Bases de données multimédia

Médias multiples : problématique et approches

Michel Crucianu (prenom.nom@cnam.fr)
http ://cedric.cnam.fr/~crucianm/bdm.html

Département Informatique
Conservatoire National des Arts & Métiers, Paris, France

19 février 2018

Plan du cours

Contenus multimédia et usages associés

- Contenus multimédia : mettent en œuvre plusieurs « médias » dont image fixe ou animée, son, texte ou hypertexte
- Types d'usages :
  1. Pluri-médias : utilisation conjointe de plusieurs médias
     - Recherche tenant compte du texte et de l'image
     - Catégorisation tenant compte du texte et de l'image
  2. Trans-média : passage d'un média à un autre
     - Classiques : annotation d'images, illustration de textes
     - Plus récentes : *image captioning* (génération de phrase descriptive à partir d'une image), *visual question answering* (réponse à une question à partir d'une image)

Médias multiples : redondance et complémentarité

Exemple : image + mots-clés

1. Objets identifiables (et caractéristiques visuelle de ces objets) : char, voiture, immeuble, personnes → redondance
2. Caractéristiques visuelles de la scène : extérieur, jour, ville, place, foule passive → redondance
3. Interprétation de la scène : fraternisation → selon le cas, redondance ou complémentarité
4. Contexte : Lisbonne, révolution des œillets, 25 avril 1974, Caetano, Salazar → complémentarité

Recherche conjointe texte + image : illustration

- Combiner la similarité entre textes et la similarité entre images

FIG. – Requête → résultats

(sources : `http ://www.miramas.org`, `https ://www.latribune.fr`, `https ://www.rds.ca`)

Annotation d'images : illustration

- Image → mots (objets identifiables et caractéristiques visuelles de la scène)
  - Base Pascal VOC 07 :

Images				
Tags	buildings, windows, cars, tree, streetlight, road	window, chair, table, image	cloth, cat	bus, banner, road, car
Labels	car	chair, diningtable, pottedplant	cat	car, bus, person

  - Base NUS-WIDE :

Images				
Tags	monochrome, car, vintage, mercedes, benz, cilest, kurt, vehicle	dog, husky, wolf, perro, lobo, roja, caperucita, siberiano, vorfas	blue, summer, sky, umbrella, holidays, ysplix, flickelite, colorartaward, platinumheartaward	ocean, sunset, sea, water, animals, whale, whales, orca, whalewatching, whaletail, orcawhale, whalephotos
Labels	vehicle	dog	sky	ocean, sunset, water, whales

Illustration de textes

- Texte → image(s) adéquate(s) pour illustrer le texte

BBC Sign in | News | Sport | Weather | Shop | Earth | Travel | More | Search

NEWS

Home | Video | World | UK | Business | Tech | Science | Magazine | Entertainment & Arts | Health | World News TV | More

Business | Market Data | Markets | Economy | Companies | Entrepreneurship | Technology of Business | More

Canada explores purchase of 18 interim Boeing Super Hornets

22 November 2016 | US & Canada

Share

Canada will explore buying 18 new Boeing Super Hornet fighter jets.

The Liberal government says the jets will close the "capability gap" in Canada's air power as it seeks a permanent replacement to its CF-18 fleet.

It will also launch a five-year procurement process in 2017 to find a replacement, which could include F-35s.

Ottawa will continue to participate in the F-35 Joint Strike Fighter (JSF) program.

Federal Defence Minister Harjit Sajjan said on Tuesday that Ottawa will immediately begin discussions with the US government and Boeing, which manufactures the Super Hornet fighter aircraft, on the purchase of the stop-gap fleet.

Canada plans to add 18 new Super Hornets to its RCAF fleet

Public Services and Procurement Minister Judy Foote was unable to provide an estimated cost for the 18 new jets, saying it would depend on the negotiations.

Image captioning : illustration

- Image → phrase(s) décrivant le contenu de l'image

A little girl sitting on a bed with a teddy bear.

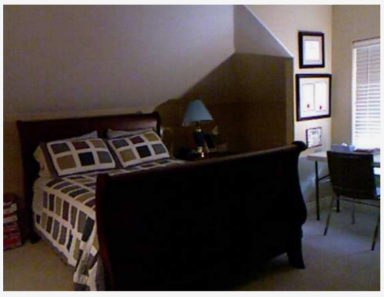
A group of people sitting on a boat in the water.

(source [19])

Visual question answering : illustration

- Image + question (phrase) → réponse sous forme de mot(s)





	What is on the right side of the cabinet?	How many drawers are there?	What is the largest object?
Neural-Image-QA:	bed	3	bed
Language only:	bed	6	table

(source [10])

Difficultés majeures

- Le fossé sémantique : le niveau des informations n'est pas le même
  - Contenu textuel : niveau sémantique élevé (∼concepts)
  - Image ou vidéo : niveau sémantique très bas (pixels…)
- Le fossé d'hétérogénéité : l'espace de représentation n'est pas le même
  - Contenu textuel : représentations vectorielles de données textuelles
  - Image ou vidéo : matrice de pixels (ou séquence de matrices), ensembles de points d'intérêt, …

→ Comment combiner les deux ou passer de l'un à l'autre ?

Plan du cours

Extraction d'entités primaires, identification d'entités nommées

- Entités primaires : mots simples ou composés (« chauve-souris »), éventuellement locutions nominales (« chemin de fer »), verbales (« arrondir les angles »)...
  - Utilisation d'un lexique : les *lemmes* (∼mots) d'une langue et/ou d'un domaine particulier
    - Informations additionnelles : morphologiques (formes possibles, à partir de racine et suffixes, préfixes), parfois syntaxiques
    - Peut souvent être enrichi par des lemmes et des locutions spécifiques au domaine
  - L'extraction de certaines locutions peut exiger des traitements plus complexes que la recherche dans un lexique
- Entité nommée : élément du langage qui fait référence à une entité unique, appartenant éventuellement à un domaine spécifique
  - Exemples : noms de personnes (« Barack Obama Jr. »), de lieux (« Mont Blanc »), d'organisations (« Mouvement international de la Croix-Rouge et du Croissant-Rouge », « ICRC »), dates...
  - Approches : règles explicitement définies, apprentissage à partir d'un *corpus* annoté, approches hybrides
  - Difficultés : polysémie (« Washington » : la ville ou l'état ou la personne ?), métonymie (« l'Elysée » : la présidence de la République Française ou le palais ?) ⇒ le contexte doit être pris en compte

Lemmatisation ou racinisation

- Objectif : représentation unique pour les formes fléchies d'un même lemme
- Racinisation :
  - Remplacer un mot (par ex. « pensons ») par sa racine (« pense »)
  - En général basée sur des règles et un lexique
  - Peut engendrer des confusions, par ex. « mange » pour « mangeable » comme pour « immangeable », « organ » pour « organe » comme pour « organisation »)
- Lemmatisation :
  - Remplacer un mot (par ex. « pensons ») par sa forme canonique (« penser »)
  - Basée sur l'analyse lexicale, utilise l'étiquetage grammatical
- Racinisation suffisante pour l'anglais, lemmatisation mieux adaptée au français

Représentation vectorielle des données textuelles

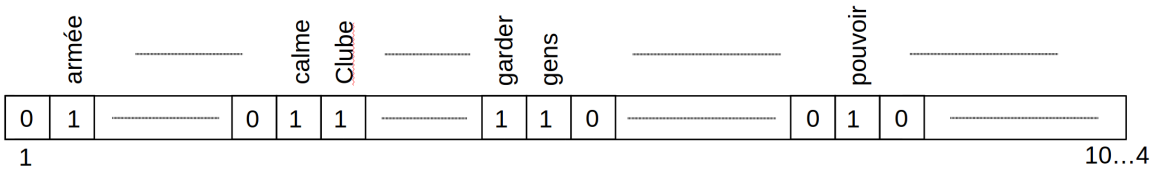
- Objectif : pouvoir manipuler des données textuelles avec les nombreux outils disponibles pour les espaces vectoriels
- Au préalable : lemmatisation ou racinisation, suppression des « mots ignorés » (*stop words*) : prépositions, conjonctions, articles, verbes auxiliaires…
  - L'ensemble des mots à ignorer peut dépendre de l'objectif de l'analyse !
  - Doit être appliquée seulement *après* l'extraction de locutions (ex. « chemin *de* fer »)
- Exemple :

  « Tôt le 25 avril 1974, au Portugal, des capitaines en rupture avec le système de Salazar se révoltent et prennent le pouvoir. La voix calme d'un mystérieux « Commandement du Mouvement des Forces armées » transmise par les radios de Lisbonne, Renascenta et Radio Clube donnant le signal de la révolte aux capitaines mutins, exhorte les gens à rester chez eux et à garder leur calme. »
  (`http://la-revolution-des-oeillets.ifrance.com/`)

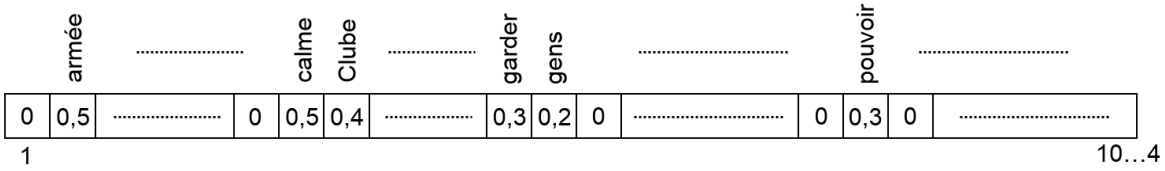
Représentation vectorielle des textes (2)

- Principe : affecter une dimension de l'espace à chaque lemme trouvé dans la base de documents

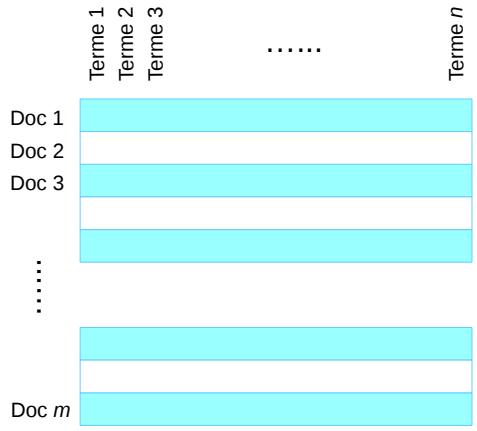
⇒ un texte (document de la base, requête) est représenté par un vecteur de (très) grande dimension, (très) creux, qui indique les mots présents dans le texte :



- Des pondérations (voir TF-IDF plus loin) sont souvent utilisées :



Représentation vectorielle des textes (3)



- Comparaison des vecteurs avec la distance cosinus : la norme du vecteur étant proportionnelle à la longueur du texte, mieux vaut mesurer l'angle entre vecteurs que la distance euclidienne
- Évolutions de la représentation vectorielle de base : pondération TF-IDF, sélection de mots (test du χ^2), analyse sémantique latente (LSA [3]), analyse sémantique explicite (ESA [4]), Word2Vec [12, 13], Sent2Vec [14]...

Pondérations *TF-IDF*

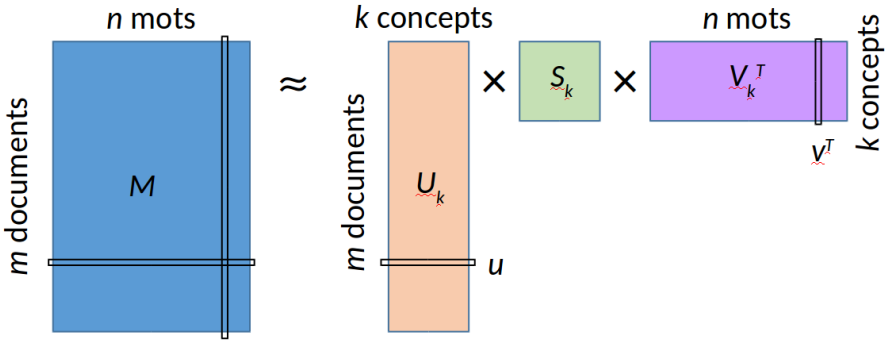
- Objectif : pondérer les termes suivant leur « importance » ; surtout pour la recherche d'informations
- Fréquence des termes (*term frequency*) : l'importance d'un terme pour un document est proportionnelle au nombre d'occurrences du terme dans le document, $tf_{ij} = \frac{n_{ij}}{\|d_j\|}$, n_{ij} étant le nombre d'occurrences du terme i dans le document j et $\|d_j\|$ la longueur du document d_j
- Inverse de la fréquence dans les documents (*inverse document frequency*) : l'importance d'un terme pour tous les documents est inversement proportionnelle au nombre de documents dans lequel il apparaît (les termes présents dans peu de documents sont plus discriminants que les termes présents dans beaucoup de documents), $idf_i = \log\left(\frac{n}{n_i}\right)$, n étant le nombre total de documents et n_i le nombre de documents contenant le terme i
  - Le logarithme permet d'éviter une sur-pondération des termes très rares, d'ailleurs en général on exclut les termes dont le nombre d'occurrences dans la collection de documents est inférieur à un seuil
  - De nombreuses variations existent, notamment pour *idf*
- Au total, la pondération du terme i dans le document j sera $tf_{ij} \cdot idf_i$

Analyse sémantique latente [3]

- Objectif : recherche de « concepts », correspondant à des corrélations entre termes, pour représenter les documents d'une collection
  - Suppression de « bruit » présent dans les données, comme l'utilisation accidentelle de mots inappropriés (ou dont un homonyme est beaucoup plus fréquent)
  - Remplacer les lemmes individuels par des concepts correspondant à des usages similaires (identifiés par LSA et non à travers une ontologie)
- Approche : décomposition en valeurs singulières (SVD) appliquée à la matrice documents-termes, suivie par une réduction de rang
  - Matrice documents-termes $\mathbf{M}$: 1 ligne par document, 1 colonne par terme
  - SVD : $\mathbf{M} = \mathbf{U} \cdot \mathbf{S} \cdot \mathbf{V}^T$, $\mathbf{U}$ et $\mathbf{V}$ étant des matrices orthogonales et $\mathbf{S}$ une matrice diagonale
  - En calculant $\mathbf{M}\mathbf{M}^T$ et $\mathbf{M}^T\mathbf{M}$ on constate que $\mathbf{S}$ contient les racines carrées des valeurs propres de $\mathbf{M}\mathbf{M}^T$ (ou $\mathbf{M}^T\mathbf{M}$), $\mathbf{U}$ contient les vecteurs propres de $\mathbf{M}\mathbf{M}^T$ et $\mathbf{V}$ les vecteurs propres de $\mathbf{M}^T\mathbf{M}$
  - Réduction de rang : on considère les k plus grandes valeurs propres et les vecteurs propres associés, alors $\mathbf{M}_k = \mathbf{U}_k \cdot \mathbf{S}_k \cdot \mathbf{V}_k^T$

Analyse sémantique latente (2)

- Décomposition matricielle pour l'analyse sémantique latente :



- Évolutions de l'idée : analyse sémantique latente probabiliste (*PLSA*, [5]), allocation de Dirichlet latente (*LDA*, [1])

Analyse sémantique explicite [4]

- Représentations basées sur la matrice documents-termes → relations entre documents (et entre termes) très dépendantes de l'ensemble de documents dont la matrice est issue → nombreux biais de modélisation
- Analyse sémantique explicite (*Explicit Semantic Analysis*, ESA [4]) :
  - Utiliser un corpus très grand (Wikipedia, en général) pour construire une matrice documents-termes

  ⇒ représentations générales et riches des termes, indépendantes d'ensembles spécifiques de documents applicatifs
  - Tout autre document est représenté par le centre de gravité de l'ensemble des mots qu'il contient (pondérations TF-IDF prises en compte)
  - Contrairement aux résultats de LSA, avec ESA les dimensions des vecteurs sont *interprétables* (dans Wikipedia chaque document décrit un concept)
- Possibilité d'utiliser un corpus multi-langues (par ex. Wikipedia) pour définir des représentations de documents indépendantes de la langue (exprimées comme des vecteurs dans l'espace des concepts Wikipedia) → *Cross-language* ESA
- Difficultés : biais présents dans Wikipedia (domaines, terminologie), disponibilité d'un corpus suffisant (et peu biaisé) dans une langue particulière

Word2Vec [12, 13]

- Objectif : obtenir, à partir d'un grand corpus de textes, des représentations vectorielles des mots qui incorporent des caractéristiques sémantiques et syntaxiques
- Modèle Skip-gram [12] : trouver des représentations permettant de prédire le mieux possible le contexte des mots
  - Étant donnée une séquence de mots $w_1, w_2, \ldots, w_T$, maximiser

$$\frac{1}{T}\sum_{t=1}^{T}\sum_{j=-k}^{j=k}\log p(w_{t+j}|w_t)$$

    k étant la largeur du contexte autour de chaque mot w_t
  - Plusieurs solutions permettent de réduire le coût de calcul
- Constats :
  - Avec cette représentation, les mots se regroupent par similarité de contexte
  - Une forme de « additivité » : la représentation la plus proche du résultat du calcul `vec(Madrid) - vec(Spain) + vec(France)` est `vec(Paris)`

Word2Vec pour des textes ?

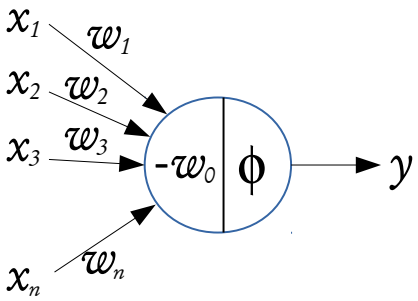
Représentation d'un texte :

1. Texte court ou petit ensemble de mots clés (< 10) : centre de gravité des vecteurs Word2Vec représentant ses mots, en général en supprimant les *stop words* et parfois en pondérant les mots (par ex. avec TF-IDF)
2. Texte long : si le nombre de mots est assez élevé alors le centre de gravité devient peu spécifique, d'autres méthodes sont préférables
   - Sent2Vec [14] : extension du contexte d'un mot à une phrase entière (ou un paragraphe, voire un document), vecteur phrase = centre de gravité des vecteurs représentant les mots *et* n-grammes de la phrase
   - Vecteurs de Fisher construits à partir des représentations Word2Vec des mots [7] :
     - À partir des vecteurs (ici Word2Vec) issus des textes de la base, un modèle de mélange (en général gaussien) est construit
     - Le vecteur de Fisher d'un ensemble de vecteurs (ici Word2Vec) est la concaténation des gradients de la log-vraisemblance de cet ensemble par rapport aux paramètres du modèle ($\rightarrow$ très grande dimension)

Plan du cours

Neurone « formel »

- Introduit dans [11] comme modèle simplifié d'un neurone réel

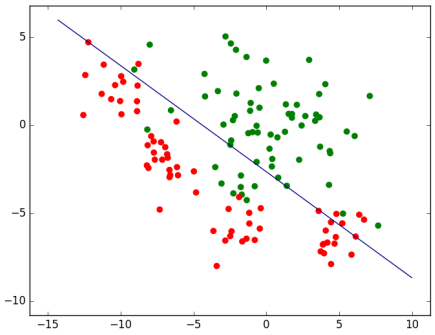


  - Entrées x_i ($i \in \{1, \ldots, n\}$)
  - Poids w_i ($i \in \{1, \ldots, n\}$), « seuil » w_0
  - Fonction d'activation ϕ ([11] : « marche » de Heaviside)
  - Sortie y

$$y = \phi\left(\sum_{i=1}^{n} w_i x_i - w_0\right)$$

Perceptron et Adaline

- Apprendre les poids pour obtenir une association entrées $\rightarrow$ sortie désirée
  - Perceptron (Rosenblatt, 1957) : $\mathbf{w}_{t+1} = \mathbf{w}_t - \eta\,\mathbf{x}$ ssi $\mathbf{x}$ mal classé
  - Adaline [18] : $\mathbf{w}_{t+1} = \mathbf{w}_t + \eta(y - \hat{y})\mathbf{x}$ (règle obtenue en minimisant l'erreur quadratique par descente de gradient)
- Le Perceptron et l'Adaline permettent de résoudre seulement des problèmes linéairement séparables :

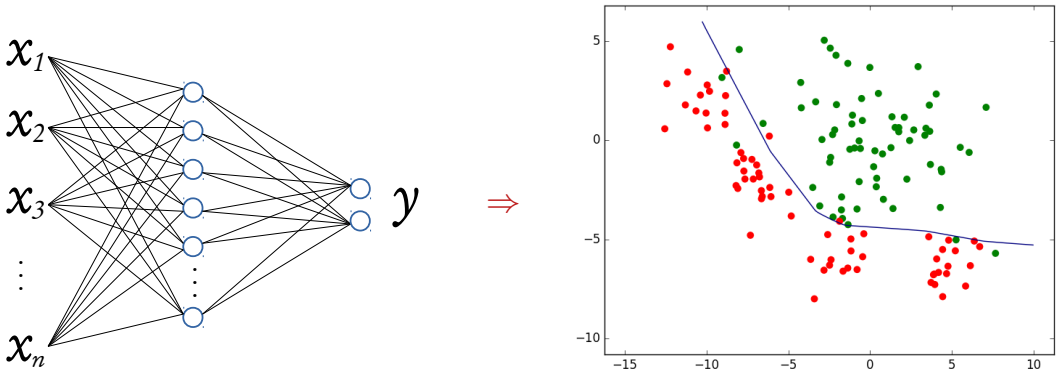


Perceptron multi-couches (PMC)

- Comment aller au-delà des séparations linéaires ?

$\rightarrow$ En ajoutant une (ou plusieurs) couche(s) cachée(s)

- L'activation des neurones se propage de l'entrée vers la sortie



Perceptron multi-couches (2)

- Résultat d'approximation universelle [2, 6] : quelle que soit une fonction continue sur des compacts de $\mathbb{R}^n$, une approximation aussi bonne que souhaité peut être obtenue avec un PMC ayant un nombre fini de neurones dans la couche cachée et une fonction d'activation « appropriée »

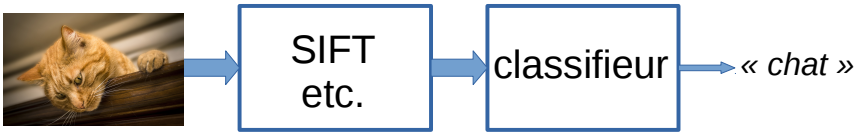
- Comment apprendre les poids dans ce cas ?

→ Toujours par descente de gradient, en calculant le gradient (les dérivées partielles) de l'erreur en sortie E par rapport aux poids de toutes les couches

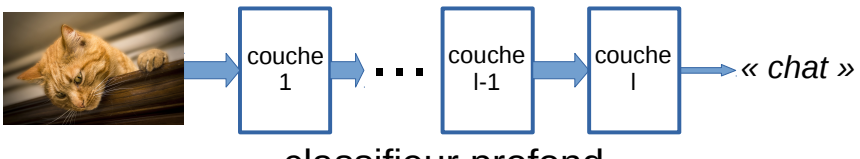
  - Dérivation de fonction composée : $E = E(y(\mathbf{w})) \Rightarrow \frac{\partial E}{\partial w_{ij}} = \frac{\partial E}{\partial y}\frac{\partial y}{\partial w_{ij}}$ ...
  - Passage d'une couche à la précédente : $\frac{\partial E}{\partial o_j} = \sum_{l \in \text{couche}+1} \frac{\partial E}{\partial o_l} \phi' w_{jl}$ (le calcul des dérivées partielles est fait de la sortie vers l'entrée)
  - Modification des poids par descente de gradient : $\Delta w_{ij} = -\eta \frac{\partial E}{\partial w_{ij}}$

Représentations apprises *vs* représentations non apprises

- Approche classique : représentations *handcrafted* + classifieur
  - Représentations de bas niveau sémantique
  - Représentations qui ne peuvent pas être optimisées | tâche, données



- Apprentissage profond : représentations apprises dans classifieur profond
  - Représentations dont le niveau sémantique progresse entre entrée et sortie
  - Représentations développées (donc optimisées) par apprentissage

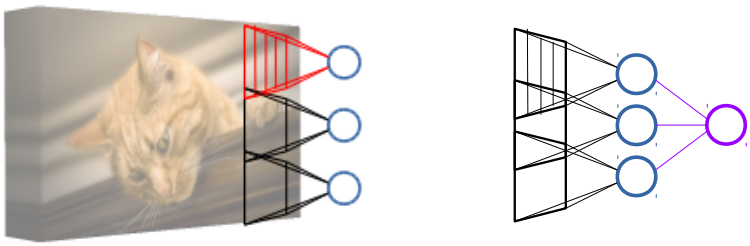


Comment apprendre des représentations ?

- PMC (connectivité complète entre couches successives) profond ⇒ très grand nombre de paramètres
  - Par ex., en entrée image de $1000 \times 1000 = 10^7$ pixels, suivie de 10 couches de 1000 neurones $\rightarrow 1000 \times 1000 \times 1000 + 10 \times 1000 \times 1000 \sim 10^9$ connexions
  - $\rightarrow$ Une bonne généralisation exige un très grand volume de données d'apprentissage
  - $\rightarrow$ Les invariances spécifiques au domaine doivent être apprises...

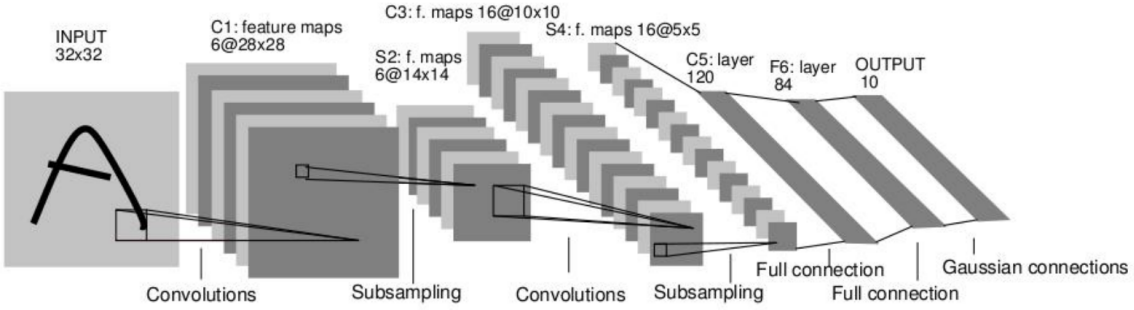
$\rightarrow$ Tirer profit des spécificités du domaine pour réduire fortement la connectivité

- Un « détecteur » devrait être invariant à la translation de l'objet dans l'image
- $\Rightarrow$ couche à connectivité locale et poids partagés entre neurones (convolution + nonlinéarité)
  suivie de couche d'agrégation spatiale (*pooling*)



Réseaux convolutionnels

- Introduits dès 1989 [8, 9]
  - LeNet 5 :



- Principe : succession de blocs [convolution + nonlinéarité + *pooling* (*subsampling*)] + couche(s) à interconnexion complète en sortie

Réseaux convolutionnels (2)

- Explosion du nombre de couches cachées (avec amélioration des performances) : 6 (LeNet5) $\rightarrow$ 7 (AlexNet 2012) $\rightarrow$ 18 (VGG, 2014) $\rightarrow$ 22 (GoogLeNet, 2014)) $\rightarrow$ 152 (ResNet, 2015)...
- Permise par de nombreuses évolutions, parmi lesquelles
  - Données : très grands volumes (ImageNet 14×10^6 images annotées, 20000 classes) ; *data augmentation*
  - Activation : ReLU plutôt que sigmoïde $\rightarrow$ éviter *vanishing gradient*
  - Convolution : fenêtres moins larges mais plusieurs couches ; plusieurs largeurs de fenêtre
  - *Pooling* : `max` plutôt que `avg`, recouvrement
  - Architecture : couches de classification intermédiaires $\rightarrow$ régularisation, éviter *vanishing gradient* ; court-circuits (*shortcuts*)
  - Régularisation : *dropout*, normalisation par *batch*

Réseaux profonds et représentation des données multimédia

- Réseau convolutionnel (CNN) profond qui apprend à résoudre un problème assez général sur un très grand corpus (par ex. reconnaissance sur ImageNet) $\rightarrow$ représentations discriminantes, plutôt générales et d'assez haut niveau sémantique développées dans les couches cachées proches de la sortie

$\Rightarrow$ Il est possible de se servir de ces représentations pour d'autres tâches

- Recherche d'images par similarité combinant aspect visuel et sémantique
- Classement de nouvelles images dans de nouvelles classes
- *Semantic segmentation* : segmentation avec reconnaissance

- Par ex., pour classement de nouvelles images dans de nouvelles classes :
  1. Remplacement de la couche de sortie par une autre (ou par SVM), apprentissage de cette seule couche sur les nouvelles classes
  2. Rétropropagation dans les couches cachées avec vitesse d'apprentissage faible (*fine-tuning*, seulement si la nouvelle base n'est pas trop petite)

Plan du cours

Recherche avec utilisation conjointe de plusieurs médias

- Approches classiques
  1. Fusion précoce : concaténation (avec pondérations) des représentations spécifiques des différents médias
  2. Fusion tardive : recherche mono-média pour chaque média, fusion des classements
- Difficultés : choix des pondérations et adaptation à la requête (au contenu), perte du lien entre médias, *curse of dimensionality* aggravé

→ développement et utilisation d'un espace commun de représentation

Plan du cours

Trans-média : passage d'un média à un autre

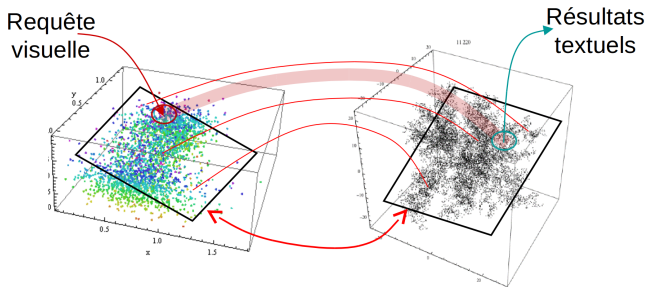
- Moteurs de recherche : recherche d'images à partir de textes (ou mots-clés), recherche de descriptions textuelles à partir d'images
- Autres applications
  - Classiques : annotation d'images, illustration de textes
  - Plus récentes (souvent *showcase*) : *image captioning* (génération de phrase descriptive à partir d'une image), *visual question answering* (réponse à une question à partir d'une image)
- Indispensable d'établir et exploiter des liens entre contenus des médias différents
- Avoir un espace commun de représentation permet à la fois de passer d'un média à un autre et d'utiliser conjointement plusieurs médias
- Voir [17] pour une synthèse de 2016

Qu'est-ce qu'un espace « commun » ?

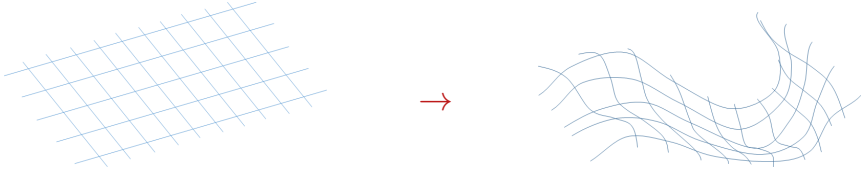
- Au départ, chaque média emploie son propre espace de représentation
  - Par ex. Sent2Vec ou vecteurs de Fisher à partir de Word2Vec pour les textes
  - Typiquement représentations issues de CNN profonds pour les images
- Mise au point d'un espace commun dans lequel un même vecteur (point) représente aussi bien un contenu textuel et un contenu visuel
  - Tous les documents textuels et toutes les images de la base sont représentées par des vecteurs dans cet espace
- Exemples d'utilisations possibles :
  1. Recherche trans-média : une requête (visuelle ou textuelle) est d'abord représentée dans cet espace, ensuite la recherche par similarité retourne des documents des deux médias
  2. Recherche pluri-médias : les composantes visuelles et textuelles du document requête sont combinées dans l'espace commun (par ex. centre de gravité), ensuite recherche par similarité
  3. Classement pluri-médias ou trans-média : un classifieur est appris dans l'espace commun, il permet alors de classer aussi bien des textes, des images et des documents mixtes

Comment construire l'espace commun ?

1. Analyse canonique des corrélations (CCA) : sous-espaces des espaces mono-média, maximisant la corrélation entre représentations visuelles et textuelles de documents mixtes



2. Analyse canonique à noyaux (KCCA) : extension dans laquelle les espaces corrélées ne sont plus linéaires

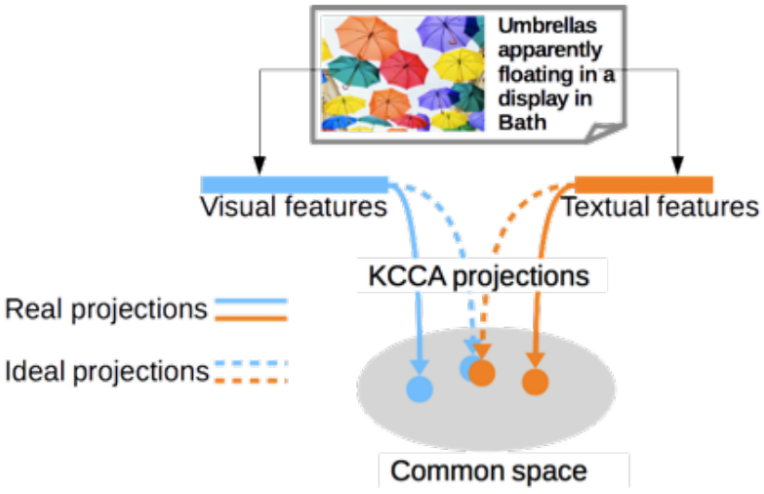


Comment construire l'espace commun ? (2)

- Espace commun $\subset \mathbb{R}^n$
  - Maximiser la corrélation entre représentations visuelles et textuelles de documents mixtes : (K)CCA, *Latent Dirichlet Allocation*...
  - Minimiser l'écart entre projections de représentations visuelles et textuelles : *Multimodal Deep Autoencoder*, *Deep* CCA...
  - Pour recherches trans-média, minimiser le rang des correspondants de l'autre média dans les réponses : DeViSE, *Deep Fragment*...
  - Lorsque les documents mixtes appartiennent à des classes, réduire l'inertie intra-classe et augmenter l'inertie inter-classes sur l'espace commun : CCA généralisé, *Cluster* CCA...
- Espace commun $\subset \{0\,,\,1\}^n$: méthodes de hachage sensible à la similarité $\rightarrow$ utilisation directe pour le passage à l'échelle de la recherche pluri-médias ou trans-média [17]

Biais de l'espace commun

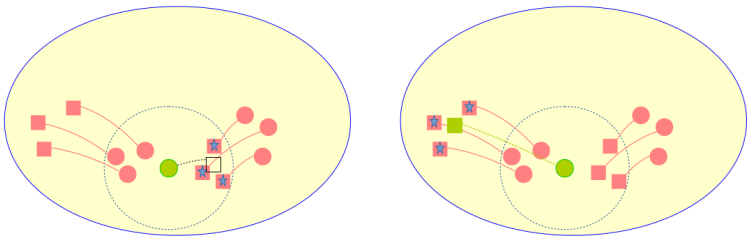
- Contenus visuels et textuels d'un document mixte ne sont pas parfaitement redondants $\Rightarrow$ projections distinctes sur l'espace commun



Biais de l'espace commun (2)

- Comment compenser ce biais ?
  - Modéliser localement l'impact du biais grâce aux documents mixtes disponibles [15, 16]
  - Avec une requête uni-média, ne pas retourner directement les *k*ppv qui sont des représentations de l'autre média (figure de gauche)

  → chercher plutôt les *k*ppv dans les représentations de même média, puis utiliser le modèle local pour retourner les correspondants dans l'autre média (figure de droite)



  - « Compléter » la représentation de données uni-média par une composante correspondant à l'autre média, obtenue par ce mécanisme

  → permet de faire de la classification trans-média, par ex. apprendre des classes sur des textes et les reconnaître dans des images [16]

Image captioning

- Objectif : génération de phrase décrivant le contenu d'une image (et non simplement de la liste des mots désignant les objets)

A little girl sitting on a bed with a teddy bear.

A group of people sitting on a boat in the water.

- Approche de [19] : se focaliser (mécanisme « attentionnel ») successivement sur des parties de l'image et générer le mot correspondant
  - Dans l'illustration ci-dessus ([19]), le côté droit de chaque exemple indique sur quelle partie de l'image porte l'attention (*soft*) lors de la génération du mot souligné
  - Mécanisme d'attention déterministe *soft* (haut) ou stochastique *hard* (bas) :

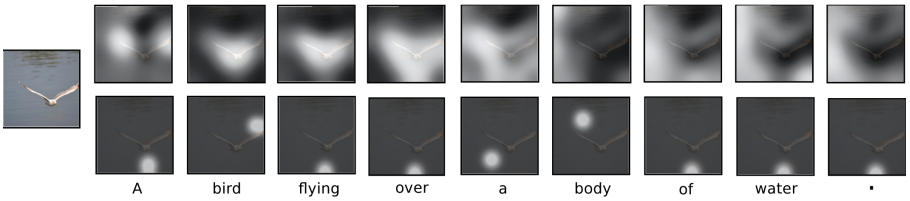
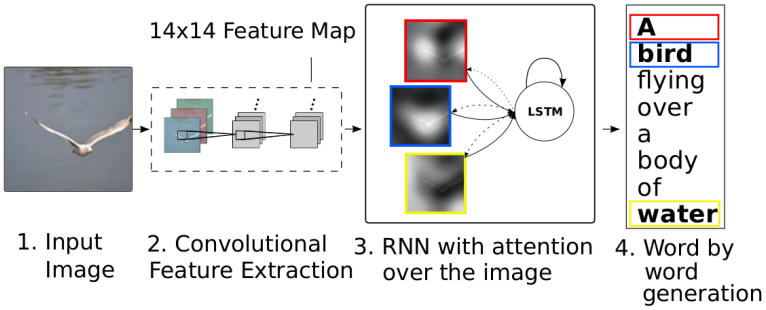


Image captioning (2)

- Architecture générale de [19] :



- Méthode de [19] :
  - Des couches intermédiaires d'un CNN pré-entraîné génèrent des représentations pour différentes régions recouvrantes de l'image
  - Une région sélectionnée par le mécanisme attentionnel sert d'entrée à un réseau récurrent (LSTM) qui génère le(s) mot(s) correspondant(s)
  - Le réseau récurrent modifie la région sur laquelle se porte l'attention dans l'étape suivante

Image captioning (3)

- Exemples d'erreurs ([19]) : visualiser l'attention permet de mieux les comprendre

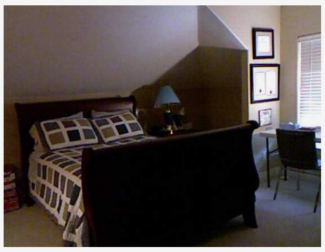
A woman holding a clock in her hand.

A woman is sitting at a table with a large pizza.

Visual question answering

- Objectif : répondre par un ou plusieurs mots à une question à partir du contenu d'une image



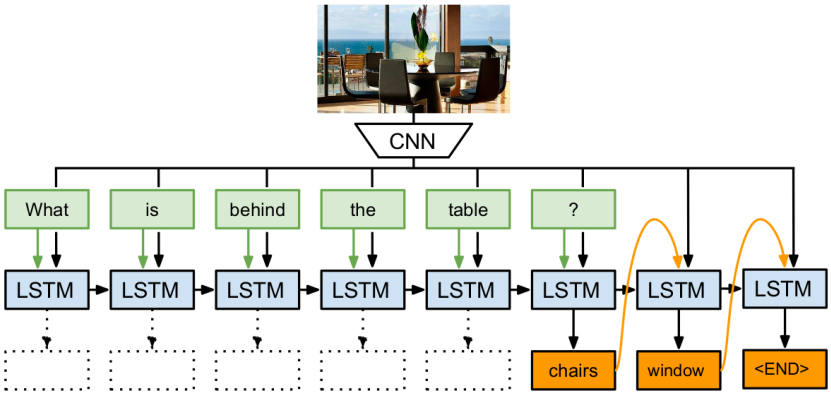


What is on the right side of the cabinet?	How many drawers are there?	What is the largest object?
Neural-Image-QA: bed	3	bed
Language only: bed	6	table

- Approche de [10] : CNN profond pour produire une représentation du contenu de l'image, réseau récurrent (LSTM) qui génère la séquence de mots de la réponse à partir de cette représentation et de la question

Visual question answering (2)

- Architecture générale de [10] :



- Méthode de [10] :
  - Types de questions : quelle couleur, combien, relations spatiales...
  - CNN profond pré-entraîné (sur ImageNet) → représentation de l'image dans la dernière couche cachée
  - Le réseau récurrent construit une représentation de la question à partir de la séquence de mots correspondante et la combine avec la représentation par CNN de l'image pour générer le mot (ou séquence de mots) de la réponse

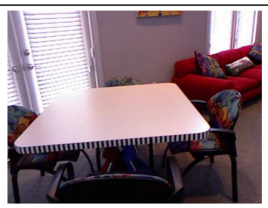
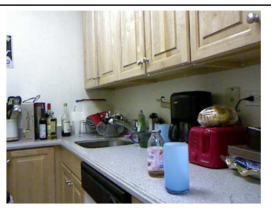
Visual question answering (3)

- Exemples de bonnes réponses composées de plusieurs mots ([10]) :



	What is on the refrigerator?	What is the colour of the comforter?	What objects are found on the bed?
Neural-Image-QA:	magnet, paper	blue, white	bed sheets, pillow
Language only:	magnet, paper	blue, green, red, yellow	doll, pillow

- Exemples d'erreurs ([10]) :

	How many chairs are there?	What is the object fixed on the window?	Which item is red in colour?
Neural-Image-QA:	1	curtain	remote control
Language only:	4	curtain	clock
Ground truth answers:	2	handle	toaster

Références I

D. M. Blei, A. Y. Ng, and M. I. Jordan.
Latent dirichlet allocation.
J. Mach. Learn. Res., 3 :993–1022, Mar. 2003.

G. Cybenko.
Approximations by superpositions of sigmoidal functions.
Mathematics of Control, Signals, and Systems, 2 :303–314, 1989.

S. C. Deerwester, S. T. Dumais, T. K. Landauer, G. W. Furnas, and R. A. Harshman.
Indexing by latent semantic analysis.
JASIS, 41(6) :391–407, 1990.

E. Gabrilovich and S. Markovitch.
Computing semantic relatedness using wikipedia-based explicit semantic analysis.
In *Proceedings of the 20th International Joint Conference on Artifical Intelligence*, IJCAI'07, pages 1606–1611, San Francisco, CA, USA, 2007. Morgan Kaufmann Publishers Inc.

Références II

T. Hofmann.
Probabilistic latent semantic indexing.
In *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '99, pages 50–57, New York, NY, USA, 1999. ACM.

K. Hornik.
Approximation capabilities of multilayer feedforward networks.
Neural Netw., 4(2) :251–257, Mar. 1991.

B. Klein, G. Lev, G. Sadeh, and L. Wolf.
Associating neural word embeddings with deep image representations using fisher vectors.
In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pages 4437–4446, 2015.

Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel.
Back-propagation applied to handwritten zip code recognition.
Neural Computation, 1(4) :541–551, 1989.

Références III

Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner.
Gradient-based learning applied to document recognition.
Proceedings of the IEEE, 86(11) :2278–2324, November 1998.

M. Malinowski, M. Rohrbach, and M. Fritz.
Ask your neurons : A neural-based approach to answering questions about images.
CoRR, abs/1505.01121, 2015.

W. S. McCulloch and W. Pitts.
A logical calculus of the ideas immanent in nervous activity.
Bulletin of Mathematical Biophysics, 5 :115–133, 1943.

T. Mikolov, K. Chen, G. Corrado, and J. Dean.
Efficient estimation of word representations in vector space.
CoRR, abs/1301.3781, 2013.

Références IV

T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean.

Distributed representations of words and phrases and their compositionality.

In C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 3111–3119. Curran Associates, Inc., 2013.

M. Pagliardini, P. Gupta, and M. Jaggi.

Unsupervised learning of sentence embeddings using compositional n-gram features.

CoRR, abs/1703.02507, 2017.

T. Q. N. Tran, H. L. Borgne, and M. Crucianu.

Aggregating image and text quantized correlated components.

In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 2046–2054, 2016.

T. Q. N. Tran, H. Le Borgne, and M. Crucianu.

Cross-modal classification by completing unimodal representations.

In *Proceedings of the 2016 ACM Workshop on Vision and Language Integration Meets Multimedia Fusion*, iV&L-MM '16, pages 17–25, New York, NY, USA, 2016. ACM.

Références V

K. Wang, Q. Yin, W. Wang, S. Wu, and L. Wang.

A comprehensive survey on cross-modal retrieval.

CoRR, abs/1607.06215, 2016.

B. Widrow and M. E. Hoff.

Adaptive switching circuits.

In *1960 IRE WESCON Convention Record, Part 4*, pages 96–104, New York, 1960. IRE.

K. Xu, J. Ba, R. Kiros, K. Cho, A. C. Courville, R. Salakhutdinov, R. S. Zemel, and Y. Bengio.

Show, attend and tell : Neural image caption generation with visual attention.

CoRR, abs/1502.03044, 2015.