Espace conceptuel: de l'homme à la machine

Adèle Mortier

9 janvier 2019

Introduction

Il semble séduisant de penser que les concepts ont peu à voir avec le raisonnement, en tant qu'il s'agit d'entités figées, culturellement acquises, et qui somme toute ne posent pas question. Cependant, ce sont bien les concepts qui structurent et rendent possible notre vie mentale ; et en particulier, nos raisonnements dépendront des concepts que nous nous sommes donnés.
Loin d'être figés, les concepts sont en fait apparus comme plus mouvants qu'on tendait à le penser : certains éléments semblent appartenir à plusieurs concepts, alors que d'autres, plus *borderline* semblent difficilement catégorisables. Ce sentiment va de pair avec l'idée que chaque concept pourrait être en quelque sorte "résumé" par un élément prototypique, qui du même coup dicterait l'appartenance ou non d'autres élément périphériques à ce concept.
Mais paradoxalement, cette incertitude inhérente vis-à-vis des limites des concepts n'empêche des structurations de haut niveau entre les concepts et leurs éléments constitutifs. Les éléments d'un même concept peuvent entretenir entre eux des relations, et ces schémas de relation peuvent être totalement ou partiellement retrouvés au travers de différents concepts. C'est ce que l'on appelle l'analogie entre concepts.
Dans ce mémoire, on se propose de comparer deux approches des concepts : une approche centrée sur l'humain – l'approche cognitiviste – et une approche plus proche de l'ingénierie – l'approche statistique. Nous verrons que ces deux approches entretiennent des liens quant à la définition de concept et d'analogie ; et nous proposerons des moyens de les confronter.

1 Les concepts en psychologie

Dans cette section nous aborderons comment les domaines de la psychologie et de la philosophie ont abordé la notion de concept au cours de l'histoire. Nous proposerons un approche formelle de ces idées.

1.1 L'approche définitoire

Une tradition philosophique née dans l'Antiquité définissait les concepts comme équivalent à un ensemble de conditions nécessaires et suffisantes (CNS). Ces conditions étaient d'appartenance à une classe – la classe dénotant ce concept :

$$C \equiv \{c_1 \dots c_k\} \tag{1}$$

$$C = \left\{ x \mid \bigwedge_{i \in [1,k]} c_i(x) \right\} \tag{2}$$

Cette vision des concepts paraissait satisfaisante dans la mesure ou elle répondait à un certain idéal philosophique :

— les concepts formaient des classes homogènes ;

— leurs frontières étaient bien claires ;

— ils représentaient ces valeurs fixes, objectives.

Cela dit, cette approche ne semble pas réellement coïncider avec la réalité des concepts au quotidien. Dans ce qui suit, nous allons passer en revue deux approches plus récentes qui relaxent différents paramètres du modèle définitoire.

1.2 Continuité des attributs : approche statistique

1.2.1 *Sharpness* du modèle définitoire

Le modèle définitoire pourrait également se résumer ainsi : une entité x est un vecteur binaire dont chaque composante correspond à une condition. Une composante quelconque de x vaut 1 si la condition qui lui correspond est vérifiée par x. Elle vaut 0 si la condition correspondante n'est pas vérifiée.

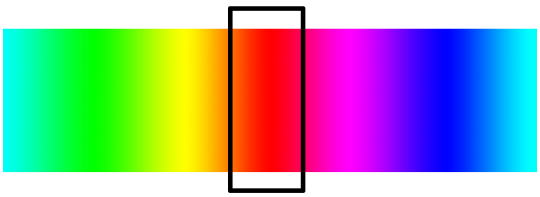
$$x = \begin{bmatrix} c_1(x) \\ \vdots \\ c_n(x) \end{bmatrix} \tag{3}$$

Dans cette implémentation du modèle définitoire, un concept est une fonction qui prend une entité x est vérifie que les composantes de x correspondant aux CNS du concept sont toutes à 1.

$$C \equiv \{c_{i_1} \dots c_{i_k}\} \tag{4}$$

$$C = \lambda x.\ \forall j \in [1,k],\ x[i_j] = 1 \tag{5}$$

Un problème majeur réside dans le caractère binaire des composantes de x. Dans le modèle définitoire, chaque entité vérifie ou ne vérifie pas une condition donnée ; il n'y a pas d'entre-deux. Pourtant, E. H. Rosch 1973 remarque que les attributs des entité du monde varient souvent continûment ; il en va ainsi de la couleur et de la forme d'un objet par exemple.



Certains éléments sont à “peu près carrés”, d'autres sont “un peu rouges” ; et on peut normalement dire si un élément est “plus carré” ou “plus rouge” qu'un autre. Tout cela conduit à penser que le vecteur binaire du modèle définitoire devrait être remplacé par un vecteur réel – ou, en d'autres termes, qu'il faudrait passer d'une logique bivalente à une logique polyvalente.

$$C \equiv \{c_{i_1} \dots c_{i_k}\} \tag{6}$$

$$C = \lambda x.\ \forall j \in [1,k],\ x[i_j] \geq \theta_j \tag{7}$$

Il découle de cette définition que les éléments appartenant à un concept peuvent être sujets à variation. Par exemple, les élément répondant au concept rouge ne sont pas forcément catégoriquement rouge : ils peuvent être rouge-orangé ou rosés. En revanche, ils ne seront pas bleus ou verts. Les entités, vues comme des vecteurs, sont donc étalées dans un espace relativement restreint qui correspond au concept. Or, si l'on admet que ces entités sont étalées sur un spectre, ce la signifie que certaines entités sont plus près du "gold standard" que d'autres. Il devient possible d'extraire une entité "caractéristique", qui se démarque des autres en tant qu'elle entre le mieux en adéquation avec le concept. E. H. Rosch 1973 parle à ce sujet de prototype ou d'exemplaire.

$$C \equiv \{c_{i_1} \dots c_{i_k}\} \tag{8}$$

$$ex(C) = argmax_x F(x[i_1], \dots x[i_k]) \tag{9}$$

$$= argmax_x \sum_{j=1}^{k} (x[i_j])^{\alpha} [\text{par exemple}] \tag{10}$$

Cet exemplaire pourrait faire office de nouvelle définition du concept. Un concept serait alors l'ensemble des entités suffisamment proches de l'exemplaire, suivant certaines dimensions (ou caractéristiques).

$$x \in C \iff dist(x, ex(C)) \leq \theta_C \tag{11}$$

1.2.2 Confirmation expérimentale

La notion de prototype dérivée de l'approche statistique a-t-elle une réalité cognitive ? E. H. Rosch 1973 montre que les prototypes sont effectivement appris plus facilement, attirent plus l'attention, et son réellement "attachés" à un nom de catégorie. Les catégories "artificielles" avec des prototypes centraux sont plus faciles à apprendre... du coup serait-il possible que l'on ai construit les catégories de notre langage autour des prototypes ? Ou que l'on ait choisi les prototypes au milieu des catégories que l'on s'est donné ? E. H. Rosch 1973 montre que les catégories générée par un élément "focal" (c'est-à-dire un élément que nous considérons comme prototypiques d'un concept, par exemple, la couleur rouge, la forme carrée) à l'aide d'une série de transformations, s'apprennent plus facilement et plus rapidement que les catégories ne contenant pas d'élément focal ou celle où l'élément focal n'est pas central (catégories générées à partir d'un élément non focal). De plus, l'élément focal au sein d'une catégorie est appris plus vite et donne lieu à moins d'erreurs, qu'il soit central ou non. Le test a été mené sur des populations archaïques ne verbalisant pas ou peu les couleurs et le formes (ces concepts et leurs sous-concepts étaient donc tout à fait nouveaux pour ces populations). Les participants ont été capables de faire du transfert de connaissances sur un training set d'éléments appartenant aux catégories apprises mais jamais rencontrés lors de l'apprentissage. Pour le cas des formes (plus contrôlé que l'expérience mettant en jeu les couleurs), ils ont aussi réussi à identifier les "bons" prototypes (les éléments focaux) pour chaque catégorie contenant ledit prototype (ie les catégories ou le prototype était central et celle où il ne l'était pas).

1.3 Flexibilité des CNS : un "air de famille"

Un problème dual au problème soulevé précédemment réside dans la définition des CNS d'un concept donné. Wittgenstein 1953 a par exemple remarqué qu'un concept comme celui de jeu était très difficile à définir par des CNS, car les différentes instances du jeu semblent vérifier des conditions pour le moins contradictoires :

— on pourrait dire que le jeu provoque un sentiment de plaisir (jeu video) ; pour autant certains jeux évoquent des situations d'affrontement, de compétition (football) ;

— on pourrait dire que le jeu nécessite de l'adresse (tennis de table) ou de la stratégie (go) ; pour autant certains jeux sont entièrement régis par le hasard (échelles et serpents).

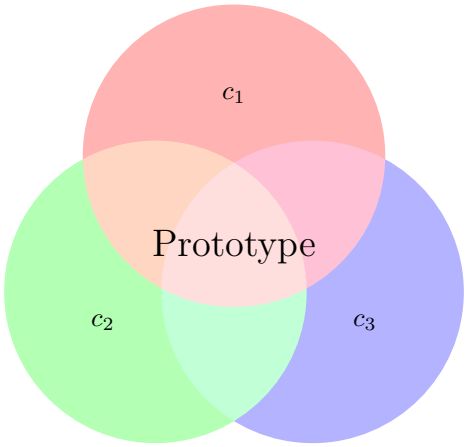
Les différents types de jeu ne semblent donc pas pouvoir vérifier *ensemble* le même ensemble de contraintes. Malgré cela, le concept de jeu nous semble très naturel et nous sommes capables de classifier des activités nouvelles comme des jeux sans trop de problèmes (par exemple, lorsque la mode du sudoku est arrivée en France, ou lorsqu' un nouveau jeu apparaît dans les cours d'école). Pigliucci 2003 fait la même remarque concernant le concept d'espèce qui fait débat en biologie : "it is impossible to define species, but it is certainly feasible to recognize them when you see them". Par ailleurs, E. Rosch et Mervis 1975 relève dans ses expériences qu'une évidente absence d'homogénéité des concepts (comme le concept de meuble, de véhicule, de légume...) ne choque pas fondamentalement les sujets et que ceux-ci continuent à croire, malgré l'évidence, que les concepts on des propriétés générales vérifiées par toutes leurs instances.

Pour résoudre ce problème, Wittgenstein introduit la notion d'"air de famille". Pour que deux entités soient liées entre elles et se réduisent à un même concept, il faut qu'elles partagent un certain nombre de traits en commun. Mais, contrairement au modèle définitoire, très strict, ces traits communs ne sont pas forcément tout à fait fixés ; il doivent juste être en nombre suffisant. Pour en revenir à notre interprétation vectorielle du modèle définitoire, la notion d'air de famille se traduit par une relaxation du quantificateur $\forall$. Par exemple, on pourrait penser à la formulation alternative suivante (très simpliste) :

$$C \equiv \{c_{i_1} \dots c_{i_k}\} \tag{12}$$

$$C = \lambda x.\ |\{j|x[i_j] = 1\}| > |\{j|x[i_j] = 0\}| \tag{13}$$

Les éléments d'un concepts seraient les éléments vérifiant une majorité d'attributs liés au concept. Ou, inversement, un concept serait défini par les attributs fréquemment représentés au sein des membres de ce concept, et peu fréquemment représenté en dehors du concept E. Rosch et Mervis 1975. Cette dernière définition tend à inverser le paradigme, en tant qu'elle dérive le concept des éléments qui lui appartiennent, et non plus l'inverse.
Mais que dire alors de la notion de prototype introduite précédemment ? Un prototype devrait vérifier un maximum de conditions liées au concept ; par maximum, on entend pas forcément toutes le conditions (ce serait le cas idéal), mais plus de conditions que n'importe quel autre entité appartenant au concept. Si l'on trace un diagramme de Venn correspondant aux attributs du concept, le prototype doit se trouver dans le secteur correspondant au maximum d'intersections.



E. Rosch et Mervis 1975 valide expérimentalement cette hypothèse et montre que les éléments considérés comme prototypiques d'un concept sont les éléments partageant le plus d'attributs avec les autres éléments du concept. Autrement dit, la notion de centralité adéquate serait une centralité de degré en théorie des graphes, sur un graphe dont les sommets sont les entités, et dont les arêtes pondérées définissent une relation de partage d'attributs entre entités :

$$c(x) = \sum_{i=1}^{k} c_i(x) \sum_{x' \neq x \in C} c_i(x') = \sum_{c|c(x)} |c \cap C| - 1 \tag{14}$$

Cela dit, cette notion est insuffisante à définir un critère de "bon prototype". En effet, les prototypes très généraux partageant beaucoup d'attributs avec les entités de leur concept, mais aussi beaucoup d'attributs avec des entités hors de leur concept, sont susceptibles d'être considérés comme "bons" également. C'est pourquoi E. Rosch et Mervis 1975 introduit un tradeoff entre aire de famille interne au concept et absence d'air de famille externe au cluster. Le prototype doit être général tout en étant distinctif. Ce tradeoff pourrait être modélisé par la notion de centralité suivante :

$$c(x) = \sum_{i=1}^{k} c_i(x) \sum_{x' \neq x} \mathbb{1}_{\{x' \in C\}} c_i(x') - \mathbb{1}_{\{x' \notin C\}} c_i(x') = \sum_{c|c(x)} |c \cap C| - |c \cap \complement C| \tag{15}$$

Autre problème : grue bleen (goodman). Les CNS temporelles peuvent donner des puzzles. Argument plus théorique qu'autre chose.

1.4 L'analogie en psychologie

1.4.1 Définition formelle

L'analogie est une opération d'ordre supérieur sur les concepts abstraits ; par exemple, le concept de procrastination ou d'excuse. Une analogie peut être vue comme une fonction d'une situation, problème, scénario vers un autre.
Le modèle n'analogie le plus simple date d'Aristote ; c'est ce que l'on appelle l'analogie à 4 termes, ou l'analogie proportionnelle :

$$a : b :: c : d \tag{16}$$

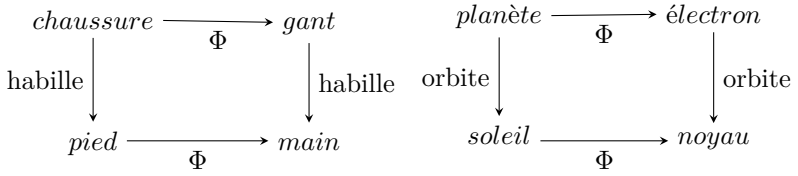
Ce qui se lit : a est à b ce que c est à d. Par exemple :

(i) “le pied est à la chaussure ce que la main est au gant” [analogie sémantique] ;
(ii) “le roi est à l’État ce que Dieu est au cosmos” [analogie sémantique] ;
(iii) “journaliste est à journal ce que guitariste est à guitare” [analogie sémantique/morphologique] ;
(iv) “dansait est à danser ce que jouait est à jouer” [analogie morphologique] ;

Dans ce cadre très simple, deux éléments (a et b) entretiennent la même relation que deux autres éléments (c et d). Par exemple (i), la relation est une relation de partie du corps à vêtement ; dans (ii), c’est une relation de pouvoir, dans (iii), c’est une relation de métier à objet de travail ; dans (iv), c’est une relation de temps. A partir de cet exemple, on peut définir une analogie comme une fonction Φ :

— de concepts à concepts ;
— de relations entre concepts à relations entre concepts ;

Avec la contrainte que, si une relation R lie a et b, alors $\Phi(R)$ lie $\Phi(a)$ et $\Phi(b)$ (conservation des flèches). Il est important de noter que cette définition s’étend à des problèmes de taille arbitraire, faisant intervenir un grand nombre d’objets qui interagissent entre eux.



On pourra alors définir :

— une analogie de structure comme une fonction Φ égale à l’identité sur l’espace des relations entre concepts ($\Phi(R) = R$) ;
— une analogie de surface comme une fonction Φ égale à l’identité sur l’espace des concepts (ou tout au moins, un sous-espace des concepts).

	même structure ↓	
même surface →	reporter un rdv chez le médecin	prendre un rdv chez le médecin
	reporter l’envoi d’un mail	

1.4.2 D’un point de vue cognitif

D’un point de vue plus pragmatique, deux problèmes analogues en structure auront le même schéma de résolution. En revanche, deux problèmes simplement analogues en surface ont des schémas de résolution différents. C’est donc l’analogie de structure qui est seule susceptible d’aider à la résolution de problèmes nouveaux, par transposition d’un problème déjà rencontré.
Cela dit, il a été montré que certaines transformations de problèmes sont plus faciles que d’autres. Dans un cadre expérimental, des sujet confronté Lorsque les gens sont devant deux problemes, ou lorsqu’on leur demande expressement de faire des analogies, ils ne voient ou ne genrent que des analogies de surface. c’est le relational gap. Cependant, Raynal, Clément et Sander 2018, ayant mis au point un nouveau paradigme de remémoration libre, montrent que les sujets sont capable de générer des analogies de structure lorsqu’ils sont libre de convoquer des domaines plus familiers. Les exemples utilisés s’appuyaient sur les concepts très courants d’excuse et de procrastination. La capacité à transposer un problème dépend donc en partie des champs sémantiques des objets :

— il est plus facile de transposer vers un domaine plus familier (e.g. lorsque l'on peut utiliser ses propres expériences) ;

— il est plus facile de transposer entre deux domaines proches (par exemple, entre l'orbite Terre/Soleil et l'orbite Lune/Terre, mais pas entre l'orbite Terre/Soleil et l'orbite Électron/Noyau).

cela est peut etre du au fait que les elements familiers et les relations qu'ils entretiennent entre eux activent de façon répétitive des schémas abstraits qui sont bien intériorisés. Il devient alors plus facile pour les sujets de les réutiliser et des les transposer. dans ce cas les analogies structurelles sont grlobablement plus fréquentes, et les individus faisant une majorité d'analogies structuerelles sont majoriteires. l'effet est donc très fort. Cela prouve que l'analogie structurelle est faite de façon assez automatique pourvu que l'on connaisse bien les concepts en présence (ils sont preactives dans la mémoire)

2 Machine learning

2.1 Embedding

L'"embedding" (ou plongement en français) consiste à traduire des données brutes en vecteurs, ou, de façon équivalente, en points dans un espace multidimensionnel. Les données brutes peuvent être des images, des échantillons sonores, des mots ou une combinaison de ces différentes sources ; on parle alors d'*embedding multimodal*. Dans ce mémoire, on se focalisera sur les données textuelles et visuelles, qui ont plus à voir avec la notion de concept que les données sonores, plus difficilement exploitables.
Le but principal d'un embedding est donc de peupler un espace par des vecteurs *signifiants*. En d'autres termes, on cherche à regrouper dans le même voisinage les vecteurs correspondant à des données proches sémantiquement : par exemple, des images de voiture avec des images de voitures, des termes liés à l'hygiène avec des termes liés à l'hygiène etc. Dans un embedding, la distance sémantique se traduit donc, entre autres, par une distance au sens topologique, (la norme précise à utiliser restant à déterminer). Un telle notion d'espace sémantique avait déjà été soulevée par E. Rosch et Mervis 1975 : "we can predict that items with the greatest family resemblance should fall in the center of the semantic space defined by proximity scaling of the items in a category [...] a semantics space in which the distance of items from the origin of the space is determined by their degree of family resemblance". Nous nous proposerons dans cette section de vérifier si les intuitions soulevées et testées par les psychologues trouvent un pendant dans les espaces d'embedding.

2.1.1 Embedding de mots

Le but est d'obtenir, par apprentissage non-supervisé, une représentation vectorielle des mots. Deux grands types de méthodes concurrentes permettent d'aboutir à des embeddings textuels comme le rappellent Pennington, Socher et Manning 2014 :

— les méthodes prédictives (skip-gram, CBOW, cf. Mikolov et al. 2013) basées sur des fenêtres de mots

— les méthodes basées sur le décompte de co-occurrences de mots et sur la factorisation de matrices de co-occurrence (LSA).

Ces deux approches reposent sur l'idée fondamentale que des mots en présence dans les mêmes contextes doivent être proches, et doivent donc avoir des traits sémantiques en commun.

2.1.2 Méthodes prédictives

La première implémentation de méthodes prédictives pour générer des embeddings est due à Mikolov et al. 2013. Les différents modèles sont regroupés sous le nom de Word2vec. Word2vec implémente la méthode du skip-gram et la méthode du sac de mots continu (CBOW) :

— la méthode du skip-gram vise à prédire le contexte d'un mot ;

— la méthode du sac de mots continus vise à prédire un mot étant donné son contexte.

Ces deux méthodes sont très simples et ne requièrent pas un grand nombre de couches : une couche cachée suffit.

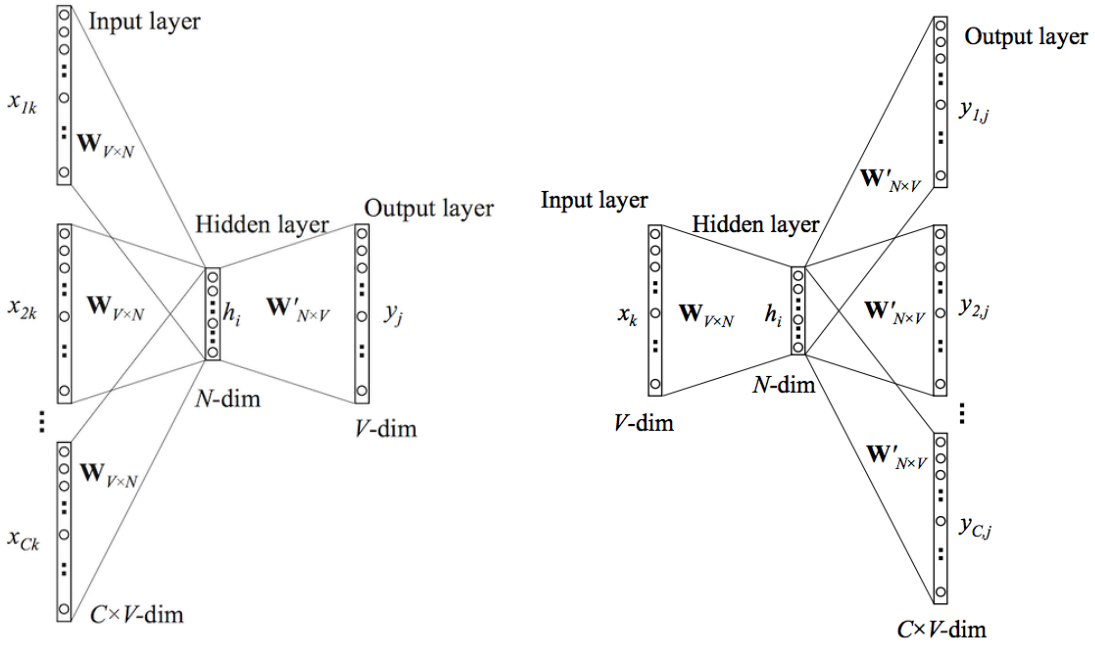


Figure 1 – Modèle CBOW (gauche) et skip-gram (droite)

On peut à première vue se demander quel est le lien entre la prédiction de mots et la création d'un espace sémantique. Le fait est que la prédiction de mots *nécessite* un espace sémantique ; autrement dit, les modèles prédictifs doivent apprendre cette représentation de façon implicite pour faire de meilleures prédictions. L'intérêt des méthodes prédictives pour l'embedding ne réside donc pas dans leur output, mais dans leur couche cachée. C'est dans cette couche, et à mesure que le réseau s'entraîne, que se crée l'espace sémantique. Pour mieux illustrer ce fait, nous allons détailler l'algorithme lié à la méthode CBOW. Nous utilisons pour cela Socher, Chaubard et Mundra 2016. L'idée derrière le CBOW est donc :

(i) d'encoder en *one-hot* les mots du contexte. Un encodage one-hot transforme un mot issu d'un lexique de taille N en un vecteur de taille N, possédant une seule coordonnée à 1 (celle que l'on a choisie pour correspondre au mot) et les autres coordonnées à 0 ;

(ii) **de plonger les vecteurs one-hot dans l'espace d'*embedding*, par changement de base (multiplication par une "matrice d'embedding")** ;

(iii) de calculer le barycentre (somme pondérée) de ces vecteurs – c'est l'application d'un principe simplifié de compositionnalité sémantique ;

(iv) de donner un score à ce barycentre (via une multiplication matricielle), puis de transformer ce score en vecteur de probabilités à l'aide d'un *softmax* – cette opération permet en fait un *remapping* dans l'espace de départ (espace "one-hot") ;

(v) de comparer le vecteur de probabilités avec le vecteur one-hot correspondant au mot central avéré (par exemple avec une fonction de coût de type entropie croisée) ;

(vi) **de rétropropager l'erreur, ce qui a notamment pour effet d'optimiser la matrice d'embedding**.

Les méthodes prédictives n'utilisent pas de façon optimale les statistiques globales des corpus, pourtant elles ont des bonnes performances sur les tâches d'analogie. les méthodes de décompte tirent très bien parti des statistiques des corpus, mais ont de moins bonnes performances sur les tâches d'analogie. Ce la dit, les deux méthodes ne sont pas si différentes, car elles exploitentn toutes deux de façon plus ou moins explicite des statistiques de co-occurrence.

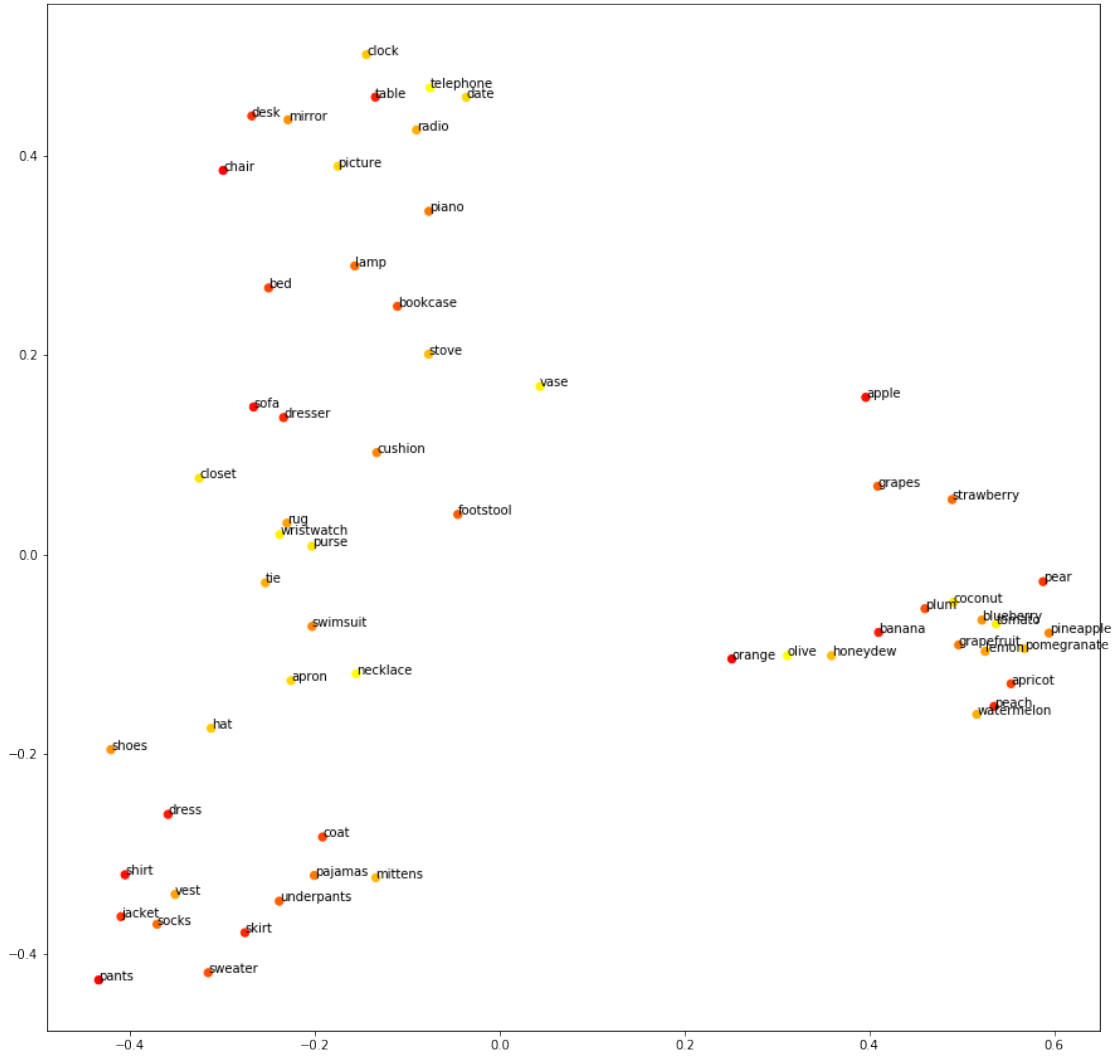


FIGURE 2 – *Embedding* de mots obtenu avec GloVe (6B, 300 dimensions) après PCA (2 dimensions), pour 3 des concepts testés par E. ROSCH et MERVIS 1975. Les points de couleur rouge sont des points censés être prototypiques ; les points jaunes sont des points *borderline*

2.1.3 Embedding d'images

2.1.4 Embedding multimodal

On peut reprocher aux modèles d'*embedding* n'exploitant qu'un seul type de donnée de ne pas avoir beaucoup de réalité cognitive. En effet, lorsque nous somme confrontés à des instances de concepts (sous forme textuelle, ou sous forme visuelle), notre mémoire active généralement des traces de types variés :

— lorsque le stimulus est écrit, une image correspondant au mot est souvent convoquée ;

— lorsque le stimulus est visuel, d'autres mots – plus ou moins généraux, plus ou moins adéquats – mais liés sémantiquement à l'image peuvent être convoqués.

Dans les deux cas, le stimulus peut également évoquer le souvenir d'une situation vécue (mémoire épisodique). L'*embedding* multimodal à pour but de simuler ces phénomènes d'association, dans le but d'augmenter la précision et la pertinence de la représentation. Frome et al. 2013 développent para exemple un modèle multimodal visuel et sémantique très performants sur des tâches souvent difficiles pour les réseaux de neurones :

— vocabulaire "ouvert" : le modèle est capable de catégoriser des images avec un vocabulaire qui peut facilement être agrandi, sans surcoût dans la phase d'entraînement ;

— *zero-shot learning* : capacité de catégoriser des images faisant partir d'un concept inédit ;

— pertinence dans l'erreur : propension des erreurs de catégorisation à ne pas être trop "absurdes" (i.e. éloignées du concept cible)

Le réseau se base sur deux embeddings pré-entraînés : un embedding visuel à l'état de l'art en 2013 (AlexNet), et un embedding textuel formé à l'aide du modèle skip-gram de Word2vec. L'*embedding* visuel est ensuite *fine-tuné* afin d'obtenir un *mapping* avec l'*embedding* textuel. L'architecture du réseau est donc assez proche des processus cognitifs opérants lors d'une catégorisation d'image par un humain : d'abord il s'agit de repérer les attribut importants l'image (embedding visuel), puis de faire correspondre ces attributs au mot le plus adéquat de notre lexique (embedding textuel).
En cas d'erreur, un bon mapping textuel/visuel garantit de tomber dans le voisinage du label correct, qui s'avère être un voisinage sémantique. Cette propriété du réseau explique se "pertinence dans l'erreur". Dans le cas d'images inédites, le réseau est capable de catégoriser en abstrayant ou en se montrant plus spécifique en fonction des images semblables déjà catégorisées. Là encore, cette capacité vient de la qualité de l'espace sémantique qui est à même de traduire en distances les relations de similarité entre labels.

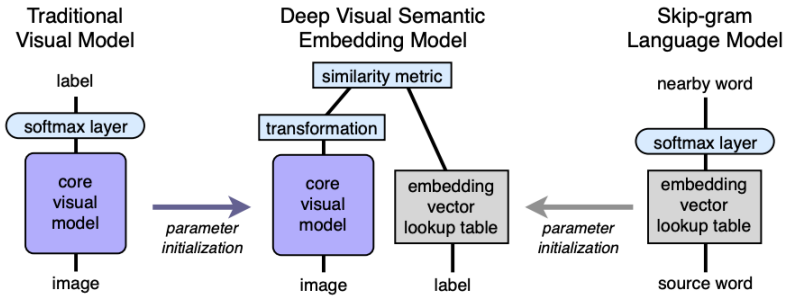


Figure 3 – Architecture du réseau développé par Frome et al. 2013 (au centre : réseau siamois final ; à gauche, composante visuelle pré-entraînée du réseau ; à droite, composante textuelle pré-entraînée du réseau)

2.2 Topologie des embeddings

2.2.1 Reduction de dimension

Les espaces d'embedding ont généralement une assez grande dimensionnalité (entre 100 et 1000). Cela ne rend pas très commode une analyse qualitative de ces espaces. De plus, une trop grande dimensionnalité peut rendre les calculs sur l'espace (similarité, clustering...) difficilement tractables. C'est pourquoi il existe des algorithmes de réduction de la dimensionnalité :

— l'analyse en composantes principales (PCA), qui consiste à repérer les dimensions très corrélées de l'espace de départ, et à les "fusionner" par des opérations linéaires (multiplication

par une matrice). Cet algorithme permet de conserver la majeure partie de l'information statistique présente dans l'espace de départ (conservation de la variance notamment). Cela fait de cet algorithme un candidat idéal pour le postprocessing.

— l'algorithme t-SNE, qui consiste à restituer au mieux dans un espace 2D ou 3D les relations de distance existant dans une espace de dimension bien supérieure. Autrement dit, les points éloignés dans l'espace multidimensionnel le seront toujours dans l'espace bidimensionnel, et de même les points proches dans l'espace multidimensionnel le seront toujours dans l'espace bidimensionnel. Contrairement à la PCA, cet algorithme est non-linéaire, il est donc surtout adapté à la visualisation.

2.2.2 Catégories

Un embedding vectoriel bien réalisé joue le rôle d'un espace sémantique. Il est donc naturel de rechercher dans cet espace des groupes de données qui sont susceptible de former des concepts. Cette tâche peut être menée à bien grâce à des algorithmes de *clustering*, qui ont pour rôle d'associer chaque donnée à un groupe, de sorte que les données proches dans l'espace en question se retrouvent dans le même groupe, et que les données éloignées forment des groupes différents. Un cluster, à l'image de l'espace duquel il est extrait, est donc multidimensionnel. Cette propriété rend en fait les clusters très proches des concepts psychologiques dont la cohésion est assurée par un "air de famille" : deux éléments d'un cluster on un faisceau d'attributs en commun (autrement dit leur projection sur certaines dimensions sont très proches), mais n'ont pas nécessairement *tous* leurs attributs en commun. En termes mathématiques, cela se traduit par le fait que les clusters de forment jamais d'hyperplan au sein de l'espace d'embedding. Dans un cluster, le "prototype" est souvent défini comme le barycentre du cluster (ou la donnée réelle la plus proche du barycentre théorique) :

$$proto(C) = argmin_{x \in C} d\left(x, \frac{1}{|C|} \sum_{x' \in C} x'\right) \tag{17}$$

Où d est une certaine distance (souvent, la norme L2 ou la similarité cosinus).

Par ailleurs, il est courant d'évaluer la qualité d'un clustering à l'aide de deux critère principaux :

— la cohésion intra-cluster (compacité), qui mesure à quel point les éléments d'un cluster sont concentrés autour du barycentre du cluster. Une mesure typique est la somme des carré des distances entre les points et le barycentre ;

— la séparabilité inter-cluster (isolation), qui mesure à quel point les clusters sont éloignés les uns des autres. Une mesure typique est la distance moyenne entre barycentres.

Ces mesures sont encore une fois cohérentes avec la vision de E. Rosch et Mervis 1975, selon laquelle les prototypes doivent être assez proches d'un maximum d'éléments du concept, et assez loin des autres prototypes.

Lsutering qui rend compte du notre propension à la convexité

2.2.3 Relations

Si les points de l'espace d'*embedding* désignent des entités, il apparaît naturel de définir les relations ("flèches") entre ces points comme des vecteurs. Par exemple, un vecteur pointant d'une entité "capitale" vers une entité "pays", pourra être vu, sous certaines conditions liées au contexte, comme la relation "est la capitale de". Il convient alors de se demander si ces relations

sont bien constantes dans l'espace ; autrement dit, si deux couples d'entités liées par la même relation sont aussi liés par le même vecteur de translation :

$$\exists ?V,\ \forall (x,y),\ xRy \iff y = x + V \tag{18}$$

2.3 Propositions d'expériences

2.3.1 Conservation des distances comme proxy de l'analogie

On demande à des participants de placer des photos sur une table.

2.3.2 Centralité comme proxy du prototype

On calcule un embedding, on fait du clustering, on regarde le centre des clusters et on compare avec le prototype "réel" donné par les gens. Tracer un tableau de l'évolution de la pensée sur les concepts en psychologie et en intelligence artificielle. Des propriétés binaires au densités sur les propriétés.

$$a : b :: c : d \iff d = argmax_{d \in V} sim(d, c - a + b) \tag{19}$$

$$a_i : b_i :: c : d \iff d = argmax_{d \in V} \mathbb{P}[d \in T]\mathbb{P} \tag{20}$$

$$sim(u,v) = cos(u,v) = \frac{u.v}{||u||||v||} \tag{21}$$

Références

Supports principaux

Mikolov, Tomas et al. (2013). « Efficient Estimation of Word Representations in Vector Space ». In : *CoRR* abs/1301.3781.

Pennington, Jeffrey, Richard Socher et Christopher D. Manning (2014). « Glove: Global vectors for word representation ». In : *In EMNLP*.

Rosch, Eleanor H. (mai 1973). « Natural categories ». In : *Cognitive Psychology* 4.3, p. 328-350. doi : 10.1016/0010-0285(73)90017-0. url : https://doi.org/10.1016/0010-0285(73)90017-0.

Rosch, Eleanor et Carolyn B Mervis (oct. 1975). « Family resemblances: Studies in the internal structure of categories ». In : *Cognitive Psychology* 7.4, p. 573-605. doi : 10.1016/0010-0285(75)90024-9. url : https://doi.org/10.1016/0010-0285(75)90024-9.

Socher, Richard, Francois Chaubard et Rohit Mundra (avr. 2016). *CS 224D: Deep Learning for NLP*. Note de cours, Université de Stanford.

Wittgenstein, Ludwig (1953). *Philosophical Investigations*. Translated by G.E.M. Anscombe. Oxford : Blackwell.

Supports secondaires

Drozd, Aleksandr, Anna Gladkova et Satoshi Matsuoka (2016). « Word Embeddings, Analogies, and Machine Learning: Beyond king - man + woman = queen ». In : *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*. Osaka, Japan : The COLING 2016 Organizing Committee, p. 3519-3530. url : http://aclweb.org/anthology/C16-1332.

Erk, Katrin (oct. 2012). « Vector Space Models of Word Meaning and Phrase Meaning: A Survey ». In : *Language and Linguistics Compass* 6.10, p. 635-653. doi : 10.1002/lnco.362. url : https://doi.org/10.1002/lnco.362.

Frome, Andrea et al. (2013). « DeViSE: A Deep Visual-Semantic Embedding Model ». In : *Advances in Neural Information Processing Systems 26*. Sous la dir. de C. J. C. Burges et al. Curran Associates, Inc., p. 2121-2129. url : http://papers.nips.cc/paper/5204-devise-a-deep-visual-semantic-embedding-model.pdf.

Lakoff, George et Mark Johnson (1980). *Metaphors we Live by*. Chicago : University of Chicago Press. isbn : 978-0-226-46800-6.

Pigliucci, Massimo (mai 2003). « Species as family resemblance concepts: The (dis-)solution of the species problem? » In : *BioEssays* 25.6, p. 596-602. doi : 10.1002/bies.10284. url : https://doi.org/10.1002/bies.10284.

Raynal, Lucas, Evelyne Clément et Emmanuel Sander (2018). « Structural similarity superiority in a free-recall reminding paradigm ». In : 40th Annual Meeting of the Cognitive Sciences Society. Madison (USA).