mathbf{x} &= \mathbf{x}_k \\ \mathbf{z} &= f(\mathbf{W}^{(1)}\mathbf{x} + \mathbf{b}^{(1)}) \\ \hat{\mathbf{x}} &= f(\mathbf{W}^{(2)}\mathbf{z} + \mathbf{b}^{(2)}) \end{aligned}$$

$\mathbf{W}^{(1)} \in \mathbb{R}^{S \times J_k}, \mathbf{b}^{(1)} \in \mathbb{R}^{S}, \mathbf{W}^{(2)} \in \mathbb{R}^{J \times S}, \mathbf{b}^{(2)} \in \mathbb{R}^{J}$

AE pour données multi modales

1. Mul-mod AE [Ngiam et al., 2011]
   - capter les liaisons entre données multi-modales en fusionnant des couches latentes
   - valeurs manquantes gérées par AE débruité



2. Deep canonically correlated AE [Wang et al., 2015]
   - Analyse canonique des corrélations : identifier les couples de combinaisons linéaires des variables (une pour un groupe, une pour le second groupe) les plus corrélées entre elles.
   - Extension aux AE : AE au lieu de combinaisons linéaires
   - Proposé uniquement pour clustering

Challenges

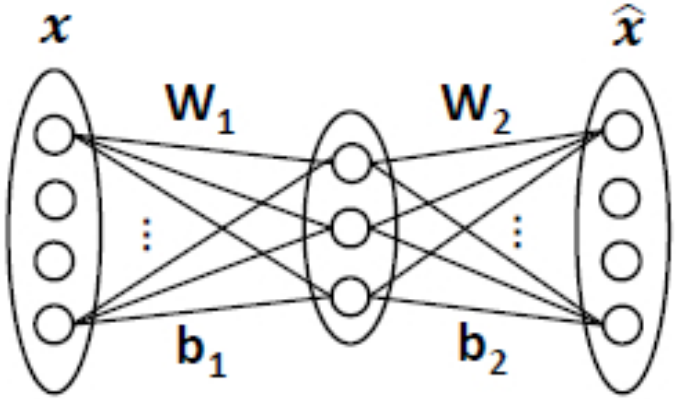
- En présence de données manquantes par blocs, les relations entre variables sont difficiles à identifier
- L'imputation par AE débruité multicouche n'est pas performante
- Hypothèse avancée : les paramètres du réseau sont difficiles à estimer

- Proposer un autre type de réseau et une procédure d'estimation adaptée

Plan

Residual AE

- Autoencoder



$$\begin{aligned} \mathbf{a}^{(0)} &= \mathbf{x} = (x_1, \ldots, x_J) \\ \mathbf{a}^{(\ell+1)} &= f(\mathbf{W}^{(\ell)}\mathbf{a}^{(\ell)} + \mathbf{b}^{(\ell)}) \\ \mathbf{a}^{(L)} &= \hat{\mathbf{x}} \end{aligned}$$

- Residual AE : la couche de sortie est $\Delta_x = \hat{\mathbf{x}} - \mathbf{x}$
  N.B. : En sommant entrée et sortie, on retrouve $\hat{\mathbf{x}}$

RAE pour images 2D

Images

- grande quantité d'information
- structure spatiale

(R)AE peu performants

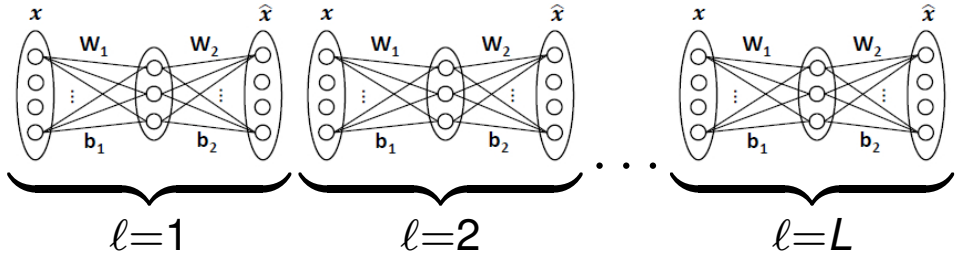
- explosion du nombre de paramètres du réseau ($J = 10000$ pour une image 100×100)
- pas de prise en compte des ressemblances locales entre les pixels

Convolutional RAE : $\mathbf{a}^{(\ell+1)} = f(\mathbf{a}^{(\ell)} * \mathbf{W}^{(\ell)} + \mathbf{b}^{(\ell)})$

- transformation à partir des pixels voisins
- $\mathbf{W}$ ne contient que les différents poids

Cascaded residual AE

- principe : succession d'autoencoders



- en termes de residus

$$\begin{aligned}\hat{\mathbf{x}} &= \hat{\mathbf{x}}^L + \Delta_{\mathbf{x}^L} \\ &= \hat{\mathbf{x}}^{L-1} + \Delta_{\mathbf{x}^{L-1}} + \Delta_{\mathbf{x}^L} \\ &= \dots \\ &= \hat{\mathbf{x}}^0 + \sum_{\ell=0}^{L} \Delta_{\mathbf{x}^\ell}\end{aligned}$$

avec $\hat{\mathbf{x}}^0 = x_k$ et $\Delta_{\mathbf{x}^0} = \mathbf{0}$

Optimisation

Fonction de coût

$$\mathbb{L} = \frac{1}{2} \| \mathbf{x} - \hat{\mathbf{x}} \|_2^2 \quad = \quad \| \mathbf{x} - \left(\hat{\mathbf{x}}^0 + \sum_{\ell=0}^{L} \Delta_{\mathbf{x}^\ell} \right) \|_2^2$$

Optimisation par “rétro-propagation" pour Cascaded RAE

- initialiser les paramètres
- mise à jour par descente de gradient pour le RAE L (répéter jusqu’à convergence)
- passer à la couche suivante $L - 1$, etc

Deux stratégies :

- one-shot : les paramètres estimés pour la couche ℓ reste **fixés** durant l’estimation des couches suivantes
- aggresive : les paramètres estimés pour la couche ℓ servent de valeurs initiales pour la couche suivante mais sont **remises à jour** ensuite

Plan

Complétion de matrices

$$\widehat{\mathbf{X}} \quad = \quad argmin_{\mathbf{M} \in \mathbb{R}^{I \times J}} \parallel X - M \parallel_F^2 \;\; avec \; \phi(\mathbf{M}) < c \qquad (1)$$

$\parallel . \parallel_F^2$ la norme de Frobenius.

$\phi(\mathbf{M}) \leq c$ est une contrainte sur M

- $\parallel \hat{\mathbf{X}} \parallel_{\ell_1} \leq c$ (SVT [Cai et al., 2010] / Soft Impute [Mazumder et al., 2010])
- rang $(\hat{\mathbf{X}}) < c$ (OptSpace [Keshavan et al., 2010])

Les solutions sont données par la SVD de $\mathbf{X}$ (seuillage doux / seuillage dur)

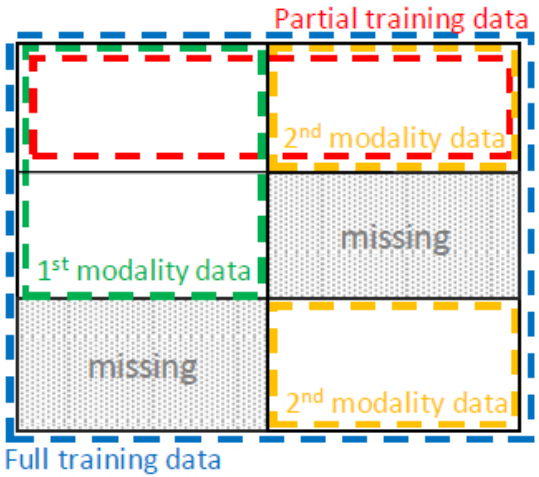
On sait résoudre le problème en présence de données manquantes

$$\widehat{\mathbf{X}} \quad = \quad argmin_{\mathbf{M} \in \mathbb{R}^{I \times J}} \parallel \mathbf{W} \star (X - M) \parallel_F^2 \;\; avec \; \phi(\mathbf{M}) < c \quad (2)$$

Protocole

A partir de 4 jeux réels incomplets

1. imputation
   - Echantillon d'apprentissage : individus sans données manquantes
   - Echantillon test : échantillons avec des modalités manquantes
   - Critères d'erreur NMSE : $\frac{\|\mathbf{X}-\widehat{\mathbf{X}}\|}{\|\mathbf{X}\|}$ et PSNR : $10 log_{10} \frac{J}{(\mathbf{x}-\widehat{\mathbf{x}})^\top (\mathbf{x}-\widehat{\mathbf{x}})}$
2. classification



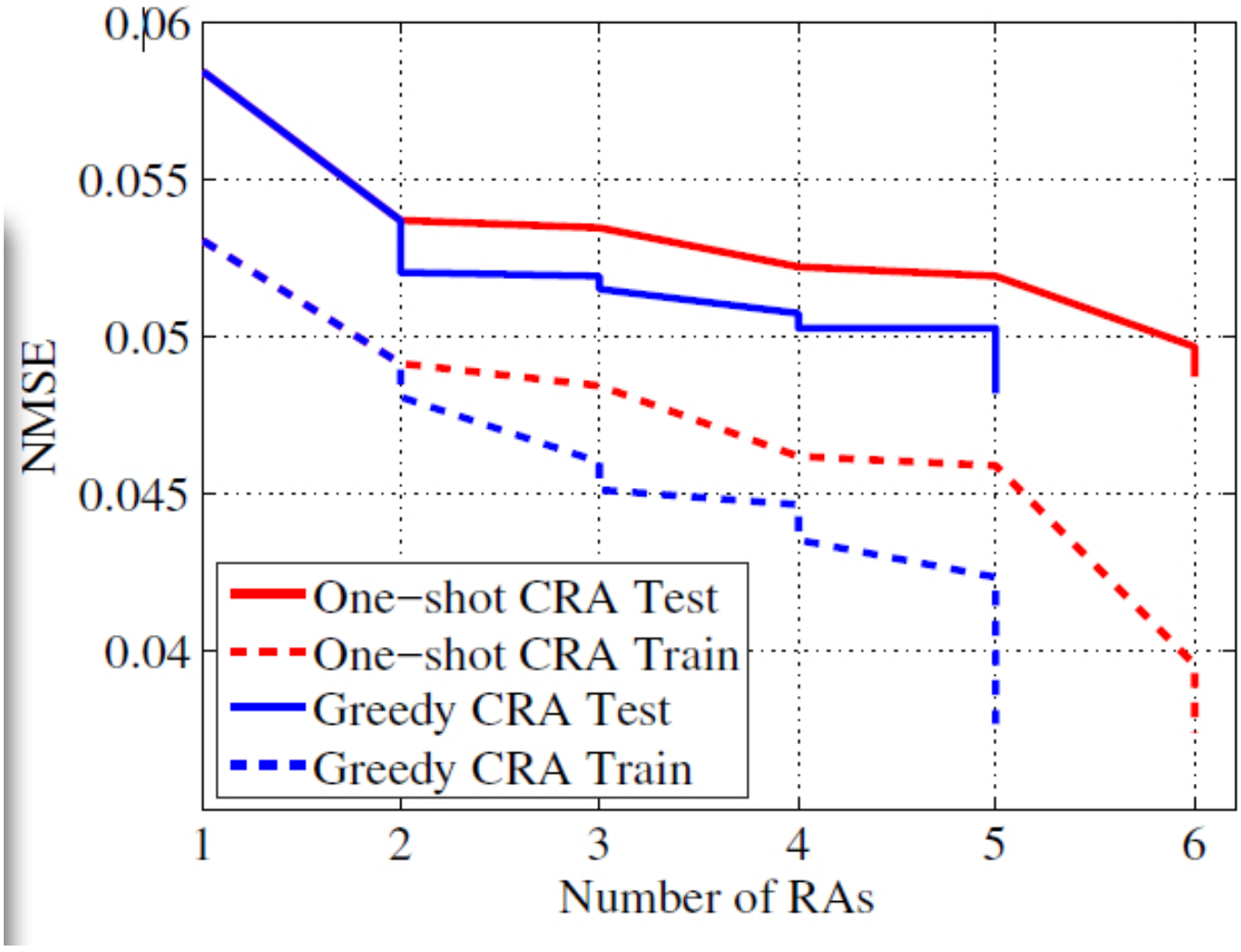
- Méthode de classification adaptée aux données (SVM, Convolutional Neural Network, ...)
- Comparaison à des méthodes sans imputation
- Critères d'erreur : taux de bon classement

Jeux de données

- GRSS : HSI et LiDAR d'un campus universitaire
- RGB-D : 41 877 images RGB-D (images et position entre objet et pixel)
- MTPIE : 754 200 portraits sous 5 angles différents
- HSFD : HSI de portraits (24 bandes)

Dataset	m	d	n_t	n_e	$r(\%)$	c	n	k
GRSS	2	[111, 37]	~ 20	~ 180	45	15	~ 200	~ 1000
RGB-D	2	[2500, 2500]	~ 170	~ 513	40	51	~ 683	~ 138
MTPIE	5	[1024,...,1024]	1529	729	50	137	1529	781
HSFD	24	[625, ..., 625]	76	38	40	38	76	$2-5$

Résultats imputation : one shot vs agressive



Résultats imputation : nombre de couches

Table 2: Affects of CRA depth on HSFD performance.

Depth	PSNR	NMSE	Recognition rate (%)
1	26.27	0.233	76.42
2	27.42	0.223	77.79
3	27.69	0.218	77.88
4	27.91	0.213	77.95
5	28.01	0.209	78.03

- fixé à 5 par la suite

Résultats imputation : pattern NA

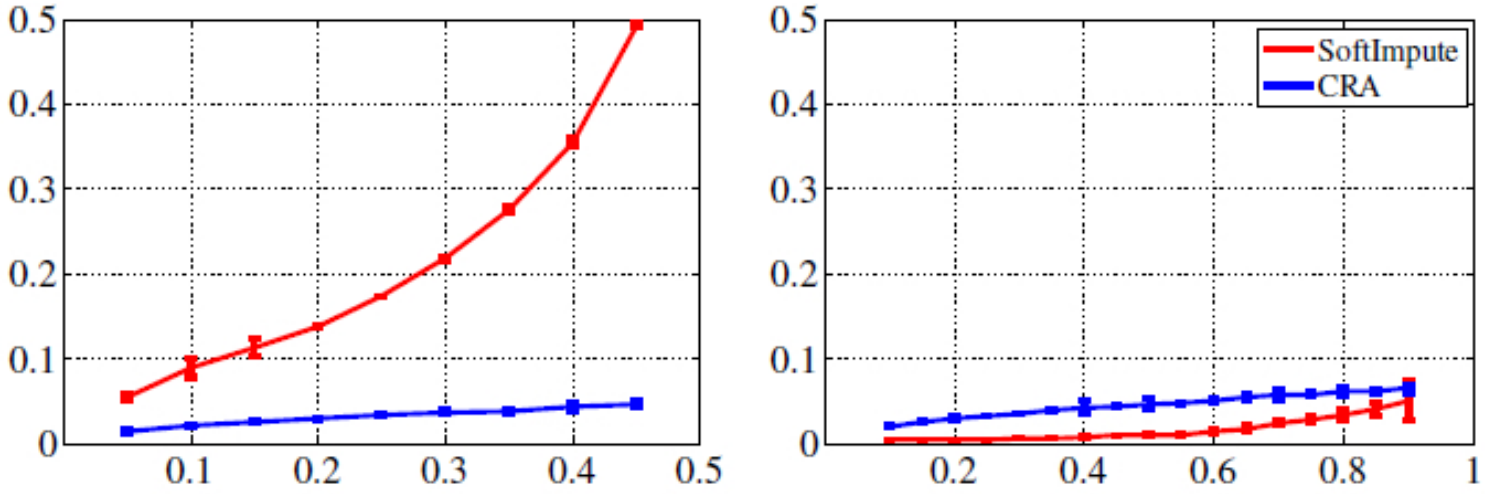


Figure 8: Compare blockwise (left) and elementwise (right) missing data imputation with different missing ratios on GRSS class of soil. In the blockwise corruption, we have to keep at least 1 of 2 modalities, hence the missing ratio must be less than 0.5.

Résultats imputation : comparaisons

		PSNR				NMSE			
		GRSS	RGB-D	MTPIE	HSFD	GRSS	RGB-D	MTPIE	HSFD
Methods	SVT [3]	-	21.51	17.02	14.47	-	0.548	0.307	0.633
	SoftImpute [26]	17.27	22.39	19.86	*26.33*	0.271	0.401	0.241	*0.232*
	OptSpace [20]	12.88	20.17	-	23.82	0.421	0.611	-	0.353
	AE&GA [1]	26.81	23.01	20.73	18.23	0.105	0.362	0.225	0.401
	DA [33]	29.62	24.05	22.09	24.38	0.079	0.341	0.171	0.276
	sDA [33]	29.74	24.07	23.31	20.13	*0.079*	*0.331*	0.151	0.372
	Mul-mod AE [30]	30.01	*24.81*	*25.91*	18.23	0.105	0.362	*0.113*	0.401
	DCCAE [34]	*30.05*	24.11			0.079	0.341		
	CRA w/o opt	30.80	25.25	25.99	26.63	0.077	0.265	0.112	0.225
	CRA w/ opt	**31.04**	25.93	26.55	27.51	**0.076**	0.248	0.105	0.222
	Conv CRA		**26.12**	**27.05**	**28.01**		**0.234**	**0.093**	**0.209**
Bounds	Full data								
	1^{st} modality only								
	2^{nd} modality only								
	Partial data								
	Relative margin	3.29	5.28	4.40	6.38	3.79	2.93	17.70	9.91

Résultats classification : comparaisons

		Recognition rate (%)			
		GRSS	RGB-D	MTPIE	HSFD
Methods	SVT [3]	-	21.17	38.73	74.24
	SoftImpute [26]	86.22	23.38	40.52	*76.36*
	OptSpace [20]	83.97	20.79	-	75.02
	AE&GA [1]	84.01	68.83	42.33	74.24
	DA [33]	85.77	69.17	44.34	76.02
	sDA [33]	86.02	69.54	45.32	75.96
	Mul-mod AE [30]	86.01	*70.21*	*54.24*	76.20
	DCCAE [34]	*86.25*	69.94		
	CRA w/o opt	86.38	70.63	54.32	76.58
	CRA w/ opt	**86.42**	71.04	56.42	77.88
	Conv CRA		**71.81**	**57.10**	**78.03**
Bounds	Full data	**88.46**	**78.32**	**59.24**	**80.00**
	1^{st} modality only	79.52	*68.72*		
	2^{nd} modality only	63.91	64.48		
	Partial data	*83.24*	56.51	*41.14*	*72.21*
	Relative margin	3.26	12.50	16.02	21.43

Conclusion

En résumé

- méthode d'imputation par autoencoder
- démontre l'intérêt des approches deep learning pour l'imputation d'images

Limites

- apport du residual AE ?
- rôle de l'hypothèse MAR ?
- protocole différent pour méthodes de complétion
- d'autres méthodes plus performantes [Verbanck et al., 2013] en complétion
- Random-forest non évoqué
- pas de vraie gestion de la multimodalité

References I

Mussa Abdella and Tshilidzi Marwala. The use of genetic algorithms and neural networks to approximate missing data in database. In *Computational Cybernetics, 2005. ICCC 2005. IEEE 3rd International Conference on*, pages 207–212. IEEE, 2005.

Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. Extracting and composing robust features with denoising autoencoders. In *Proceedings of the 25th International Conference on Machine Learning*, ICML '08, pages 1096–1103, New York, NY, USA, 2008. ACM. doi : 10.1145/1390156.1390294.

Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, and Andrew Y Ng. Multimodal deep learning. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pages 689–696, 2011.

Weiran Wang, Raman Arora, Karen Livescu, and Jeff Bilmes. On deep multi-view representation learning. In *Proceedings of the 32nd International Conference on Machine Learning (ICML-15)*, pages 1083–1092, 2015.

Jian-Feng Cai, Emmanuel J. Candes, and Zuowei Shen. A singular value thresholding algorithm for matrix completion. *SIAM Journal on Optimization*, 20(4) :1956–1982, 2010. doi : 10.1137/080738970.

Rahul Mazumder, Trevor Hastie, and Robert Tibshirani. Spectral regularization algorithms for learning large incomplete matrices. *Journal of machine learning research*, 11(Aug) :2287–2322, 2010.

References II

R. H. Keshavan, A. Montanari, and S. Oh. Matrix completion from a few entries. *IEEE Transactions on Information Theory*, 56(6) :2980–2998, 2010. ISSN 0018-9448. doi : 10.1109/TIT.2010.2046205.

Marie Verbanck, Julie Josse, and Francois Husson. Regularised PCA to denoise and visualise data. *Statistics and Computing*, pages 1–16, 2013. ISSN 0960-3174, 1573-1375.