Un modèle multimodal d'apprentissage de représentations de phrases qui préserve la sémantique visuelle

Patrick Bordes, Eloi Zablocki, Laure Soulier, Benjamin Piwowarski, Patrick Gallinari

Sorbonne Université, CNRS, Laboratoire d'informatique de Paris 6, LIP6, F-75005 Paris, France - prenom.nom@lip6.fr

RÉSUMÉ. *L'ancrage visuel est un domaine de recherche actif dont le but est d'enrichir les représentations vectorielles textuelles à l'aide d'informations visuelles. La plupart des travaux du domaine s'appuient sur des projections inter-modales qui alignent les éléments de deux modalités différentes. Cette technique s'avère problématique car elle impose que tous les objets aient une correspondance directe. Dans ce papier, nous proposons un modèle d'apprentissage de représentation de phrases qui transfère la structure d'un espace de représentation visuel à un espace textuel tout en préservant les deux espaces. Notre approche multimodale est générique dans la mesure où l'ancrage visuel est modélisé via une fonction objectif qui assure que (1) des phrases associées à un contenu visuel similaire doivent être proches dans l'espace textuel et que (2) les similarités entre éléments doivent être préservées entre les modalités. Nous démontrons la qualité de nos représentations de phrases sur des tâches de similarité de phrases et recherche inter-modale.*

ABSTRACT. *In this paper, we tackle visual grounding, an active field aiming to enrich textual representations with visual information, at the sentence level. Our model transfers the structure of a visual representation space to the textual space without using inter-modal projections, which are inherently problematic since modalities do not have a one-to-one correspondence. Our new multimodal approach can build upon any sentence representation model and can be implemented in a simple fashion by using objectives ensuring that (1) sentences associated with the same visual content should be close in the textual space and (2) similarities between related elements should be preserved across modalities. We demonstrate the quality of the learned representations on semantic relatedness, classification and cross-modal retrieval tasks.*

MOTS-CLÉS : *représentation de phrases, intermodalité, ancrage*

KEYWORDS: *sentence embeddings, multimodal, grounding*

1. Introduction

La représentation de données textuelles est un enjeu primordial pour de nombreuses tâches de Traitement Automatique du Language Naturel (Mikolov *et al.*, 2013 ; Peters *et al.*, 2018 ; Bojanowski *et al.*, 2017) ou de Recherche d'Information (Amer *et al.*, 2016 ; Lioma *et al.*, 2015 ; Manotumruksa *et al.*, 2016). Les premiers travaux dans ce domaine se sont appuyés sur une représentation pondérée des termes, comme la pondération TD-IDF (Salton et McGill, 1986) ; puis, les avancées en intelligence artificielle, et plus particulièrement en apprentissage profond, ont ouvert de nombreuses perspectives pour l'apprentissage de représentations textuelles. Les Modèles Distributionnels Sémantiques (Mikolov *et al.*, 2013 ; Peters *et al.*, 2018) sont des efforts récents pour parvenir à cet objectif. Ces modèles reposent sur l'*hypothèse de la sémantique distributionnelle* (Harris, 1954), qui stipule que des mots ayant des contextes similaires dans un corpus de texte ont en général un sens similaire. À un autre niveau de granularité, avoir des représentations de phrases de bonne qualité et simples d'utilisation est très utile pour tous les modèles qui représentent des phrases comme des vecteurs encodant leur sémantique – par exemple, les modèles de traduction automatique (Bahdanau *et al.*, 2014) ou de réponse automatique à des questions (Sagara et Hagiwara, 2014). La difficulté d'encoder la sémantique des phrases réside dans le problème de la composition des sens de sous-séquences de mots, et dans la prise en compte des relations implicites entre les mots (Norman, 1972).

L'apprentissage de représentations textuelles se fait à partir de co-occurrences statistiques entre les mots dans de larges corpus de textes. De nombreux travaux ont démontré que les modèles de Traitement Automatique du Langage dont l'apprentissage repose uniquement sur de l'information textuelle, peuvent conduire à des représentations biaisées et des prédictions erronées ; par exemple, des modèles purement textuels peuvent prédire que "le ciel est vert" (Baroni, 2016). D'autre part, des études de psychologie cognitive ont montré que la compréhension du langage est toujours *ancrée* dans la réalité physique et les expériences perceptuelles (Fincher-Kiefer, 2001). Pour permettre aux modèles de représentation de capturer cette information, une approche émergente est donc d'*ancrer* les modèles de langage dans la réalité physique à l'aide de l'information visuelle[1]. Dans le cas des mots, l'ancrage visuel produit de meilleures représentations, notamment pour les évaluations de similarité sémantique (Bruni *et al.*, 2014 ; Silberer et Lapata, 2014 ; Lazaridou *et al.*, 2015). Cependant, ce domaine de recherche demeure largement sous-exploré pour les représentations de phrases. A notre connaissance, le seul article s'intéressant à l'*ancrage visuel des phrases* est proposé par (Kiela *et al.*, 2018). Leur modèle est séquentiel : des vecteurs purement linguistiques sont concaténés à des vecteurs appris à l'aide d'information visuelle. Leur technique d'acquisition d'information visuelle repose sur une série de projections inter-modales, ce qui est le moyen usuel pour traiter l'information

1. Le mot *ancrage* peut avoir un autre sens dans la littérature. En effet, il peut aussi correspondre à d'autres tâches : identifier à quel région d'une image (ou quel segment d'une vidéo) correspond une phrase ou à un mot donné.

multimodale (Lazaridou *et al.*, 2015). Cependant, cette approche est limitée, puisque les modalités n'ont pas de correspondance directe entre leurs éléments (Gordon et Van Durme, 2013) : de nombreuses images peuvent correspondre à une même phrase, et de nombreuses phrases peuvent correspondre à la même image. Cette intuition est confirmée par le travail de (Collell et Moens, 2018), où les auteurs démontrent expérimentalement que les projections (et de manière plus générale les réseaux de neurones) ne sont pas suffisantes pour transférer la structure de voisinage entre modalités.

Pour traiter ce problème, nous proposons un modèle d'apprentissage de représentations de phrases multimodal, dans lequel nous exploitons l'espace visuel en *préservant* sa structure, c'est-à-dire en conservant les similarités entre les éléments correspondants des modalités textuelles et visuelles. Notre modèle joint est basé sur l'hypothèse que les contextes visuels et textuels d'une phrase *ancrent* son sens. Dans notre cas, le contexte textuel d'une phrase correspond aux phrases adjacentes dans un corpus de texte. Le contexte visuel correspond à des vidéos ainsi que les autres descriptions. Le contexte visuel est exploité en distinguant deux types d'information complémentaires : (1) l' *information de groupe*, correspondant à la connaissance implicite du fait que des phrases associées à la même vidéo font référence à la même réalité physique ; (2) l'*information perceptuelle*, qui construit des représentations à partir du contenu d'une vidéo. Nous nous posons les deux Questions de Recherches (QR) suivantes : • **QR1** : Quel est l'impact de l'ancrage visuel sur les représentations de phrases comparé aux modèles reposant exclusivement sur de l'information textuelle ? • **QR2** : Est-ce-que les informations de groupe et perceptuelle sont complémentaires, et comment des modèles les utilisant se comparent-ils avec d'autres reposant sur des projections inter-modales ?

Nos contributions sont les suivantes : nous proposons un modèle multimodal pour apprendre des représentations de phrases ancrées visuellement, tels (1) les informations textuelles et visuelles sont utilisées de manière jointe ; (2) la structure de l'espace visuel est transférée à l'espace textuel sans recourir à des projections inter-modales ; (3) nous utilisons des vidéos à la place des image à cause de leur aspect temporel (les phrases décrivent souvent des actions ancrées dans le temps) et car plusieurs images d'une vidéo constituent a priori un meilleur contexte visuel comparé à une image unique ; (4) nous conduisons une série d'expériences quantitatives et qualitatives sur de nombreuses taches de TALN afin de comparer notre modèle avec les autres méthodes d'ancrage visuel. Nous montrons par exemple que nos représentations de phrases permettent d'améliorer la performance en recherche d'images. Notre but premier est d'étudier l'impact de l'information visuelle sur les représentations de phrases ; notre objectif n'est donc pas de surpasser des modèles textuels pré-existants qui ont des performances état de l'art sur ces tâches. Nous montrons plutôt que n'importe quel modèle textuel peut bénéficier d'ancrage visuel de manière très simple : en ajoutant des fonctions objectifs et en ne changeant rien aux modèles originels. Ajouter ces fonctions objectifs permet d'améliorer les performances des représentations, comme nous le montrons avec des modèles standards tels que SkipThought (Kiros *et al.*, 2015) ou FastSent (Hill *et al.*, 2016).

2. État de l'art

Représentations de phrases : De nombreuses approches ont été proposées dans les dernières années afin de construire des représentations sémantiques de phrases. Tout d'abord, les techniques supervisées produisent des plongements de phrases qui dépendent de tâches spécifiques. Par exemple, dans le cas des tâches de classification, les représentations peuvent être apprises en utilisant des réseaux récurrents comme les LSTM (Hochreiter et Schmidhuber, 1997), des réseaux récursifs (Socher *et al.*, 2013a), de convolution (Kalchbrenner *et al.*, 2014), ou auto-attentifs (Lin *et al.*, 2017). D'autre part, les méthodes non supervisées produisent des représentations qui ne dépendent pas d'une tâche en particulier. Par exemple, SkipThought (Kiros *et al.*, 2015) et FastSent (Hill *et al.*, 2016) reposent sur l'hypothèse distributionelle (Harris, 1954) appliquée aux phrases : *des phrases qui apparaissent dans des contextes textuels similaires doivent avoir un sens similaire*. Dans le cas de SkipThought, la phrase est encodée à l'aide d'un réseau GRU, et deux décodeurs GRU sont entrainés à reconstruire les phrases adjacentes. Pour FastSent, la représentation d'une phrase est la somme de ses plongements de mots ; le but de l'apprentissage est de prédire les mots des phrases adjacentes à l'aide d'un objectif d'échantillonage négatif (*negative sampling*). Le présent article étend ces travaux en ajoutant une composante d'ancrage visuel.

Ancrage visuel : Pour définir le sens d'un texte, l'approche traditionnelle consiste à considérer le langage comme un système purement symbolique, fondés sur des mots et des règles syntactiques (Chomsky, 1980). Cependant, (Fincher-Kiefer, 2001 ; W. Barsalou, 1999) insistent sur l'intuition que le langage doit être ancré dans la réalité physique et l'expérience perceptuelle. L'importance de l'ancrage dans le monde réel est souligné par (Gordon et Van Durme, 2013), où un important biais est mis en évidence : la fréquence à laquelle les objets, les relations ou les événements sont mentionnés dans le langage naturel sont radicalement différents de leur fréquence d'apparition dans la réalité. En conséquence, utiliser des ressources visuelles en plus de ressources textuelles est prometteur pour acquérir du "bon sens" (Lin et Parikh, 2015 ; Yatskar *et al.*, 2016) et tenter de réduire le biais entre le texte et la réalité.

Les Modèles Sémantiques Distributionnels Multimodaux ont été développés pour résoudre le manque d'ancrage perceptuel des Modèles Sémantiques Distributionnels qui s'appuient uniquement sur de l'information textuelle comme (Mikolov *et al.*, 2013) ou (Pennington *et al.*, 2014). Deux approches peuvent être distinguées. Premièrement, l'approche *séquentielle* combine des représentations textuelles et visuelles qui ont été apprises séparément (Bruni *et al.*, 2014 ; Silberer et Lapata, 2014). Deuxièmement, la méthode *jointe* apprend une représentation multimodale commune à partir de plusieurs sources de manière simultanée ; l'avantage étant que l'information visuelle provenant des mots concrets est transférée à des mots plus abstraits qui n'ont pas forcément de données visuelles associées. Plus proche de notre travail, (Lazaridou *et al.*, 2015) présentent le modèle Skip-Gram Multimodal, où l'objectif Word2vec (Mikolov *et al.*, 2013) est optimisé conjointement à un objectif de classement, ce qui permet de rapprocher les vecteurs de mots correspondant à des objets concrets à

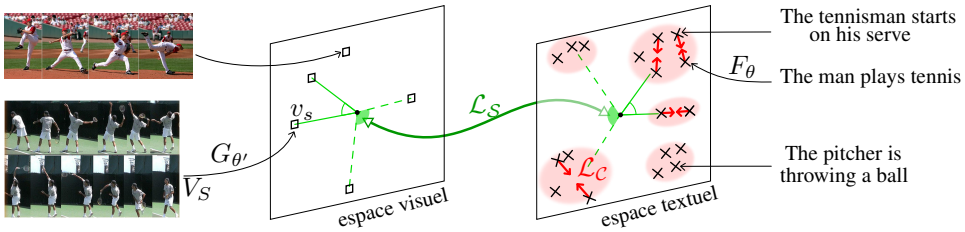

Figure 1. *Description du modèle. Les cercles rouges représentent les groupes visuels, composés des phrases associées à une vidéo donnée. Les flèches rouges décrivent le gradient de l'objectif de groupe, qui rassemble les phrases visuellement équivalentes — le terme contrastif de l'objectif $\mathcal{L}_C$ n'est pas représenté. La flèche et les angles verts illustrent l'objectif perceptuel, qui s'assure que les similarités soient corrélées entre les modalités. L'origine est au centre de chaque espace.*

ceux de leurs caractéristiques visuelles. De manière similaire, (Zablocki *et al.*, 2018) montrent que, dans les images, le contexte visuel des objets est tout aussi important que l'apparence visuelle des objets. Cependant, ces modèles apprennent des représentations de mots, alors que notre modèle s'attaque aux phrases. Récemment, (Kiela *et al.*, 2018) ont posé les bases des représentations de phrases multimodales en proposant une méthode séquentielle : des vecteurs appris a partir de données purement textuelles (Toronto BookCorpus) sont concaténés avec des vecteurs obtenus à partir d'un dataset de légendes d'images (MS COCO (Lin *et al.*, 2014)). Un encodeur LSTM de phrases est entraîné à prédire la représentation d'une image correspondante en prédisant si l'image correspond à la phrase et/ou en prédisant les autres descriptions de l'image.

La plupart de ces travaux s'appuient sur des projections inter-modales entre les espaces textuels et visuels afin d'intégrer l'information visuelle (Kiela *et al.*, 2018 ; Lazaridou *et al.*, 2015). Cependant, (Collell et Moens, 2018) démontrent que, lorsqu'une fonction inter-modale est apprise (de type projection ou réseaux de neurones), la projection de la modalité source ne ressemble pas à la modalité objectif, en termes de plus proches voisins. Cela suggère que les projections inter-modales ne sont peut-être pas appropriées pour incorporer la sémantique visuelle dans des représentations textuelles. Nous nous attaquons à ce problème en distinguant l'information de groupe et perceptuelle afin de préserver la structure de l'espace visuel.

3. Modèle joint pour apprendre des représentations de phrases

3.1. *Description du modèle*

Nous apprenons des représentations de phrases ancrées dans le réel en utilisant de manière jointe les contextes textuels et visuels d'une phrase. La ressource textuelle est un large corpus C_T de phrases ordonnées. La ressource visuelle est un corpus de vidéos C_V, où chaque vidéo a plusieurs descriptions qui la décrivent.

Une phrase S est représentée par $s = F_\theta(S)$ et sa vidéo correspondante V_S par $v_s = G_{\theta'}(V_S)$, où F (resp. G) est un encodeur de phrases (resp. de vidéos) paramétré par θ (resp. θ'). Nous proposons d'utiliser une approche où F_θ est appris en optimisant de manière jointe une fonction objectif textuelle $\mathcal{L}_\mathcal{T}(\theta)$ sur C_T et une fonction objectif visuelle $\mathcal{L}_\mathcal{V}(\theta, \theta')$ sur C_V. Jusqu'à maintenant, cette méthode a uniquement été appliquée aux mots, avec des bons résultats (Lazaridou *et al.*, 2015 ; Zablocki *et al.*, 2018). Notez que C_T and C_V peuvent être des jeux de données distincts mais que θ est partagé entre les deux objectifs. En d'autres termes, les représentations de phrases sont influencées par leurs contextes textuels et visuels. N'importe quel encodeur de phrase F_θ et objectif textuel $\mathcal{L}_\mathcal{T}$ peut être utilisé, comme SkipTought (Kiros *et al.*, 2015), FastSent (Hill *et al.*, 2016) ou encore QuickThought (Logeswaran et Lee, 2018). Dans cet article, nous nous focalisons sur SkipThought, et nous présentons des preuves que notre approche améliore également les performances de FastSent (cf. section 5.6).

Contrairement à la plupart des travaux d'ancrage visuel, qui se basent sur des projections inter-modales (Kiela *et al.*, 2018 ; Lazaridou *et al.*, 2015) pour intégrer de l'information visuelle, nous proposons un nouveau moyen de traiter l'information visuelle en préservant la structure de l'espace visuel (section 3.2).

3.2. *Traitement du contexte visuel*

Dans cette section, nous introduisons des hypothèses et les objectifs qui en découlent afin de modéliser la fonction objectif $\mathcal{L}_\mathcal{V}$.

Information de groupe. Sans même considérer le contenu des vidéos, le fait que des phrases décrivent ou non la même réalité est une source implicite d'information que nous nommons l'*information de groupe*. Deux phrases sont dites *visuellement équivalentes* (resp. *visuellement différentes*) si elles sont associées à la même vidéo (resp. à des vidéos différentes), i.e. si elles décrivent ou non la même réalité. Nous appelons *groupe (visuel)* un ensemble de phrases visuellement équivalentes. Utiliser l'information de groupe peut être utile pour améliorer la structure de l'espace textuel ; intuitivement, les représentations de phrases visuellement équivalentes (resp. différentes) doivent être proches (resp. éloignées) ce qui nous permet de définir l'hypothèse de groupe visuel : *Une phrase donnée doit être plus proche d'une phrase visuellement équivalente que d'une phrase visuellement différente.*

Nous traduisons cette hypothèse avec la contrainte $\cos(s, s^+) \leq \cos(s, s^-)$, où s^+ (resp. s^-) est une phrase visuellement équivalente (resp. différente) par rapport à s. En suivant l'exemple de (Karpathy et Li, 2015), nous utilisons un objectif de classement avec marge afin d'assurer que la distance entre les deux termes soit supérieure à une marge fixée m (cf. les éléments rouges dans la Figure 1) :

$$\mathcal{L}_\mathcal{C} = \sum_{(s,s^+,s^-)} \max(0, m - \cos(s, s^+) + \cos(s, s^-))$$

où s^+ (resp. s^-) est une représentation d'une phrase qui est visuellement équivalente (resp. différente) par rapport à s. Ces phrases sont échantillonnées de manière uniforme.

Information perceptuelle. L'hypothèse de groupe ignore la structure de l'espace visuel et n'utilise la modalité visuelle que comme un intermédiaire pour déterminer si deux phrases sont visuellement équivalentes ou différentes. De plus, l'objectif de classement $\mathcal{L}_\mathcal{C}$ sépare simplement les phrases visuellement différentes, ce qui peut être un problème si deux vidéos ont un contenu proche. Par exemple, les vidéos de basketball et de tennis dans la Figure 1, bien que différentes, ont en commun le fait d'être des vidéos de sport ; de fait, les représentations de leurs phrases respectives doivent être rapprochées dans l'espace textuel.

Afin de régler ce problème, nous tenons compte de la structure de l'espace visuel et nous utilisons le contenu des vidéos ; nous proposons également une nouvelle approche qui évite les projections inter-modales. L'intuition est que la structure de l'espace textuel doit être calquée sur celle de l'espace visuel afin d'extraire la sémantique visuelle. Nous choisissons de préserver, entre les modalités, les *similarités* entre éléments correspondants (cf les éléments verts dans la Figure 1). Nous définissons ainsi l'hypothèse perceptuelle : *La similarité entre deux phrases dans l'espace textuel doit être corrélée avec la similarité entre les vidéos correspondantes dans l'espace visuel*.

Nous traduisons cette hypothèse avec l'objectif perceptuel :

$$\mathcal{L}_\mathcal{P} = -\rho(\cos(s, s'), \cos(v_s, v_{s'}))$$

où ρ est la corrélation de Pearson.

Objectif global. La fonction objectif multimodale finale est une combinaisons linéaire des objectifs décrits précédemment, pondérés par les hyperparamètres α_T, α_P and α_C :

$$\mathcal{L}(\theta, \theta') = \underbrace{\alpha_T.\mathcal{L}_\mathcal{T}(\theta)}_{\text{textual context}} + \underbrace{\alpha_P.\mathcal{L}_\mathcal{P}(\theta, \theta') + \alpha_C.\mathcal{L}_\mathcal{C}(\theta)}_{\text{visual context } \mathcal{L}_\mathcal{V}}$$

3.3. *Modélisation des vidéos*

Afin d'évaluer l'impact de la sémantique visuelle sur les représentations de phrases, nous examinons différents types de contexte visuels. Comme fait dans (Yao *et al.*, 2016 ; Guo *et al.*, 2016), des caractéristiques visuelles sont extraites en utilisant l'avant-dernière couche d'un CNN pré-entraîné. Une vidéo est représentée par un ensemble de n images $(I_k)_{k\in[1,n]}$. Soit $(i_k)_{k\in[1,n]}$ les représentations de ces images obtenues avec le CNN. Nous présentons trois façons très simples de représenter la vidéo V. Notez que notre modèle peut se généraliser à tout autre modèle de représentation vidéo plus complexe, comme (Qiu *et al.*, 2017) par exemple.

• *Frame* (F) : cette configuration très simple revient à garder la première image de la vidéo et d'ignorer le reste de la séquence (toute autre image de la vidéo pourrait être utilisée). Le vecteur de contexte visuel est $v = i_1$.
• *Average* (A) : l'aspect temporel est ignoré ; la scène est représentée par la moyenne des caractéristiques des images : $v = \frac{1}{n}\sum_{k=1}^{n} i_k$ (Zha *et al.*, 2015).
• *Temporal* (T) : l'intuition est que, dans une vidéo, toutes les images ne sont pas forcéments pertinentes pour comprendre la phrase associée. Un mécanisme d'attention permet de se focaliser sur les images importantes. Nous posons : $v = \sum_{k=1}^{n} \beta_k i_k$, où $\beta_k = \text{softmax}(< \sum_w u_w, N.i_k >)$. Nous sommons sur tous les mots w de la description s, u_w étant une représentation fixée et pré-entraînée de w, et N une matrice de projection à apprendre.
Afin de mesurer l'information apportée par la vidéo, nous introduisons un modèle de base (R) où les vecteurs visuels sont échantillonnés selon une distribution normale.

4. Protocole d'évaluation

4.1. *Datasets*

Dataset textuel. Comme (Kiros *et al.*, 2015 ; Hill *et al.*, 2016), nous utilisons le Toronto BookCorpus dataset comme corpus textuel C_T. Ce corpus est composé de 11K livres ; un total de 74M phrases ordonnées, avec en moyenne 13 mots par phrase.

Dataset visuel. Nous nous servons du *MSVD dataset* (Chen et Dolan, 2011) comme corpus visuel C_V. Ce dataset de légendes de vidéo contient 1970 vidéos et 80K descriptions en anglais, chaque phrase décrivant la totalité de la vidéo associée. Notez que le total de phrases de C_V ne représente que 0.1% du total de phrases présentes dans C_T, ce qui est négligeable. Ainsi, les différences de performances entre les modèles ancrés visuellement et les modèles purement textuels ne peuvent pas être expliquées par la quantité de données textuelles présentes dans C_V.

4.2. *Modèles de Base et Scénarios*

Dans les expériences, nous choisissons comme modèle textuel un des modèles de phrases les plus établis : SkipThought (Kiros *et al.*, 2015) - que nous notons **ST**- i.e. l'encodeur de phrases est un GRU et $\mathcal{L}_\mathcal{T}$ correspond à l'objectif SkipThought). Nous démontrons la généralisabilité de notre modèle en utilisant FastSent (Hill *et al.*, 2016) dans la section 5.6. Une fois le modèle textuel choisi, nous en dérivons plusieurs modèles de base et scénarios, chacun représentant une approche différente de l'ancrage visuel. Comme notre objectif est d'étudier les différentes techniques d'ancrage visuel et leur influence sur les représentations de phrases, tous les baselines et scénarios partagent la même dimension de représentation et sont entraînés sur les mêmes datasets (4.1).

Scénarios du modèle. Nous testons différentes variantes de notre modèle d'ancrage visuel présenté en section 3. Nous notons ces variantes $\mathbf{M}^I_V(\alpha_T, \alpha_P, \alpha_C)$, qui dépendent de :
• l' *initialisation* $I \in \{p, \varnothing\}$: l'encodeur de phrases F_θ est soit pré-entraîné en utilisant l'objectif textuel $\mathcal{L}_\mathcal{T}$ ($I = p$), ou initialisé de façon aléatoire ($I = \varnothing$).
• la *représentation visuelle* $V \in \{F, A, T, R\}$: où F, A, T and R sont les différentes modélisations vidéos décrites en section 3.3.

Modèles de base. Nous transposons aux phrases deux modèles d'ancrage visuel originellement conçus pour les mots :
• *Projection* (**P**) : Inspirée de (Lazaridou *et al.*, 2015), ce modèle de base projette les vidéos dans l'espace textuel. Nous utilisons un objectif de classement :

$$\sum_{s,v_-} max(0, m' - \cos(s, W.v_-) + \cos(s, W.v_s))$$

où W est une matrice de projection à apprendre et m' une marge fixée. Nous notons $\mathbf{P}^I_V(\alpha_T)$ les variantes de ce modèle de base utilisant la fonction objectif $\mathcal{L} = \alpha_T.\mathcal{L}_\mathcal{T} + \mathcal{L}_\mathcal{V}$.
• *Séquentiel* (**SEQ**) : En s'inspirant de (Collell Talleda *et al.*, 2017), nous apprenons une régression linéaire (W, b) pour prédire la représentation visuelle de la vidéo à partir de la représentation SkipThought. Le vecteur de phrase ancré est la concaténation du vecteur SkipThought originel et de sa représentation visuelle "imaginée" : $\mathbf{ST} \oplus W\mathbf{ST} + b$, que nous projetons à l'aide d'une PCA afin d'avoir une représentation de la même taille que notre modèle.

Modèles de base externes. Nous notons **E** notre ré-implémentation du seul travail portant sur l'ancrage visuel des phrases : le modèle *GroundSent-Both model* de (Kiela *et al.*, 2018), avec une couche linéaire de projection entre les modalités ; nous utilisons le même encodeur de phrases que notre modèle.

4.3. *Tâches*

A la suite des travaux traitant de représentations de phrases (Kiros *et al.*, 2015 ; Hill *et al.*, 2016), nous considérons plusieurs benchmarks pour évaluer la qualité des nos représentations ancrées :

Similarité sémantique. Nous utilisons deux benchmarks bien connus de similarité sémantique pour les phrases : STS (Cer *et al.*, 2017) et SICK (Marelli *et al.*, 2014), qui sont composés de paires de phrases associées à un jugement de similarité déterminé par des humains. STS est subdivisé en trois sources textuelles : *Captions* contient des phrases concrètes décrivant des actions de la vie de tous les jours, tandis que les autres contiennent des phrases plus abstraites : des gros titres d'information pour *News* et des posts de forum pour *Forum*. Les corrélations de Spearman et Pearson sont mesurées entre par le cosinus des représentations de phrases et les scores de similarités

humains correspondant. Les hyperparamètres sont déterminés sur SICK/trial (les résultats reportés dans les tableaux sont calculés sur SICK/train+test). Notez que nous n'apprenons pas de modèle en plus de nos vecteurs de phrases, contrairement aux scores SICK reportés dans (Kiros *et al.*, 2015).

Benchmarks de classification. Nous utilisons sept benchmarks de classification de phrases : polarité d'opinions (MPQA) (Wiebe et Cardie, 2005), critique de films (MR) (Pang et Lee, 2005), subjectivité/objectivité (SUBJ) (Scott *et al.*, 2004), classification de type de questions (TREC) (Voorhees, 2001), revue de consommateurs (CR) (Hu et Liu, 2004), analyse de sentiment sur SST (Socher *et al.*, 2013b), et le corpus d'identification de paraphrase MSRP (Dolan *et al.*, 2004). Pour chaque jeu de données, un classifieur de régression logistique est appris à partir des représentations de phrases, et nous reportons la précision de classification.

Mesures structurelles. Afin d'étudier l'*information perceptuelle*, nous introduisons $\rho_{vis} = \rho(\cos(s, s'), \cos(v_s, v_{s'}))$: cela mesure si les similarités entre phrases corrèlent bien avec les similarités entre leurs vidéos correspondantes. Pour étudier l'*information de groupe*, nous introduisons $C_{intra} = \mathbb{E}_{v_s = v_{s'}}[cos(s, s')]$, qui mesure l'homogénéité des groupes (i.e. la similarité moyenne entre phrases à l'intérieur des groupes), et $C_{inter} = \mathbb{E}_{v_s \neq v_{s'}}[cos(s, s')]$, qui mesure si les groupes sont bien séparés les uns des autres (i.e. similarité moyenne pour des paires de phrases appartenant à des groupes différents).

4.4. *Détails d'Implémentation*

Les vidéos sont échantillonnées à une fréquence de 3 images par seconde. Ensuite, les images sont traitées par un réseau VGG pré-entraîné (Simonyan et Zisserman, 2014). La fonction objectif multimodale $\mathcal{L}$ est optimisée avec Adam (Kingma et Ba, 2014) et un taux d'apprentissage $\lambda = 8.10^{-4}$. Avec SkipThought (resp. FastSent) comme modèle textuel, nous utilisons la même dimension de représentation (2400 pour SkipThought, 100 pour FastSent) et les mêmes hyperparamètres que (Kiros *et al.*, 2015) (resp. (Hill *et al.*, 2016)). Les autres hyperparamètres sont déterminés avec la corrélation de Pearson sur SICK/trial : $\alpha_P = 0.1$, $\alpha_C = 1$, $m = m' = 0.5$.

5. Résultats expérimentaux

Dans cette partie expérimentale, nous présentons les résultats obtenus avec le modèle SkipThought comme modèle textuel. La généralisabilité de notre modèle à un autre modèle de phrases (FastSent) est discuté en section 5.6.

	STS/All	STS/Cap.	STS/News	STS/For.	SICK
ST	40/41	44/42	38/42	21/22	52/55
$\mathbf{M}_R$	51/53	62/62	39/43	23/24	56/57
$\mathbf{M}_F$	57/59	75/75	**41/46**	24/26	**60**/61
$\mathbf{M}_A$	**58/60**	75/75	**41**/45	23/26	59/**63**
$\mathbf{M}_T$	57/**60**	**76/76**	**41/46**	**25/27**	**60/63**

Tableau 1. *Video modelings comparison on semantic relatedness. We use model $\boldsymbol{M}^p_{\bullet(0,1,0)}$, noted $\boldsymbol{M}_\bullet$ for simplicity's sake. Results are given in the form $\rho_{Spearman}/\rho_{Pearson}$.*

5.1. *Préliminaire : discussion sur la modélisation des vidéos*

Afin de comparer les différents modélisations de vidéo, nous reportons les scores de similarité sémantiques dans le Tableau 1. Comme l'objectif perceptuel est le seul composant de notre modèle qui exploite le contenu des vidéos, nous n'utilisons pas les informations textuelles et groupe : nous posons donc : $\mathcal{L} = \mathcal{L}_\mathcal{P}$ (i.e. le modèle $\mathbf{M}^p_\bullet(0,1,0)$). Nous observons que : (1) $\mathbf{M}_\bullet >$ **ST** : ce qui montre l'importance et l'utilité de l'information perceptuelle ; (2) $\mathbf{M}_R >$ **ST** : dans ce cas, l'objectif perceptuel a pour effet d'éloigner les phrases visuellement différentes ; en effet, lorsque qu'on utilise des représentations aléatoires et indépendantes pour les vidéos, on a en moyenne $\cos(v_s, v_{s'}) \approx 0$; (3) utiliser plus d'une image est légèrement mieux que d'en utiliser une seule - e.g, A a $+3.3\%$ d'amélioration relative par rapport à F pour la corrélation de Pearson sur SICK ; (4) sélectionner les images importantes de la vidéo - i.e. modèle T - plutôt que de considérer toutes les images avec une importance égale - modèle A - améliore la qualité des représentations. On peut noter que les différences de performances suivant les modélisations vidéos F, A, T sont relativement faibles. Cela peut s'expliquer par le fait que les vidéos du jeu de données MSVD sont courtes (10 secondes en moyenne) et ne contiennent que très peu de changements de plans. Ainsi, quasiment toutes les images peuvent constituer un contexte visuel pertinent pour les descriptions associées. Nous pensons que de plus grandes différences de performances pourraient être observées pour un jeu de données contenant de plus longues vidéos. Dans le reste de la partie expérimentale, nous utilisons la représentation A pour les vidéos, ce qui est la plus simple et offre un bon compromis entre efficacité (T) et temps de calcul (F).

5.2. *L'impact de l'ancrage visuel (QR1)*

Dans cette section, nous nous intéressons à l'effet de l'ancrage visuel sur les représentations de phrases. Nous donnons dans le Tableau 2 les mesures structurelles (calculées en pourcentage sur le jeu de test de MSVD) et les scores de similarité sémantique sur trois versions de notre modèle — $\mathbf{M}_c$ (information de groupe), $\mathbf{M}_p$ (information perceptuelle) et $\mathbf{M}_b$ (la combinaison des deux), et les baselines **ST** et **P** ;

Model		Mesures Struct.			Similarité sémantique				
		ρ^{vis}	C^{intra}	C^{inter}	STS/All	STS/Cap.	STS/News	STS/For.	SICK
Texte	**ST**	16	43	25	40/41	44/42	38/42	21/22	52/55
Proj. **P**	$\mathbf{P}^p_A(0)$	37	63	06	62/67	82/84	43/48	**29/31**	61/75
Cluster $\mathbf{M}_c$	$\mathbf{M}^p_A(0, \quad 0, 1)$	38	**66**	**01**	62/66	83/84	41/46	22/24	**62/76**
Percep. $\mathbf{M}_p$	$\mathbf{M}^p_A(0, \quad 1, 0)$	**53**	44	18	58/60	75/75	41/45	23/26	59/63
Combiné $\mathbf{M}_b$	$\mathbf{M}^p_A(0, 0.1, 1)$	46	63	02	**64/68**	**84/85**	**44/49**	27/29	**62/76**

Tableau 2. ***QR1,2*** *: Mesures structurelles et performances de similarité pour différentes hypothèses d'ancrage visuel.*

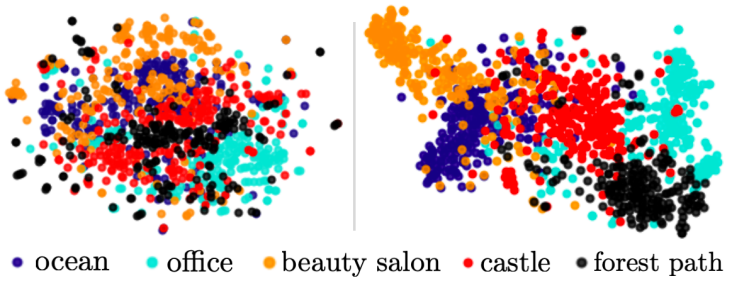

Figure 2. ***QR1*** *: Visualisations t-SNE des phrases de CMPlaces pour 5 scènes visuelles. A gauche : modèle textuel* ***ST****. A droite : modèle ancré* $\boldsymbol{M}_b$*.*

nous n'utilisons pas l'objectif textuel (i.e. nous posons $\alpha_T = 0$) pour isoler l'effet des différentes hypothèses d'ancrage visuel.

Les résultats indiquent que l'ancrage visuel (notre modèle $\mathbf{M}_b$) améliore les scores de similarité sémantique par rapport au modèle purement textuel **ST**. Pour comprendre dans quel cas l'ancrage visuel est utile, nous utilisons un score de concrétude des mots et prenons sa moyenne $\bar{c}$ sur l'ensemble des phrases d'un jeu de données. Nous utilisons des jugements humains de concrétude (entre 0 et 5) pour 40,000 mots anglais (Brysbaert *et al.*, 2013). Les plus grands gains de performances Δ entre $\mathbf{M}_b$ et **ST** sont observés lorsque le jeu de données a une grande concrétude $\bar{c}$. En effet, lorsque $\bar{c}$ est élevé, l'amélioration est conséquente : $\Delta = +10/21$ pour SICK ($\bar{c} = 3.12$) et $\Delta = +40/43$ pour *Captions* ($\bar{c} = 3.10$). Pour les jeux de données ayant des scores de concrétude plus bas comme *News* ($\bar{c} = 2.61$) et *Forum* ($\bar{c} = 2.39$), l'amélioration est plus petite ($\Delta = +6/7$). Ainsi, l'ancrage visuel apporte de l'information complémentaire utile, surtout pour les phrases concrètes.

Cette observation est confirmée par une autre expérience qui démontre que l'ancrage visuel raffine la compréhension des phrases lorsque les phrases décrivent des situations visuelles. Nous utilisons les phrases de CMPlaces (Castrejon *et al.*, 2016), qui décrivent des scènes visuelles (e.g, *coast*, *shoe-shop*, *plaza*, *pond*, *cockpit*, etc.)

Query	**ST**	$\mathbf{M}_b$
A man is **horrified**	An older man in a suit is smiling	The man is holding his face and screaming
This is a **tragedy**	I think this is a huge food court	View from the survivor of a motorcycle accident
Two people are in **love**	Two people are out in the ocean kitesurfing	A couple of people that are next to each other

Tableau 3. ***QR1*** *: Plus proches voisins d'une phrase donnée parmi les phrases de MS COCO.*

	VisSim	SemSim	Simlex	MEN	WordSim
ST	49	61	48	63	59
$\mathbf{M}_b$	**50**	**62**	**49**	**65**	**60**

Tableau 4. *Comparaison de **ST** et de $\mathbf{M}_b$ sur les tâches de similarité de mots.*

et sont classifiées en 205 catégories. Nous sélectionnons aléatoirement 5 scènes visuelles et visualisons leurs phrases correspondantes (cf. Figure 2) en utilisant t-SNE (Maaten et Hinton, 2008). Nous remarquons que notre modèle d'ancrage visuel est meilleur pour regrouper les phrases qui ont un contenu visuel similaire par rapport au modèle purement textuel. Ce constat est renforcé par le calcul des mesures structurelles sur les cinq groupes de la Figure 2 : $C_{inter} = 18, C_{intra} = 23$ pour **ST**, $C_{inter} = 11, C_{intra} = 28$ pour $\mathbf{M}_b$. En effet, C_{inter} (resp. C_{intra}) est inférieur (resp. supérieur) pour le modèle ancré $\mathbf{M}_b$ comparé au modèle purement textuel **ST**, ce qui montre que les groupes associés aux différentes scènes sont plus clairement séparés dans l'espace (resp. les phrases correspondant aux mêmes scènes sont plus regroupées).

De plus, nous montrons dans le Tableau 3 que la connaissance acquise par notre modèle d'ancrage visuel peut également être transférée à des phrases plus abstraites. Pour ce faire, nous construisons manuellement des phrases composés de mots plus abstraits, i.e. ayant une concrétude basse (entre 2.5 and 3.5) et tirés du dataset USF (Nelson *et al.*, 2004). Ensuite, nous trouvons les plus proches voisins dans l'ensemble de phrases de MS COCO (Lin *et al.*, 2014). Notre modèle ancré est plus précis que le modèle purement textuel pour capturer le sens visuel, même pour les phrases qui ne sont pas font pas référence directement à la modalité visuelle. Par exemple, à la première ligne du Tableau 3, la phrase de **ST** contredit la phrase en décrivant l'homme comme "smiling" ("souriant"), tandis que la phrase de $\mathbf{M}_b$ donne une illustration concrète de l'horreur : "grabs his head while screaming" ("attrape sa tête en criant"). L'information perceptuelle des phrases concrètes se propage donc aux phrases abstraites, ce qui généralise les observations faites au niveau des mots (Hill et Korhonen, 2014).

Finalement, nous calculons les performances pour la tâche de similarité sémantique ; les corrélations de Spearman sont données dans le Tableau 4 pour les cinq jeux de données suivants : Simlex (Hill *et al.*, 2015), MEN (Bruni *et al.*, 2014), Word-

Sim353 (Finkelstein *et al.*, 2002), VisSim et SemSim (Silberer et Lapata, 2014). Ici encore, l'ancrage visuel produit également de meilleures représentations de mots ; par exemple, $\mathbf{M}_b = \mathbf{M}_A^p(0)$ bat **ST** sur tous les benchmarks.

5.3. *Influence des informations de groupe et perceptuelles sur la structure de l'espace textuel (QR2)*

Afin de répondre à **QR2**, nous étudions l'influence des informations groupe et perceptuelles sur la structure de l'espace de représentations textuelles en utilisant des mesures intrinsèques. Nous commentons donc les mesures de ρ_{vis}, C_{intra} et C_{inter} indiquées dans le Tableau 2 pour les modèles $\mathbf{M}_c$, $\mathbf{M}_p$, $\mathbf{M}_b$, **ST** et **P**. Comme nous l'avions supposé, utiliser uniquement l'information de groupe mène à la plus haute valeur pour C_{intra} et la plus basse pour C_{inter} ; cela suggère que $\mathbf{M}_c$ est le modèle le plus efficace pour séparer les phrases visuellement différentes. Utiliser uniquement l'information perceptuelle avec $\mathbf{M}_p$ mène logiquement à des espaces visuels et textuels très corrélés (plus haute valeur de ρ_{vis}), mais la structure de voisinage n'est pas bien préservée (C_{intra} le plus bas et C_{inter} le plus haut). $\mathbf{M}_b$ et **P** sont optimisés pour former des groupes bien séparés et pour capturer de l'information perceptuelle dans leurs espace de représentation ; ce qui se traduit par un C_{intra} haut et un C_{inter} bas. Cependant, la différence entre ces modèles réside dans le fait que $\mathbf{M}_b$ est meilleur à capturer l'information de groupe (C_{intra} plus haut, C_{inter} plus bas) ainsi que l'information perceptuelle (ρ_{vis} plus haut). Nous pensons que cette différence explique des meilleures performances de similarité sémantique pour le modèle combiné $\mathbf{M}_b$. Cela renforce notre postulat selon lequel les informations groupe et perceptuelles sont complémentaires. Pour les expériences décrites plus bas, nous utilisons donc systématiquement le modèle combiné (i.e. $\alpha_P = 0.1$, $\alpha_C = 1$).

5.4. *Comparaison des modèles sur les tâches de similarité et de classification*

Nous donnons dans le Tableau 5 les performances sur les tâches de similarité sémantique et de classification. Nous choisissons un poids textuel $\alpha_T = 500$ pour **P** et **M** ; nous écrivons $\mathbf{M}_A^\bullet$ (resp. $\mathbf{P}_A^\bullet$) à la place $\mathbf{M}_A^\bullet(500, .1, 1)$ (resp. $\mathbf{P}_A^\bullet(500)$) par souci de clarté. Nous observons que, de manière générale, les modèles ancrés ont des meilleures performances que le modèle purement textuel **ST**, aussi bien en similarité qu'en classification, ce qui complète la réponse à **QR1**.

Tout d'abord, nous comparons les modèles séquentiels et joints. Nous remarquons que les modèles joints **M** et **P** sont en général meilleurs que le modèle séquentiel **SEQ** en similarité et en classification. Par exemple, $\mathbf{M}_A^\varnothing$ a une amélioration relative de $6\%/16\%$ comparé à **SEQ** sur STS. En classification, $\mathbf{M}^\varnothing$ et $\mathbf{M}^p \geq$ **SEQ** sur 5 tâches sur 7. L'approche jointe a donc des résultats supérieurs a l'approche séquentielle, ce qui confirme les résultats reportés pour les représentations ancrées de mots (Zablocki *et al.*, 2018). Enfin, nos modèles entraînés depuis le début ($I = \varnothing$) se comportent légèrement mieux que les modèles pré-entraînés avec l'objectif textuel ($I = p$). Cela

Model		Similarité sémantique					Classification						
		STS/All	STS/Cap.	STS/News	STS/For.	SICK	MSRP	MPQA	MR	SUBJ	TREC	CR	SST
Text	**ST**	40/41	44/42	38/42	21/22	52/55	71.6	86.2	75.9	92.1	89.4	**82.5**	83.3
Seq.	**SEQ**	47/44	70/59	37/44	29/24	58/69	70.0	86.1	75.8	92.2	89.2	81.9	83.3
Proj.	$\mathbf{P}_A^{\varnothing}$	47/48	64/61	38/43	20/20	56/66	72.8	**86.3**	76.5	92.7	89.4	80.4	82.0
	$\mathbf{P}_A^{p}$	41/39	57/47	38/43	19/19	54/59	73.1	86.2	76.7	92.5	88.8	81.3	83.7
Ext.	**E**	43/41	57/53	38/**45**	**33/28**	51/56	72.2	86.2	76.4	92.1	88.9	79.5	83.9
Model	$\mathbf{M}_A^{\varnothing}$	**50/51**	**71/69**	37/43	20/19	**60/70**	72.2	**86.3**	**77.2**	**93.0**	**90.4**	81.0	82.1
	$\mathbf{M}_A^{p}$	44/43	61/54	**39**/44	19/19	54/60	**74.4**	86.1	76.3	92.6	88.8	81.3	**84.6**

Tableau 5. *Performances de similarité textuelles et de classification pour les différents baselines et scénarios de notre modèle.*

est peut-être dû au fait que les informations visuelles et textuelles sont intégrées de façon jointe depuis le début de l'entraînement, ce qui mène à de meilleures interactions entre les modalités.

Pour compléter notre réponse à **QR2**, nous comparons notre modèle **M** avec des modèles faisant intervenir des projections inter-modales. Notre modèle obtient de meilleurs résultats que **P** en similarité textuelle ; e.g, $\mathbf{M}^p$ a une amélioration relative de $7\%/10\%$ sur $\mathbf{P}^p$ sur le benchmark STS. En classification, $\mathbf{M}^{\varnothing} \geq \mathbf{P}^{\varnothing}$ (resp. $\mathbf{M}^p \geq \mathbf{P}^p$) sur 6 (resp. 5) tâches sur 7, et tous les maxima sont atteints par notre modèle, à l'exception de CR où c'est la **ST** (texte seulement) qui obtient le maximum (ce résultat est également observé par (Kiela *et al.*, 2018) et peut s'expliquer par la concrétude relativement basse de CR). Notre modèle surpasse également notre ré-implémentation **E** du modèle état-de-l'art de (Kiela *et al.*, 2018). Nous ne donnons pas leurs résultats originaux dans le Tableau 5, car leur modèle de base textuel est plus faible que le notre (à cause des différences d'encodeur et de dimensions). Si nous comparons, entre le Tableau 5 et les résultats donnés dans leur article, la différence Δ entre le modèle ancré le plus performant et les modèles de base textuels, nous remarquons que notre Δ est plus élevé que le leur sur MPQA, MR, SUBJ and MSRP, et il est égal pour SST. Sur MSRP, par exemple, $\Delta^{\text{Kiela et al.}} = 0.7$ et $\Delta^{\mathbf{M}_A^p} = 74.4 - 71.6 = 2.8$. Cela renforce notre postulat selon lequel préserver la structure de l'espace visuel est plus efficace que d'apprendre des projections inter-modales.

5.5. *Recherche inter-modale*

Nous considérons les tâches d'annotation d'image et de recherche d'image définies sur MS COCO (Lin *et al.*, 2014). Pour cela, nous apprenons un objectif de classement afin de rapprocher les images (représentées avec le réseau convolutionnel OxfordNet (Simonyan et Zisserman, 2014)) et leurs légendes dans un espace latent multimodal où images et phrases sont projetées (les paramètres d'apprentissage sont ces matrices de projection). Nous évaluons les modèles avec les mesures de rappel R@1, R@5, R@5 ainsi que que le rang médian (Med r). Dans le Tableau 6, nous donnons les résultats des modèles les plus performants (sur les tâches précédentes) pour la tâche

de recherche inter-modale. Nous constatons que, de manière similaire, l'ancrage visuel permet d'améliorer les performances. Cependant, l'apport de l'ancrage visuel est moins prépondérant par rapport aux autres tâches présentés plus tôt ; nous pensons que cela est dû à notre objectif qui cherche à intégrer l'apport visuel dans la représentation de mots, plutôt qu'à rajouter une composante visuelle aux mots.

	R@1	R@5	R@10	Med r	R@1	R@5	R@10	Med r
Modèle	Recherche d'Image				Annotation d'Image			
ST	35.2	69.4	82.3	2.8	29.1	64.3	79.3	3.0
$\mathbf{M}_{\mathbf{A}}^{p}(100)$	**35.8**	69.9	**83.1**	**2.6**	27.8	63.4	79.1	3.2
$\mathbf{M}_{\mathbf{A}}^{p}(10)$	35.5	**70.6**	82.6	**2.6**	27.9	63.5	**79.4**	3.2
$\mathbf{M}_{\mathbf{A}}^{\varnothing}(1)$	35.2	69.2	82.8	2.7	**29.2**	**65.0**	79.8	**2.9**

Tableau 6. *Performances des différents scénarios du modèle sur la tâche de recherche inter-modale*

5.6. *Généralisabilité*

Enfin, nous montrons que notre modèle multimodal permet d'améliorer les performances pour un autre modèle que SkipThought : ici, nous choisissons FastSent (que nous notons **FS**). Dans le Tableau 7, nous donnons les résultats de **FS** ainsi que de $\mathbf{M}_{A}^{\varnothing}(100, 0.1, 1)$, qui bat le modèle textuel sur toutes les tâches de similarité et de classification.

	STS	SICK	MSRP	MPQA	MR	SUBJ	TREC	CR	SST
FS	68/71	59/72	71.6	79.7	70.1	86.9	69.6	75.2	74.0
$\mathbf{M}_{A}^{\varnothing}(100)$	**71/74**	**64/78**	**72.5**	**81.2**	**71.0**	**87.6**	**70.6**	**76.4**	**74.6**

Tableau 7. *Evaluations de similarité sémantique et de classification avec le modèle FastSent (**FS**) comme modèle textuel.*

6. Conclusion

Dans cet article, nous avons proposé un modèle multimodal pour apprendre des représentations de phrases ancrées, qui a des performances état-de-l'art comparé aux autres approches d'ancrage visuel. Nos principaux résultats sont les suivants : (1) Les informations perceptuelles et de groupe sont utiles et complémentaires pour apprendre des représentations de phrases ; (2) Préserver la structure de l'espace visuel, en calquant les similarités textuelles sur les similarités visuelles, surpasse les stratégies basées sur les projections inter-modales. À l'avenir, nous souhaitons étudier la contribution temporelle contenue dans les vidéos et leur influence sur les représentations de phrases.

7. Remerciements

Ce travail a été effectué dans le cadre du Projet CHIST-ERA MUSTER (ANR-15-CHR2-0005) et de Labex SMART (ANR-11-LABX-65), et a bénéficié d'une aide de l'Etat gérée par l'Agence Nationale de la Recherche au titre du programme Investissements d'Avenir portant la référence ANR-11-IDEX-0004-02.

8. Bibliographie

Amer N. O., Mulhem P., Géry M., « Toward Word Embedding for Personalized Information Retrieval », *CoRR*, 2016.

Bahdanau D., Cho K., Bengio Y., « Neural Machine Translation by Jointly Learning to Align and Translate », *CoRR*, 2014.

Baroni M., « Grounding Distributional Semantics in the Visual World », *Language and Linguistics Compass*, vol. 10, n° 1, p. 3-13, 2016.

Bojanowski P., Grave E., Joulin A., Mikolov T., « Enriching Word Vectors with Subword Information », *TACL*, vol. 5, p. 135-146, 2017.

Bruni E., Tran N. K., Baroni M., « Multimodal Distributional Semantics », *J. Artif. Int. Res.*, vol. 49, n° 1, p. 1-47, January, 2014.

Brysbaert M., Beth Warriner A., Kuperman V., « Concreteness ratings for 40 thousand generally known English word lemmas », *Behavior research methods*, 10, 2013.

Castrejon L., Aytar Y., Vondrick C., Pirsiavash H., Torralba A., « Learning Aligned Cross-Modal Representations from Weakly Aligned Data », *Computer Vision and Pattern Recognition (CVPR), 2016 IEEE Conference on*, IEEE, 2016.

Cer D. M., Diab M. T., Agirre E., Lopez-Gazpio I., Specia L., « SemEval-2017 Task 1 : Semantic Textual Similarity - Multilingual and Cross-lingual Focused Evaluation », *CoRR*, 2017.

Chen D. L., Dolan W. B., « Collecting Highly Parallel Data for Paraphrase Evaluation », *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics : Human Language Technologies - Volume 1*, HLT '11, Association for Computational Linguistics, Stroudsburg, PA, USA, p. 190-200, 2011.

Chomsky N., « Rules and representations », *Behavioral and brain sciences*, vol. 3, n° 1, p. 1-15, 1980.

Collell G., Moens M., « Do Neural Network Cross-Modal Mappings Really Bridge Modalities ? », *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 2 : Short Papers*, p. 462-468, 2018.

Collell Talleda G., Zhang T., Moens M.-F., « Imagined visual representations as multimodal embeddings », *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence (AAAI-17)*, AAAI, 2017.

Dolan B., Quirk C., Brockett C., « Unsupervised Construction of Large Paraphrase Corpora : Exploiting Massively Parallel News Sources », *COLING 2004, 20th International Conference on Computational Linguistics, Proceedings of the Conference, 23-27 August 2004, Geneva, Switzerland*, 2004.

Fincher-Kiefer R., « Perceptual components of situation models », *Memory & Cognition*, vol. 29, n° 2, p. 336-343, Mar, 2001.

Finkelstein L., Gabrilovich E., Matias Y., Rivlin E., Solan Z., Wolfman G., Ruppin E., « Placing Search in Context : The Concept Revisited », *ACM Trans. Inf. Syst.*, vol. 20, n° 1, p. 116-131, January, 2002.

Gordon J., Van Durme B., « Reporting Bias and Knowledge Acquisition », *Proceedings of the 2013 Workshop on Automated Knowledge Base Construction*, AKBC '13, ACM, New York, NY, USA, p. 25-30, 2013.

Guo Z., Gao L., Song J., Xu X., Shao J., Shen H. T., « Attention-based LSTM with Semantic Consistency for Videos Captioning », *Proceedings of the 2016 ACM on Multimedia Conference*, MM '16, ACM, New York, NY, USA, p. 357-361, 2016.

Harris Z. S., « Distributional structure », *Word*, vol. 10, n° 2-3, p. 146-162, 1954.

Hill F., Cho K., Korhonen A., « Learning Distributed Representations of Sentences from Unlabelled Data », *NAACL HLT 2016, The 2016 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, San Diego California, USA, June 12-17, 2016*, p. 1367-1377, 2016.

Hill F., Korhonen A., « Learning Abstract Concept Embeddings from Multi-Modal Data : Since You Probably Can't See What I Mean », *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*, p. 255-265, 2014.

Hill F., Reichart R., Korhonen A., « SimLex-999 : Evaluating Semantic Models With (Genuine) Similarity Estimation », *Computational Linguistics*, vol. 41, n° 4, p. 665-695, 2015.

Hochreiter S., Schmidhuber J., « Long Short-Term Memory », *Neural Comput.*, vol. 9, n° 8, p. 1735-1780, November, 1997.

Hu M., Liu B., « Mining and Summarizing Customer Reviews », *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '04, ACM, New York, NY, USA, p. 168-177, 2004.

Kalchbrenner N., Grefenstette E., Blunsom P., « A Convolutional Neural Network for Modelling Sentences », *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, ACL 2014, June 22-27, 2014, Baltimore, MD, USA, Volume 1 : Long Papers*, p. 655-665, 2014.

Karpathy A., Li F., « Deep visual-semantic alignments for generating image descriptions », *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, p. 3128-3137, 2015.

Kiela D., Conneau A., Jabri A., Nickel M., « Learning Visually Grounded Sentence Representations », *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers)*, p. 408-418, 2018.

Kingma D. P., Ba J., « Adam : A Method for Stochastic Optimization », *CoRR*, 2014.

Kiros R., Zhu Y., Salakhutdinov R., Zemel R. S., Urtasun R., Torralba A., Fidler S., « Skip-Thought Vectors », *Advances in Neural Information Processing Systems 28 : Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, p. 3294-3302, 2015.

Lazaridou A., Pham N. T., Baroni M., « Combining Language and Vision with a Multimodal Skip-gram Model », *NAACL HLT 2015, The 2015 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, Denver, Colorado, USA, May 31 - June 5, 2015*, p. 153-163, 2015.

Lin T., Maire M., Belongie S. J., Hays J., Perona P., Ramanan D., Dollár P., Zitnick C. L., « Microsoft COCO : Common Objects in Context », *Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V*, p. 740-755, 2014.

Lin X., Parikh D., « Don't just listen, use your imagination : Leveraging visual common sense for non-visual tasks », *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, p. 2984-2993, 2015.

Lin Z., Feng M., dos Santos C. N., Yu M., Xiang B., Zhou B., Bengio Y., « A Structured Self-attentive Sentence Embedding », *CoRR*, 2017.

Lioma C., Simonsen J. G., Larsen B., Hansen N. D., « Non-Compositional Term Dependence for Information Retrieval », *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval, Santiago, Chile, August 9-13, 2015*, p. 595-604, 2015.

Logeswaran L., Lee H., « An efficient framework for learning sentence representations », *International Conference on Learning Representations*, 2018.

Maaten L. v. d., Hinton G., « Visualizing data using t-SNE », *Journal of machine learning research*, vol. 9, , p. 2579-2605, 2008.

Manotumruksa J., MacDonald C., Ounis I., « Modelling User Preferences using Word Embeddings for Context-Aware Venue Recommendation », *CoRR*, 2016.

Marelli M., Bentivogli L., Baroni M. G., Bernardi R., Menini S., Zamparelli R., « SemEval-2014 Task 1 : Evaluation of Compositional Distributional Semantic Models on Full Sentences through Semantic Relatedness and Textual Entailment », *SemEval@COLING*, 2014.

Mikolov T., Sutskever I., Chen K., Corrado G. S., Dean J., « Distributed Representations of Words and Phrases and their Compositionality », *Advances in Neural Information Processing Systems 26 : 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States.*, p. 3111-3119, 2013.

Nelson D. L., McEvoy C. L., Schreiber T. A., « The University of South Florida free association, rhyme, and word fragment norms », *Behavior Research Methods, Instruments, & Computers*, vol. 36, nº 3, p. 402-407, 2004.

Norman D. A., « Memory, knowledge, and the answering of questions. », 1972.

Pang B., Lee L., « Seeing Stars : Exploiting Class Relationships for Sentiment Categorization with Respect to Rating Scales », *ACL 2005, 43rd Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference, 25-30 June 2005, University of Michigan, USA*, p. 115-124, 2005.

Pennington J., Socher R., Manning C. D., « Glove : Global Vectors for Word Representation. », *EMNLP*, vol. 14, p. 1532-1543, 2014.

Peters M. E., Neumann M., Iyyer M., Gardner M., Clark C., Lee K., Zettlemoyer L., « Deep Contextualized Word Representations », *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language*

Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers), p. 2227-2237, 2018.

Qiu Z., Yao T., Mei T., « Learning Spatio-Temporal Representation with Pseudo-3D Residual Networks », *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, p. 5534-5542, 2017.

Sagara T., Hagiwara M., « Natural language neural network and its application to question-answering system », *Neurocomputing*, vol. 142, p. 201-208, 2014.

Salton G., McGill M. J., « Introduction to modern information retrieval », 1986.

Scott D., Daelemans W., Walker M. A. (eds), *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics, 21-26 July, 2004, Barcelona, Spain*, ACL, 2004.

Silberer C., Lapata M., « Learning Grounded Meaning Representations with Autoencoders », *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, ACL 2014, June 22-27, 2014, Baltimore, MD, USA, Volume 1 : Long Papers*, p. 721-732, 2014.

Simonyan K., Zisserman A., « Very Deep Convolutional Networks for Large-Scale Image Recognition », *CoRR*, 2014.

Socher R., Perelygin A., Wu J., Chuang J., Manning C. D., Ng A. Y., Potts C., « Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank », *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, EMNLP 2013, 18-21 October 2013, Grand Hyatt Seattle, Seattle, Washington, USA, A meeting of SIGDAT, a Special Interest Group of the ACL*, p. 1631-1642, 2013a.

Socher R., Perelygin A., Wu J., Chuang J., Manning C. D., Ng A. Y., Potts C., « Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank », *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, EMNLP 2013, 18-21 October 2013, Grand Hyatt Seattle, Seattle, Washington, USA, A meeting of SIGDAT, a Special Interest Group of the ACL*, p. 1631-1642, 2013b.

Voorhees E. M., « Overview of the TREC 2001 Question Answering Track », *In Proceedings of the Tenth Text REtrieval Conference (TREC*, p. 42-51, 2001.

W. Barsalou L., « Perceptual Symbol Systems », , vol. 22, p. 577-609 ; discussion 610, 09, 1999.

Wiebe J., Cardie C., « Annotating expressions of opinions and emotions in language. Language Resources and Evaluation », *Language Resources and Evaluation (formerly Computers and the Humanities*, p. 2005, 2005.

Yao L., Torabi A., Cho K., Ballas N., Pal C., Larochelle H., Courville A., *Describing videos by exploiting temporal structure*, vol. 11-18-December-2015, Institute of Electrical and Electronics Engineers Inc., United States, p. 4507-4515, 2, 2016.

Yatskar M., Ordonez V., Farhadi A., « Stating the Obvious : Extracting Visual Common Sense Knowledge », *NAACL HLT 2016, The 2016 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, San Diego California, USA, June 12-17, 2016*, p. 193-198, 2016.

Zablocki E., Piwowarski B., Soulier L., Gallinari P., « Learning Multi-Modal Word Representation Grounded in Visual Context », *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, New Orleans, Louisiana, USA, February 2-7, 2018*, 2018.

Zha S., Luisier F., Andrews W., Srivastava N., Salakhutdinov R., « Exploiting Image-trained CNN Architectures for Unconstrained Video Classification », *Proceedings of the British*

Machine Vision Conference 2015, BMVC 2015, Swansea, UK, September 7-10, 2015, p. 60.1-60.13, 2015.