Apprentissage multimodal de représentation de mots à l'aide de contexte visuel

Eloi Zablocki, Benjamin Piwowarski, Laure Soulier, Patrick Gallinari

Apprentissage multimodal de représentation de mots à l'aide de contexte visuel

Éloi Zablocki, Benjamin Piwowarski, Laure Soulier, et Patrick Gallinari

Sorbonne Université, CNRS, LIP6, F-75005 Paris, France

Résumé

Représenter la sémantique d'un mot est un défi majeur pour pouvoir traiter automatiquement le langage. Jusqu'à présent, une grande partie des méthodes déterminent le sens d'un mot via ses contextes dans un corpus de texte. Plus récemment, certains auteurs se sont intéressés à l'apparence visuelle d'un objet pour améliorer la représentation sémantique du mot correspondant. Cependant, ces travaux ignorent l'environnement et le contexte visuel dans lequel l'objet apparaît. Dans cet article, nous proposons d'apprendre la représentation des mots en bénéficiant de la complémentarité des modalités texte et image par la prise en compte simultanée des contextes textuels et visuels des mots. Nous explorons plusieurs choix de modélisation de contexte visuel, et présentons une méthode jointe qui intègre le contexte visuel dans un modèle skip-gram multimodal. Enfin, l'apport de ces représentations dans des tâches d'analyse sémantiques est évaluée sur plusieurs jeux de données. Cet article est une traduction de [ZPSG18].

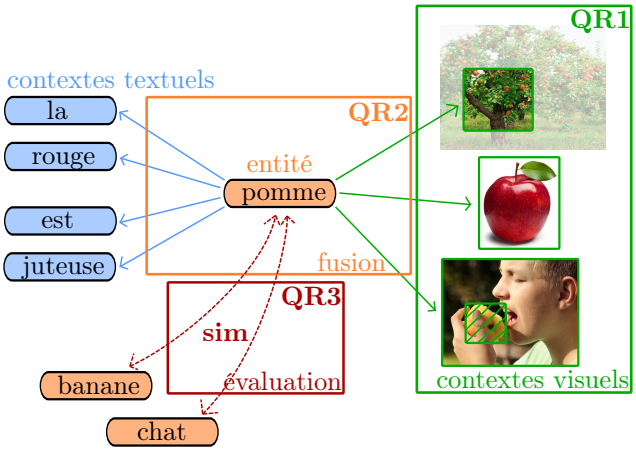

FIGURE 1 – Aperçu de notre approche et des Questions de Recherche (QR) sous-jacentes : QR1 concerne l'utilisation de contextes pour la partie visuelle du modèle, QR2 l'intégration de la partie visuelle avec la partie textuelle, et QR3 l'évaluation des représentations.

1 Introduction

La prise en compte de la sémantique d'un contenu linguistique est nécessaire pour aborder de nombreuses tâches, telles que la traduction automatique [BCB15], l'analyse de sentiments [MDP+11] ou encore le résumé de texte [RCW15]. La plupart des méthodes émergentes qui modélisent la sémantique reposent sur l'apprentissage de représentations de mots à partir de Modèles de Sémantique Distributionnels (MSD). Ceux-ci permettent de produire des représentations vectorielles pour les mots en fonction de leurs co-occurrences dans des corpus de textes. Ces modèles reposent sur l'*hypothèse distributionnelle* qui stipule que "les mots qui apparaissent dans des contextes similaires doivent avoir des sens similaires" [Har54].

Afin d'améliorer la qualité des représentations de mots, plusieurs éléments indiquent qu'il est essentiel de prendre en compte de l'information multimodale. En effet, il existe un biais linguistique entre ce qui est dit dans du texte et ce qui apparaît dans les images [GV13]. De plus, des études psychologiques ont montré que le sens des mots est ancré dans la perception [GK02, Bar08]. Ces observations soulignent les rôles complémentaires joués par l'image et le texte, et offrent de nouvelles perspectives pour des tâches de traitement automatique du langage, en complétant les informations textuelles par des informations visuelles. En outre, les avancées récentes et significatives en vision par ordinateur offrent des outils efficaces pour l'extraction d'information sémantique à partir d'images.

Dans ce contexte, des modèles multimodaux ont été proposés pour améliorer les représentations de mots en utilisant des techniques de fusion jointe [HK14, LPB15]

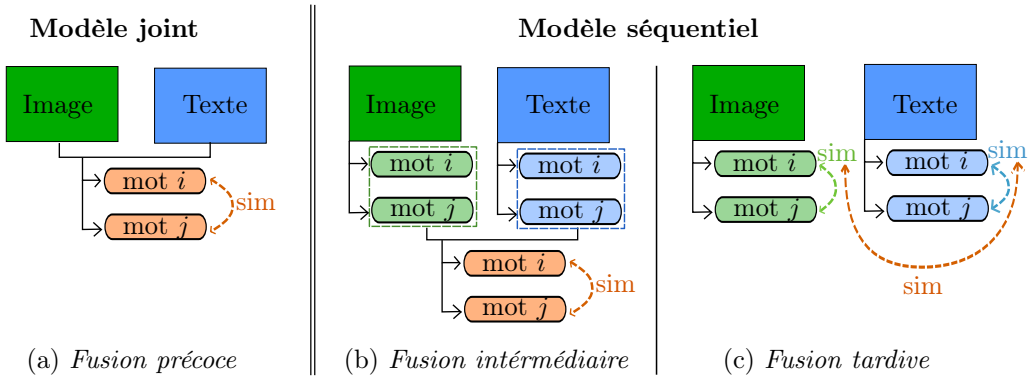

FIGURE 2 – Aperçu des différentes techniques de fusion : précoce, intermédiaire et tardive. Les rectangles aux coins arrondis représentent les représentations de mots. Le vert est relié aux images, le bleu au texte et l'orange au multimodal. "sim" est un exemple d'une tâche d'évaluation, ici *similarité de mots*.

ou séquentielle [KHKC14, BTB14]. Cependant, la plupart de ces travaux ignorent le *contexte visuel* des objets alors qu'il complète l'information sémantique présente dans les images. L'importance du contexte peut être illustrée dans un exemple simple (Figure 1). À partir d'une image d'une pomme sur un fond noir, on peut voir sa couleur, sa texture et sa forme. À partir de son contexte, par exemple si elle pousse sur un arbre, nous pouvons déduire la taille relative des pommes par rapport aux feuilles des arbres, et que les pommes sont des fruits qui poussent sur les arbres. S'il quelqu'un qui mange la pomme, on peut en déduire que les pommes sont comestibles et ainsi de suite. Cet exemple souligne l'importance de l'environnement visuel d'un objet et indique que son exploitation pourrait être bénéfique pour estimer la sémantique du mot associé.

Dans ce travail, nous présentons un modèle multimodal d'apprentissage de représentations de mots, en exploitant les contextes dans différentes modalités : texte et image. Nous proposons une extension du skip-gram où les représentations de mots sont apprises compte tenu des contextes textuels et visuels. Notre contribution porte sur trois aspects principaux :

— La formalisation du contexte visuel ainsi que sa modélisation à différents niveaux de granularité (Section 3.1) ;
— Un modèle multimodal axé sur le contexte pour apprendre conjointement des représentations à partir de textes et d'images non alignés (Section 3.2) ;
— Une analyse approfondie des résultats obtenus pour déterminer l'influence de la modalité visuelle sur les représentations multimodales apprises (sections 4 et 5).

2 Travaux connexes

2.1 Apprentissage de représentation de mots à partir de textes.

Les Modèles de Sémantique Distributionnels (MSDs) s'appuient implicitement ou explicitement sur la factorisation d'une matrice de co-occurrence de termes pour estimer les représentations de mots ; c'est le cas de GloVe [PSM14] ou *Word2Vec* [MSC+13] sur lequel nous nous basons. Dans ce dernier modèle, les mots sont prédits compte tenu de leur contexte (modèle CBOW - *"Continuous Bag Of Word"*) ou réciproquement (modèle Skip-Gram). Ces deux variantes de modèles ont pour objectif d'apprendre deux représentations pour un même mot (une pour le mot en tant qu'entité, une en tant que contexte) ; les représentations d'entités et de contextes présentant des caractéristiques complémentaires [NMCC16]. Plusieurs extensions ont été proposées au modèle de Skip-Gram, les plus proches de notre travail étant celles qui prennent en compte de l'information supplémentaire fournie par des graphes de connaissance [TGCL16, SCH17]. Comme dans notre approche, des données complémentaires au texte sont utilisées pour améliorer les représentations de mots, cependant, nous utilisons des images au lieu de bases de connaissance.

2.2 Apprentissage de représentation de mots à partir de textes et d'images.

Des études récentes motivent la construction de représentations de mots avec des données langagières et perceptuelles telles que des images. Plus précisément, chez les humains, des études psychologiques révèlent

que la signification des mots est reliée aux actions et à la perception sensorielle [GK02, Bar08]. De plus, Gordon et al. mettent en évidence la complémentarité du langage et des images [GV13]. En particulier, ils exhibent un *biais texte-image* qui montre que la fréquence avec laquelle les entités et actions sont mentionnées dans le langage ne correspond pas à leurs fréquences dans le monde réel ; ce qui est commun est généralement peu évoqué, ce qui est surprenant, et donc rare, davantage décrit. Ce biais systématique par rapport aux fréquences du monde réel motive l'exploitation d'information visuelle pour apprendre la représentation de mots, conduisant ainsi à des approches multimodales. Dans cet esprit, deux principales méthodes ont été proposées : les méthodes *séquentielles* et *jointes* (c.f. Figure 2).

Les *méthodes séquentielles* construisent séparément des représentations visuelles et textuelles, puis les combinent en utilisant différentes techniques : la *fusion intermédiaire* ou la *fusion tardive*. A partir de représentations apprises séparément dans chaque modalité, la fusion intermédiaire consiste à les combiner pour former un vecteur multimodal (Figure 2 (b)) ; plusieurs méthodes d'agrégation ont été considérées comme la concaténation [KB14], la décomposition en valeurs singulières [BBBT12], l'analyse canonique de corrélations [SL12], ou encore la projection inter-modale [CZM17]. Les techniques de fusion tardive (Figure 2 (c)) reposent également sur des représentations indépendantes pour chaque modalité. L'interaction multimodale des représentation est alors effectuée en aval de la tâche au travers, par exemple, d'une combinaison linéaire de scores de similarité respectivement obtenus à partir de données textuelles et visuelles [BTB14]. Dans la plupart des modèles séquentiels cités ci-dessus, les représentations textuelles sont des représentations pré-apprises (e.g., Glove [PSM14] ou Word2Vec [MSC+13]) et les représentations visuelles sont construites à partir de l'agrégation (par exemple la moyenne) des activations obtenues avec un réseau de neurones convolutifs pré-appris sur des images.

Alors que les fusions intermédiaires et tardives ne permettent pas de bénéficier durant l'apprentissage des interactions potentielles entre les différentes modalités, les *modèles joints* apprennent directement une représentation partagée à partir de données textuelles et visuelles (Figure 2 (a)). Cette idée est proche de la façon dont les humains apprennent le sens d'un mot, ancré dans son environnement, comme observé dans [GK02] et [Bar08]. Certains modèles joints nécessitent du texte et des images alignés. Par exemple, [RS13] utilisent une modélisation bayésienne basée sur l'hypothèse que le texte et les images associées sont générés en utilisant un ensemble partagé de sujets latents et [KVMP16] ancrent les représentations de mots dans le monde visuel en essayant de prédire la scène abstraite associée à une phrase donnée. Notre modèle suit une stratégie de fusion précoce, mais ne nécessite pas de texte et d'images alignés.

Plus proche de notre travail, des extensions de *Word2Vec* skip-gram ont été proposées. Par exemple, [HK14] basent leur modèle sur l'hypothèse que la fréquence d'apparition d'un concept concret dans un texte est corrélée avec la probabilité qu'un humain en fasse l'expérience. Les représentations de mots concrets sont entraînées à prédire les mots de contexte (comme dans le modèle classique de skip-gram) et les caractéristiques perceptuelles – des normes définies dans [MCSM05] qui décrivent les objets comme un ensemble de caractéristiques (couleur, usage, etc.). Ce travail a été complété plus tard par [LPB15] dont la méthode est conçue pour utiliser des images au lieu des normes qui sont construites à la main. Ils contraignent les représentations de mot concrets à être proches de leur représentations visuelles pré-apprises. Notre travail exploite davantage cette ligne de recherche, mais met l'accent sur l'exploitation du contexte visuel, ce qui n'a jamais été fait à notre connaissance.

2.3 Modélisation et utilisation de contextes visuels.

Plusieurs des travaux présentés ci-dessus utilisent la modalité visuelle pour contraindre la représentation textuelle à être proche de la représentation visuelle d'un objet. Une telle stratégie a deux inconvénients. Premièrement, il y a une asymétrie dans la prise en compte des modalités : le texte définit un contexte sémantique pour chaque mot - les mots qui l'entourent - tandis que les images sont utilisées pour avoir des informations visuelles sur l'objet lui-même. Deuxièmement, l'information contenue dans le contexte dans lequel les objets apparaissent, complémentaire du texte, n'est pas utilisé pour améliorer la représentation des mots. [BUBS12] proposent une approche de fusion intermédiaire où une intégration visuelle est construite en comptant le nombre de mots visuels dans les images. Cette contribution est une première tentative pour appliquer l'hypothèse distributionnelle aux images dans la mesure où les objets sémantiquement similaires ont tendance à apparaître dans des environnements similaires dans les images. Leurs expériences démontrent que l'apparence du contexte (entourant les objets) est plus informative pour la sémantique que l'apparence de

l'objet lui-même. Cette observation est renforcée par des expérimentations dans [RS13] qui propose un modèle d'allocation de Dirichlet latent, et [BTB14] qui utilise une technique de comptage pour apprendre des représentations de mots en tenant compte des contextes textuels et visuels. Après avoir construit des matrices de comptage de contexte pour le texte (nombre de co-occurrences avec contextes) et les images (en utilisant des représentations visuelles en sac de mots), ils effectuent une réduction de rang sur la concaténation des deux matrices. La fusion est considérée en second lieu, au niveau des caractéristiques ou de l'évaluation. Par rapport à notre travail, il n'y a pas d'apprentissage de représentation pour les contextes. De plus, les travaux ci-dessus utilisent des représentations d'images en sacs de mots et ne proposent pas d'apprendre conjointement des représentations à la fois visuelles et textuelles.

En plus de l'identification des entités dans leur contexte, de l'information spatiale est présente si les objets sont localisés dans l'image. [BTB14] propose d'utiliser cette information spatiale intrinsèque comme contexte en divisant l'image en boîtes 4x4 et en considérant les mots visuels séparément pour chaque région. Plus récemment, [CM18] apprennent des représentations de mots qui intègrent du sens-commun spatial, grâce à un modèle qui prédit la disposition spatiale d'une image à partir des objets et relations connus. Nous proposons d'utiliser la spatialité présente dans les images pour apprendre des représentations de mots, ce qui reste encore largement sous-exploré.

2.4 Hypothèse et questions de recherche

Contrairement à d'autres travaux sur l'apprentissage de représentations multimodales, nous proposons un modèle joint, ne nécessitant pas de texte et d'image alignés. Nous mettons l'accent sur l'exploitation du contexte visuel, ce qui n'a jamais été fait à notre connaissance. Dans ce travail, nous appelons *entité* un mot qui correspond à un objet.

Nous postulons que l'exploitation du contexte visuel peut améliorer la représentation apprise des mots. Cette hypothèse nous amène à considérer des images de scènes complexes, contenant de nombreux objets. En effet, les images d'un seul objet donnent très peu d'information sur sa taille, comment il peut interagir, etc. Au contraire, une image est beaucoup plus informative si elle montre un objet dans son environnement, interagissant ou non avec d'autres personnes ou objets. En conséquence, nous abordons les Questions de Recherche (QR) suivantes (illustrées dans la Figure 1) : (**QR1**) Dans les images, que peut-on utiliser pour apprendre des représentations sémantiques pour des objets ? En particulier, le contexte peut-il capter une partie de la sémantique d'un mot / d'une entité ? (**QR2**) Comment pouvons-nous fusionner des modules visuels et textuels pour former un MSD multimodal ? (**QR3**) Comment pouvons-nous évaluer la contribution de la modalité visuelle dans les représentations de mots multimodales ?

3 Apprentissage multimodal de représentations de mots, guidé par le contexte

Nous présentons ici un modèle multimodal exploitant contextes visuels et textuels pour satisfaire l'hypothèse distributionnelle. Pour ce faire, nous formalisons d'abord la notion de *contexte visuel* et proposons diverses instanciations (QR1).

Nous introduisons ensuite notre modèle joint multimodal basé sur le modèle skip-gram [MSC+13] (QR2). Les représentations de mots sont partagées par les modules textuels et visuels. En revanche, les contextes sont spécifiques aux modalités. Une force de notre modèle repose sur le fait qu'il ne nécessite pas de données alignées. Par ailleurs, nous supposons que les objets sont déjà détectés dans les images puisque ce n'est pas l'objet de l'article.

3.1 Apprentissage de représentation avec des contextes visuels

3.1.1 Formalisation

En nous basant sur l'algorithme skip-gram de Word2Vec qui considère les entités e (mots) et leurs contextes $\mathcal{C}_e = \{c_1,...,c_n\}$ (n mots dans une fenêtre centré sur l'entité), nous introduisons ci-dessous un modèle fondé sur l'hypothèse distributionnelle pour les images.

Dans notre cas, les contextes sont des contextes visuels que nous définirons ensuite. Le choix des éléments de contexte visuel $c \in \mathcal{C}_e$ pour une entité e n'a pas besoin de correspondre à une liste d'entités sémantiques [LG14]. Par exemple, les éléments de contexte visuels peuvent être des objets environnants, des caractéristiques de bas niveau telles que l'apparence visuelle, ou encore la localisation des objets environnants par rapport à l'entité considérée.

Dans cet esprit, nous définissons une fonction f_θ, paramétrée par θ (appris), telle que pour toute entité e et élément de contexte visuel $c \in \mathcal{C}_e$, $f_\theta(c)$ est un vecteur

de $\mathbb{R}^d$. Ces représentations sont ensuite utilisées dans le calcul de la fonction de coût :

$$\mathcal{L}_{\text{i}} = -\sum_{e\in\mathcal{D}}\sum_{c\in\mathcal{C}_e}\Big[\log\sigma(f_\theta(c)^\top t_e) + \sum_{c^-}\log\sigma(-f_\theta(c^-)^\top t_e)\Big] \quad (1)$$

où $\mathcal{D}$ est l'ensemble des entités, t_e est la représentation associée à l'entité e (appris), c^- est un contexte négatif, et σ est la fonction sigmoïde. Cette fonction objectif est très proche de la fonction de coût originale du skip-gram mais intègre l'apprentissage de f_θ qui partage des paramètres (θ) pour le calcul de chaque élément de contexte.

3.1.2 Choix de modélisation.

Étant donné une entité e, nous proposons différentes façons de modéliser une instance des contextes visuels $c \in \mathcal{C}_e$. Nous détaillons également la construction de f_θ, qui peut intégrer de l'information spatiale ou non.

Contexte de haut niveau (objets environnants). Une image I peut être représentée comme un sac d'objets : $I = \{o_1,o_2,...\}$. Étant donné une entité $e = o_i$ (pour un certain i) dans une image, nous définissons $\mathcal{C}_e = \{o_j, j \neq i\}$ comme l'ensemble de tous les autres objets apparaissant dans l'image. Chaque élément $c \in \mathcal{C}_e$ correspond à un contexte $c = o_j$ faisant référence à un objet environnant. Cette représentation simple donne des informations de haut niveau sur l'environnement dans lequel les objets apparaissent. Nous définissons $f_\theta(c) = V_c$ où $V \in \mathbb{R}^{M\times d}$ est une simple table de représentations pour les M objets, d est la dimension de l'espace de représentation, et V_c est la c-ième ligne de cette matrice.

Contexte de bas niveau (patchs d'image). L'ensemble $\mathcal{C}_e$ de tous les éléments de contexte visuel peut être vu comme des patchs de l'image complète où l'entité e est masquée par un rectangle noir. Nous l'appelons *contexte de bas niveau* car il utilise directement les valeurs des pixels de l'environnement des entités. L'utilisation d'un contexte de bas niveau est particulièrement intéressante pour deux raisons : les annotations utilisées dans la modélisation de haut-niveau peuvent être incomplètes ou biaisées [MZMG16], de plus, cela permet de tenir compte de l'apparence du contexte, ce qui apporte de l'information supplémentaire. Cependant, cela nécessite un modèle plus complexe, avec davantage de paramètres, et il est plus difficile d'extraire des informations significatives des valeurs de pixels. Nous suggérons deux possibilités pour sélectionner $c \in \mathcal{C}_e$: (1) l'instance c est l'image complète où l'entité est masquée par un rectangle noir ; (2) c est un petit patch d'image choisi au hasard autour de l'entité. En pratique, il existe plusieurs choix pour c tels que $c \in \mathcal{C}_e = \{c_1,c_2,...\}$. Dans les deux cas, le patch d'image c est représenté par la sortie d'un CNN, paramétré par θ_1, pour former un vecteur $u_c = \text{CNN}_{\theta_1}(c) \in \mathbb{R}^B$, où B correspond au nombre de cartes d'activations à la dernière couche du CNN (égal à 2048 dans nos expériences). Le vecteur de contexte visuel $f_\theta(c) = Nu_c$ est alors formé par la projection de u_c à la dimension d avec une matrice $N \in \mathbb{R}^{d\times B}$. Les paramètres à apprendre sont donc $\theta = \{\theta_1,N\}$.

Intégration d'information spatiale dans le contexte. Nous souhaitons utiliser de l'information spatiale supplémentaire lorsqu'un jeu de données fournit la localisation pour les entités (des boîtes englobantes ou des masques de segmentation). Par exemple, en regardant la position d'une tasse dans une image par rapport à une table ou à la main d'une personne, on peut en déduire que les tasses se posent sur des tables et qu'elles peuvent être tenues par des personnes. Nous souhaitons renforcer les représentations des contextes visuels présentés ci-dessus par des informations spatiales. Nous considérons deux méthodes pour modéliser ce que nous nommons la *spatialité visuelle*, sous forme d'un vecteur $s_{(e,c)}$ représentant les relations visuelles entre e et c, et deux modèles pour intégrer ce vecteur avec un élément de contexte visuel c par le biais $f^{sp}_\theta(c,s_{(e,c)}) \in \mathbb{R}^d$.

La première méthode considère les caractéristiques de bas niveau, et correspond à un vecteur spatial de quatre dimensions dont les composantes sont les positions relatives sur les axes x et y des deux boîtes englobantes de e et c (notée δ_x et δ_y), et le rapport de la largeur et de la hauteur entre les deux boîtes englobantes de e et c (δ_{largeur} et δ_{height}). La deuxième méthode forme un vecteur de caractéristiques de haut niveau, et correspond également à un vecteur spatial de quatre dimensions dont les composantes sont quatre fonctions indicatrices indiquant si le contexte c est en dessous, à côté, au-dessus, ou plus grand que l'entité e (1 si vrai, 0 sinon). D'après les conventions définies dans [LLKM16], le contexte est dit être "en dessous" si $|\delta_x| \leq \delta_y$, "au-dessus" si $|\delta_x| \leq -\delta_y$ et "à côté" sinon. Un contexte est dit être "plus grand" que son entité si $\delta_{\text{largeur}}\delta_{\text{hauteur}} \geq 1$.

Une fois le vecteur spatial $s_{(e,c)}$ construit, il est intégré à la représentation du contexte visuel $v_c = f_\theta(c)$,

pour former un contexte visuel enrichi d'information spatiale $v_c^{sp} = f_\theta^{sp}(c,s_{(e,c)})$ qui est utilisé dans les équations de skip-gram au lieu de $f_\theta(c)$. Encore une fois, deux variantes sont considérées : (1) une combinaison linéaire du contexte visuel v_c avec le vecteur spatial $s_{(e,v)}$, soit $f_\theta^{sp}(c,s_{(e,c)}) = M.(v_c \oplus s_{(e,c)})$ où $M \in \mathbb{R}^{d\times(d+4)}$ et $\oplus$ désigne l'opérateur de concaténation ; (2) une interaction bilinéaire $f_\theta^{sp}(c,s_{(e,c)}) = s_{(e,c)} M v_c$ où $M \in \mathbb{R}^{4\times d\times d}$. Ce modèle a plus de paramètres libres mais considère une interaction bilinéaire entre le vecteur spatial $s_{(e,c)}$ et le contexte visuel v_c, ce qui permet de capturer des interactions plus complexes.

3.2 Intégration dans un modèle multimodal

Nous présentons ici notre modèle d'apprentissage de représentations multimodales qui intègre le module visuel précédemment présenté avec le skip-gram textuel. L'idée principale est que les représentations de mots sont partagées entre les modalités, mais que le contexte est spécifique à chaque modalité. La contribution de chaque modalité est contrôlée par une combinaison linéaire (pondéré par α, déterminé par validation croisée) des fonctions objectifs spécifiques aux modalités, ce qui donne la fonction de coût globale suivante :

$$\mathcal{L}(T,U,\theta) = \mathcal{L}_{\text{t}}(T,U) + \alpha\mathcal{L}_{\text{i}}(T,\theta) \qquad (2)$$

où T (resp. U) désigne la table de recherche des entités (resp. contextes visuels) et $\mathcal{L}_{\text{t}}(T,U)$ est la fonction de coût de l'algorithme de *Word2Vec* [MSC+13].

Un point crucial est que ce modèle ne nécessite pas de textes et d'images alignés, ni de représentations supplémentaires pré-apprises sur des jeux de données externes. Il est en revanche nécessaire que les entités identifiées dans les images soient associées à un mot unique du vocabulaire. En outre, nous justifions l'utilisation d'un modèle joint car nous pensons qu'il est important que les représentations soient apprises aussi bien pour les entités que pour les contextes. En effet, comme les représentations d'entités sont affectées par les deux modalités, les représentations des contextes devraient être mises à jour par transitivité entre les modalités à travers les représentations partagées d'entités.

4 Protocole d'évaluation

Dans cette section, nous évaluons les représentations de mots sur différentes tâches. En particulier, nous mesurons la qualité des espaces de mots construits à partir de données visuelles (QR1) et multimodales (QR2).

4.1 Données

Nous utilisons une grande collection de textes en anglais, la base de données Wikipedia (`http://dumps.wikimedia.org/enwiki`), nettoyée et préparée avec le logiciel Gensim [ŘS]. La base totalise 4,2 millions d'articles et un vocabulaire de 2,1 millions de mots uniques. Pour les données visuelles, nous utilisons le jeu de données Visual Genome [KZG+17] car c'est une grande collection d'images (108k images) avec un grand nombre d'objets différents (4842 entités uniques avec plus de 10 occurrences) dans des scènes riches et complexes (31 instances d'objet par image en moyenne).

4.2 Scénarios et modèles de référence

Scénarios Pour évaluer les différentes composantes de notre modèle, nous proposons différents scénarios. En particulier, nous évaluons le modèle qui utilise les autres objets comme contextes visuels (noté **O**), le modèle qui utilise des patchs d'image (**P**) et des images complètes ($\mathbf{P}_{\text{full}}$).

Les modèles qui utilisent des informations spatiales sont également évalués et sont notés **Sp**(.,.,.) où le premier argument indique le type de contexte visuel (**O**, **P** ou $\mathbf{P}_{\text{full}}$), le second le contexte spatial (δ pour le bas-niveau ou $\mathbb{1}$ pour le haut-niveau), et le troisième la méthode d'intégration ($\oplus$ pour la concaténation et b pour l'intégration bilinéaire).

Toutes les combinaisons de ces modèles avec le modèle skip-gram purement textuel (noté **T**) sont entraînées et évaluées pour obtenir des représentations multimodales de mots, comme expliqué dans la section 3.2.

Modèles de référence Pour évaluer notre modèle d'apprentissage de représentations multimodales basé sur le contexte visuel (QR2), nous évaluons : 1) le modèle de texte skip-gram seul (noté **T**) et 2) un modèle séquentiel noté $\mathbf{O}\oplus\mathbf{T}$, où les représentations du modèle **T** sont concaténées avec des représentations obtenues de **O** et ensuite projetés dans un espace de plus petite dimension par analyse en composante principale. Cela sert de point de comparaison entre notre approche jointe et une approche séquentielle.

Nous nous comparons également à un modèle de référence (**L**), inspiré du modèle de [LPB15]. Les caractéristiques visuelles des objets eux-mêmes sont utilisées pour apprendre les représentations des mots contrairement au contexte visuel que nous utilisons dans notre modèle. Pour toute entité visuelle e, ils supposent qu'un vecteur visuel v_e représentant l'entité est donné. Lors de l'apprentissage du skip-gram textuel, la simila-

rité entre la représentation de l'entité et son apparence visuelle est maximisée via la fonction de coût suivante :

$$\mathcal{L}_{\text{object}} = \sum_{e \in \mathcal{D}} \sum_{v^-} \max(0, \gamma - \cos(t_e, v_e) + \cos(t_e, v^-))$$

où γ est une marge et v^- est l'apparence visuelle d'un objet "négatif" (aléatoire). Pour un objet e, v_e est maintenu fixe et des informations visuelles sont incorporées chaque fois que l'entité est rencontrée dans le texte. Nous notons ce modèle $\mathbf{L} + \mathbf{T}$ où $\mathbf{L}$ correspond à la partie visuelle et $\mathbf{T}$ à l'objectif de type skip-gram.

4.3 Tâches

Comme dans les travaux précédents [LPB15, CZM17], nous évaluons notre modèle sur trois tâches sémantiques différentes, à savoir la similarité des mots, la prédiction des caractéristiques et la prédiction du degré d'abstraction / de concrétude. Chaque tâche mesure un aspect différent de la qualité des représentations.

4.3.1 Similarité de mots

La similarité sémantique évalue le degré de similarité de paires de mots. Nous utilisons plusieurs jeux de données qui fournissent des scores pour les paires de mots : WordSim353 [FGM+02], MEN [BTB14], SimLex-999 [HRK15], SemSim et VisSim [SL14]. Une corrélation de Spearman est calculée entre la liste des scores de similarité donnée par le modèle (similarité cosinus entre les vecteurs multimodaux) et les scores fournis par les jeux de données ; plus la corrélation est élevée, meilleures sont les représentations sémantiques.

4.3.2 Prédiction de caractéristiques

Cette tâche consiste à prédire des caractéristiques (par exemple 'est_rouge', 'peut_voler') des objets à partir des représentations de mots. L'ensemble des données d'évaluation est un extrait du jeu de données McRae [MCSM05], qui comporte 43 caractéristiques regroupées en 9 catégories pour 417 entités. Nous utilisons le même protocole pour l'évaluation que celui proposé dans [CM16] : un classificateur SVM linéaire est appris et les scores de validation $k-$croisée sont indiqués ($k = 5$).

4.3.3 Estimation de degré de concrétude

Les normes USF [NMS04] donnent de degrés de concrétude pour 3260 mots anglais. Nous souhaitons savoir si le degré de concrétude d'un mot peut être déterminé à partir de sa représentation multimodale. En pratique, un SVM avec un noyau RBF est entraîné à prédire le degré de concrétude à partir des représentations de mots. Cette tâche est uniquement utilisée pour évaluer les représentations multimodales (Table 2) car les représentations purement visuelles couvrent un vocabulaire trop petit, ne comportant que des mots concrets.

4.4 Détails d'implémentation

Les expériences utilisent Tensorflow [AAB+16]. Les images sont agrandies à la taille 598×598 et sont traitées par un réseau Inception-V3 [SVI+16] pré-appris pour former un tenseur visuel spatial de forme $17 \times 17 \times 2048$ (avant le ReLU à la couche "Mixed_7c"). Un sous-tenseur de forme $1 \times 1 \times 2048$ correspond à l'activation d'une région de l'image originale. Nous utilisons 5 exemples négatifs par entité, et nos modèles sont appris par descente de gradient stochastique avec un pas d'apprentissage $l_r = 10^{-3}$ et des mini-batchs de taille 64. N et M sont régularisés avec une pénalisation L_2 pondéré respectivement par les scalaires λ et μ. Les valeurs des hyperparamètres ont été définies par validation croisée : $\lambda = 0.1$, $\mu = 0.1$, $\gamma = 0.5$, $\alpha = 0.2$.

5 Expériences et analyses

QR1 : Évaluation des représentations visuelles de mots. La table 1 rapporte les résultats des expériences pour la QR1 (partie 3.1) sur les tâches de similarité de mots et de prédiction de caractéristiques. Cela montre quel type d'information visuelle peut être utile pour former une représentation.

De façon générale, on observe que le contexte des objets est plus informatif que leur apparence visuelle. En effet, les résultats de cette tâche mettent en évidence que nos modèles dépassent généralement les modèles de référence. Par exemple, les résultats du modèle $\mathbf{P}_{\text{full}}$ sont en moyenne 29% plus élevés que ceux du modèle de référence $\mathbf{L}$. Cependant, dans la tâche de prédiction de caractéristique, les caractéristiques visuelles directes des objets (modèle $\mathbf{L}$) conviennent mieux pour les catégories qui décrivent visuellement les objets (par exemple *est_rouge* dans la catégorie 'couleur' ou *est_rond* dans la catégorie 'forme') mais pas pour les autres catégories non visuelles telles que 'encyclopédique', 'goût' et 'son'. Pour mesurer la complémentarité des informations provenant des objets et de leur environnement, nous avons également évalué un modèle d'ensemble ($\mathbf{L}+\mathbf{O}$), combinant le modèle de référence $\mathbf{L}$ et le modèle $\mathbf{O}$, où '+' indique la sommation

			VisSim	SemSim	Simlex	MEN	WordSim	Encyclopédique	Goût	Son	Taxonomique	Fonction	Tactile	Couleur	Forme	Mouvement
			Tâche de similarité					Tâche de prédiction de caractéristiques								
Référence		**L**	43	45	16	22	17	56	49	36	**76**	**56**	**17**	**41**	**60**	**58**
Nos modèles	Objets	**O**	43	54	31	64	27	48	46	35	62	48	03	21	43	36
	Patchs	**P**	28	35	17	35	22	30	51	23	48	37	04	24	38	30
		$\mathbf{P}_{\text{full}}$	35	42	19	43	28	30	48	30	46	35	06	23	35	27
	Spatial	$\mathbf{Sp}(\mathbf{O},\delta,\oplus)$	48	57	32	58	27	40	55	28	54	50	06	24	44	37
		$\mathbf{Sp}(\mathbf{O},\mathbb{1},\oplus)$	48	58	30	58	25	40	**60**	33	54	50	11	25	41	34
		$\mathbf{Sp}(\mathbf{O},\delta,b)$	46	56	**35**	54	28	37	57	27	50	50	15	24	38	32
		$\mathbf{Sp}(\mathbf{O},\mathbb{1},b)$	**51**	**61**	33	62	30	38	58	27	58	47	10	22	43	34
	Ensemble	$\mathbf{L}+\mathbf{O}$	45	57	33	**66**	**34**	**58**	52	**42**	74	**56**	02	27	53	53

TABLE 1 – Résultats pour la QR1. Les colonnes à gauche du tableau sont les corrélations de Spearman (en pourcentage) sur la tâche de similarité de mots (seulement les paires de mots correspondant à des entités visuelles sont évaluées). Les colonnes à droite sont les scores F1 (en pourcentage) pour la tâche de classification de caractéristiques (catégorisées comme proposé dans [CM16]).

des fonctions objectifs lorsque les représentations sont partagées. Il est intéressant de noter que la combinaison des contextes visuels et des informations directes ($\mathbf{L}+\mathbf{O}$) aboutit à un modèle ayant une très bonne performance moyenne, montrant la complémentarité des contextes visuels avec les représentations visuelles des entités.

Concernant l'utilisation d'information spatiale, les performances sont meilleures sur les jeux d'évaluation de similarité des mots (+9% d'amélioration en moyenne pour $\mathbf{Sp}(\mathbf{O},c,b)$ par rapport à $\mathbf{O}$) et la tâche de prédiction de caractéristiques (+20 %). Les caractéristiques spatiales de haut et de bas niveau donnent des résultats similaires. Cela renforce notre intuition que le contexte visuel, et plus particulièrement les informations spatiales, sont prometteurs pour l'apprentissage de la représentation des mots et la réduction du biais affectant les textes et les images.

Enfin, les contextes de haut niveau (dans $\mathbf{O}$) donnent de meilleurs scores (+31 %) que les contextes de bas niveau ($\mathbf{P}$ ou $\mathbf{P}_{\text{full}}$). L'utilisation de caractéristiques visuelles de bas niveau est un problème difficile. Cependant, elles sont prometteuses car elles sont peu coûteuses à collecter, ne nécessitent pas d'annotations de contexte et contiennent des informations riches si elles sont manipulées correctement. La difficulté réside dans le bruit naturel dans l'environnement des objets et le besoin de modules visuels qui extraient automatiquement des informations de haut niveau à partir des valeurs brutes des pixels.

QR2 / QR3 : Évaluation de notre modèle / analyse de l'apprentissage de la représentation multimodale multimodale axée sur le contexte. La table 2 rapporte les résultats de l'apprentissage multimodal (QR2 et QR3, partie 3.2) sur les tâches de similarité, de prédiction de caractéristiques et de concrétude. Les représentations sont initialisées avec les représentations pré-apprises obtenues à partir de modèle de référence textuel $\mathbf{T}$.

Les résultats montrent que tous les modèles multimodaux appris fournissent des représentations de meilleure qualité que le modèle textuel, sur toutes les tâches. Par exemple, $\mathbf{O}+\mathbf{T}$ montre une amélioration moyenne de 9 % sur $\mathbf{T}$. Ceci est en phase avec les conclusions des travaux connexes [HRK14]. En outre, un modèle conjoint ($\mathbf{O}+\mathbf{T}$) se compare favorablement à un modèle séquentiel ($\mathbf{O}\oplus\mathbf{T}$) construit à partir de représentations obtenues à partir de $\mathbf{O}$ et $\mathbf{T}$, avec une amélioration relative de 5%, montrant que les représentations calculées en utilisant plusieurs modalités à la fois sont bénéfiques. Comme précédemment, nous avons également évalué un modèle d'ensemble ($\mathbf{L}+\mathbf{O}+\mathbf{T}$) pour mesurer la complémentarité des caractéristiques visuelles dans le modèle multimodal. Encore une fois, nous remarquons une légère amélioration par rapport à $\mathbf{O}+\mathbf{T}$ et $\mathbf{L}+\mathbf{T}$. Finalement, cela ouvre des perspectives pour formaliser et exploiter l'information visuelle à la fois des entités et de leur contexte. Les résultats obtenus sont cohérents avec les conclusions de l'analyse QR1 : l'environnement visuel des entités est plus utile que leur apparence visuelle sur les tâches

			VisSim	SemSim	Simlex	MEN	WordSim	Encyclopedique	Goût	Son	Taxonomique	Fonction	Tactile	Couleur	Forme	Mouvement	Conc.
			Tâche de similarité					Tâche de prédiction de caractéristiques									
Réf.	Texte	$\mathbf{T}$	48	60	33	69	63	58	52	44	79	62	11	32	54	60	42.1
	Séquentiel	$\mathbf{O} \oplus \mathbf{T}$	49	62	33	71	64	63	55	40	72	59	12	35	54	58	43.7
	Joint	$\mathbf{L} + \mathbf{T}$	52	65	34	71	65	61	55	42	80	59	11	31	54	62	43.4
Nos modèles	Objets	$\mathbf{O} + \mathbf{T}$	53	66	35	**75**	**67**	62	55	46	**82**	61	13	33	55	61	42.9
	Patchs	$\mathbf{P} + \mathbf{T}$	53	65	35	72	**67**	60	56	49	**82**	60	12	32	55	61	43.1
		$\mathbf{P}_{\text{full}} + \mathbf{T}$	53	65	34	73	65	60	55	44	**82**	**63**	14	32	55	59	43.2
	Spatial	$\mathbf{Sp}(\mathbf{O},\delta,\oplus) + \mathbf{T}$	52	66	36	73	64	**64**	**59**	46	81	62	06	31	**57**	**63**	42.5
		$\mathbf{Sp}(\mathbf{O},\mathbb{1},\oplus) + \mathbf{T}$	54	66	35	72	64	62	56	**52**	80	61	13	**34**	**57**	58	43.7
		$\mathbf{Sp}(\mathbf{O},\delta,b) + \mathbf{T}$	54	**68**	**38**	73	66	63	56	48	81	60	13	32	56	**63**	42.5
		$\mathbf{Sp}(\mathbf{O},\mathbb{1},b) + \mathbf{T}$	**55**	67	34	**75**	64	61	58	46	80	**63**	**15**	**34**	**57**	62	**44.4**
	Ensemble	$\mathbf{L} + \mathbf{O} + \mathbf{T}$	54	66	35	**75**	65	63	55	50	**82**	60	10	33	55	59	43.9

TABLE 2 – Résultats expérimentaux pour la QR2 sur les jeux d'étalonnage de similarité de paires de mots, la tâche de prédiction de caractéristiques et la tâche d'estimation de degré de concrétude (Conc.) d'un mot. Les mesures de concrétude sont les coefficient de détermination (R^2) en pourcentage.

évaluées (amélioration de 3.2%)

Pour mieux comprendre les représentations apprises, nous cherchons à expliquer l'effet de la modalité visuelle sur la représentation multimodale des mots. Pour ce faire, nous estimons la corrélation entre le décalage mesuré sur la représentation et le degré de concrétude d'un mot. La mesure donne une corrélation de $\rho_{\text{Spearman}} = 0.33$, mettant en évidence que les mots visuels et concrets voient leurs représentations modifiées davantage que d'autres mots non visuels et abstraits.

6 Conclusion et perspectives

Dans ce travail, nous avons proposé une approche multimodale basée sur le contexte pour apprendre des représentations de mots. En accord avec les travaux connexes, nous avons observé la complémentarité des données visuelles et textuelles pour apprendre les représentations de mots. Plus important encore, nous avons montré que l'environnement visuel des objets et leur localisation relative sont cruciaux pour construire des représentations de mots – plus que (mais complémentaires à) l'apparence visuelle des objets eux-mêmes, telle qu'exploitée dans les travaux précédents.

Dans nos travaux futurs, nous explorerons l'utilisation de tâches en aval pour évaluer les représentations de mots multimodales, car cela peut donner une meilleure idée de la façon dont la partie visuelle du modèle contribue aux représentations d'apprentissage. De plus, nous étendrons notre travail pour apprendre des relations entre objets, en se basant sur des représentations multimodales et l'exploitation de bases de connaissance existantes.

Remerciements

Ce travail est partiellement soutenu par le projet européen CHIST-ERA MUSTER (ANR-15-CHR2-0005), FUI PULSAR (BPI France, Région Ile de France) et le Labex SMART (ANR-11-LABX-65). Nous remercions en outre Guillem Collell pour nous avoir fourni les vecteurs visuels pré-appris, nécessaires à l'évaluation du modèle de référence.

Références

[AAB+16] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, et al. Tensorflow : Large-scale machine learning on heterogeneous distributed systems. *arXiv preprint arXiv :1603.04467*, 2016.

[Bar08] Lawrence W. Barsalou. Grounded Cognition. *Annual Review of Psychology*, 59, 2008.

[BBBT12] Elia Bruni, Gemma Boleda, Marco Baroni, and Nam-Khanh Tran. Distributional semantics in technicolor. In *ACL 2012*, volume 1, 2012.

[BCB15] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *ICLR*, 2015.

[BTB14] Elia Bruni, Nam Khanh Tran, and Marco Baroni. Multimodal distributional semantics. *JAIR*, 2014.

[BUBS12] Elia Bruni, Jasper R. R. Uijlings, Marco Baroni, and Nicu Sebe. Distributional semantics with eyes : using image analysis to improve computational representations of word meaning. In *ACM*, 2012.

[CM16] Guillem Collell and Marie-Francine Moens. Is an Image Worth More than a Thousand Words ? On the Fine-Grain Semantic Differences between Visual and Linguistic Representations. In *Coling*, 2016.

[CM18] Guillem Collell and Marie-Francine Moens. Learning representations specialized in spatial knowledge : Leveraging language and vision. *TACL*, 6 :133–144, 2018.

[CZM17] Guillem Collell, Ted Zhang, and Marie-Francine Moens. Imagined visual representations as multimodal embeddings. In *AAAI*, 2017.

[FGM+02] Lev Finkelstein, Evgeniy Gabrilovich, Yossi Matias, Ehud Rivlin, Zach Solan, Gadi Wolfman, and Eytan Ruppin. Placing search in context : the concept revisited. *ACM*, 2002.

[GK02] Arthur M Glenberg and Michael P Kaschak. Grounding language in action. *Psychonomic bulletin & review*, 2002.

[GV13] Jonathan Gordon and Benjamin Van Durme. Reporting bias and knowledge acquisition. In *AKBC*, 2013.

[Har54] Zellig Harris. Distributional structure. *Word*, 1954.

[HK14] Felix Hill and Anna Korhonen. Learning abstract concept embeddings from multi-modal data : Since you probably can't see what I mean. In *EMNLP*, 2014.

[HRK14] Felix Hill, Roi Reichart, and Anna Korhonen. Multi-Modal Models for Concrete and Abstract Concept Meaning. *TACL*, 2014.

[HRK15] Felix Hill, Roi Reichart, and Anna Korhonen. Simlex-999 : Evaluating semantic models with (genuine) similarity estimation. *Computational Linguistics*, 2015.

[KB14] Douwe Kiela and Léon Bottou. Learning image embeddings using convolutional neural networks for improved multi-modal semantics. In *EMNLP*, 2014.

[KHKC14] Douwe Kiela, Felix Hill, Anna Korhonen, and Stephen Clark. Improving multi-modal representations using image dispersion : Why less is sometimes more. In *ACL*, 2014.

[KVMP16] Satwik Kottur, Ramakrishna Vedantam, José M. F. Moura, and Devi Parikh. Visualword2vec (vis-w2v) : Learning visually grounded word embeddings using abstract scenes. In *CVPR*, 2016.

[KZG+17] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Li Fei-Fei. Visual genome : Connecting language and vision using crowdsourced dense image annotations. *IJCV*, 2017.

[LG14] Omer Levy and Yoav Goldberg. Dependency-based word embeddings. In *ACL*, 2014.

[LLKM16] Oswaldo Ludwig, Xiao Liu, Parisa Kordjamshidi, and Marie-Francine Moens. Deep embedding for spatial role labeling. *CoRR*, abs/1603.08474, 2016.

[LPB15] Angeliki Lazaridou, Nghia The Pham, and Marco Baroni. Combining language and vision with a multimodal skip-gram model. In *NAACL*, 2015.

[MCSM05] Ken McRae, George S Cree, Mark S Seidenberg, and Chris McNorgan. Semantic feature production norms for a large set of living and nonliving things. *Behavior research methods*, 2005.

[MDP+11] Andrew L Maas, Raymond E Daly, Peter T Pham, Dan Huang, Andrew Y Ng, and Christopher Potts. Learning word vectors for sentiment analysis. In *ACL*, 2011.

[MSC+13] Tomas Mikolov, I. Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. Distributed Representations of Words and Phrases and their Compositionality. In *NIPS*, 2013.

[MZMG16] Ishan Misra, C. Lawrence Zitnick, Margaret Mitchell, and Ross Girshick. Seeing through the Human Reporting Bias : Visual Classifiers from Noisy Human-Centric Labels. *CVPR*, 2016.

[NMCC16] Eric T. Nalisnick, Bhaskar Mitra, Nick Craswell, and Rich Caruana. Improving document ranking with dual word embeddings. In *WWW*, 2016.

[NMS04] Douglas L Nelson, Cathy L McEvoy, and Thomas A Schreiber. The usf free association, rhyme, and word fragment norms. *BRMIC*, 2004.

[PSM14] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove : Global vectors for word representation. In *EMNLP*, 2014.

[RCW15] Alexander M. Rush, Sumit Chopra, and Jason Weston. A neural attention model for abstractive sentence summarization. In *EMNLP*, 2015.

[ŘS] Radim Řehůřek and Petr Sojka. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*.

[RS13] Stephen Roller and Sabine Schulte im Walde. A Multimodal LDA Model Integrating Textual, Cognitive and Visual Modalities. 2013.

[SCH17] Robert Speer, Joshua Chin, and Catherine Havasi. Conceptnet 5.5 : An open multilingual graph of general knowledge. In *AAAI*, 2017.

[SL12] Carina Silberer and Mirella Lapata. Grounded models of semantic representation. In *EMNLP*, 2012.

[SL14] Carina Silberer and Mirella Lapata. Learning grounded meaning representations with autoencoders. In *ACL*, 2014.

[SVI+16] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *CVPR*, 2016.

[TGCL16] Fei Tian, Bin Gao, Enhong Chen, and Tie-Yan Liu. Learning better word embedding by asymmetric low-rank projection of knowledge graph. *JCST*, 2016.

[ZPSG18] Eloi Zablocki, Benjamin Piwowarski, Laure Soulier, and Patrick Gallinari. Learning multi-modal word representation grounded in visual context. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, New Orleans, Louisiana, USA, February 2-7, 2018*, 2018.