Julien	RESUME D'ARTICLE	MS IA
GUEGAN		29/04/2020

Multimodal Machine Learning: A Survey and Taxonomy

Le monde qui nous entoure est multimodal : on voit des objets, on entend des sons, on ressent des textures, on sent des odeurs … En un sens plus général, une modalité désigne la manière qu'une expérience se passe. Un problème est donc multimodal quand il se base sur des données faisant intervenir au moins 2 modalités différentes. Ce domaine de recherche apporte de réels challenges étant donné la nature hétérogène des données mais offre la possibilité de capturer des correspondances entre modalités et une compréhension profonde des phénomènes naturels. L'article dresse un portrait des différentes techniques actuelles du machine learning multimodal en présentant les 5 challenges suivants : la **représentation** (résumer les données en exploitant leur complémentarité), la **translation** (passer d'une modalité à une autre), **l'alignement** (identifier les relations entre modalités), la **fusion** (unir les informations pour la prédiction) et le **co-apprentissage** (transférer les connaissances entre les modalités).

L'un des premiers exemples de recherche multimodale est la reconnaissance vocale audio-visuelle (en 1989) qui s'appuie sur l'interaction entre l'ouïe et la vision durant la perception d'un discours (on entend un son tout en voyant des lèvres en prononcées un autre, on perçoit alors un 3ème son). Les résultats ont montré que la combinaison du son et de l'image a permis d'améliorer en particulier la robustesse des prédictions quand le signal était bruité. Puis, du fait des fortes avancées techniques en détection de visage et reconnaissance d'expression, le domaine de la compréhension des comportements humains lors d'interaction sociales s'est fortement développée et essaie notamment de modéliser la tâche complexe qu'est la reconnaissance d'émotion lors d'échanges entre des orateurs et des auditeurs. Plus récemment (années 2010), une nouvelle catégorie d'application multimodale a émergé : la description de contenu multimédia. Avec en particulier la génération automatique de texte pour décrire une image donnée.

1. Représentations

La représentation des données brutes dans un format qui peut permettre de faire fonctionner un modèle de machine learning est une étape cruciale de modélisation et encore plus pour des données multimodales. Représenter des données multimodales pose plusieurs difficultés : comment combiner différentes sources, comment gérer différents niveaux de bruits ou encore comment traiter les données manquantes. Des travaux ont montré qu'une bonne représentation possède les propriétés suivantes : régularité, cohérence spatiale et temporelle, sparsité et regroupement naturel. Deux catégories de représentation multimodale peuvent être distinguer : conjointe et coordonnée. La représentation **conjointe** combine les features unimodales en un seul espace de représentation, elle est souvent utilisée lorsque les données multimodales sont disponibles durant l'étape d'entrainement et la prédiction. Pour réaliser ce type de représentation, on peut utiliser des réseaux de neurones profonds où les couches cachées projettent les modalités sur un espace conjoint, ils atteignent souvent d'excellentes performantes mais requièrent un important volume de données et gèrent mal les données manquantes. Les modèles probabilistes graphiques sont également des méthodes qui permettent de construire ce type de représentations en se basant sur l'utilisant de variables aléatoires latentes, ils ont pour avantage de ne pas avoir besoin de données étiquetées et donc gère mieux que les réseaux de neurone les données manquantes mais ils sont plus difficiles à entrainer. D'autre part, la représentation **coordonnée** utilise les features séparément mais leur imposent certaines contraintes de similarités, elle est adaptée pour des applications où une seule modalité est présente à l'étape de test. Par exemple, les modèles de similarités minimisent la distance entre les modalités dans l'espace coordonné (la représentation du mot *chien* est plus proche d'une image de *chien* qu'une image de *voiture*).

2. Translation

La translation d'une modalité à une autre consiste à générer une donnée disponible dans une modalité différente. Par exemple, ayant une image on voudrait générer une phrase la décrivant (ou l'inverse). Parmi les méthodes connues, on différencie deux types d'approches : basée sur exemples et générative. Les modèles **basés sur exemples** s'appuient sur un ensemble d'apprentissage de données dictionnaire qui permet simplement de récupérer l'échantillon le plus proche dans le dictionnaire et l'utiliser comme résultat translaté, la récupération peut être faite directement dans l'espace unimodal ou un espace intermédiaire. Le principal défaut des modèles basé sur exemple est que le dictionnaire entier représente le modèle ce qui le rend lent et que l'exemple à translater n'existe pas toujours dans le dictionnaire. Les approches **génératives** nécessitent la capacité à la fois de comprendre la modalité source et de générer la modalité cible. Une première sous-approche est d'utiliser une grammaire (un ensemble de règles) prédéfinie pour générer une modalité en particulier, par exemple pour passer d'une vidéo à une texte, on repère des concepts haut niveau dans la vidéo puis on génère une description de la forme : *qui* fait *quoi* et *où* et *comment* il le fait. Une autre sous approche possible est les modèles encodeur-décodeur qui se base sur réseaux de neurones entrainés de telle sorte d'encoder la modalité source dans un espace de représentation vectoriel et ensuite de décoder dans la modalité cible, le tout par une seule passe dans le réseau (les GANs ont récemment montré de bons résultats pour générer des images à partir de texte).

Un challenge majeur de la tâche de translation est la difficulté de l'évaluer. En effet, des tâches de synthèse de parole et de description d'image n'ont pas une unique et correcte translation, plusieurs réponses sont souvent possibles et décider quelle translation est la meilleure est subjectif. Un moyen d'évaluer une telle tâche peut être fait en utilisant le jugement humain qui peut noter la naturalité, le réalisme l'exactitude, la pertinence sur l'échelle de Likert. Mais cette façon de faire est longue, coûteuse et requière de faire attention aux biais qui peuvent être introduit par les personnalités qui jugent. D'autre part, il existe tout de même des métriques (BLEU, ROUGE, Meteor, CIDEr) qui mesurent la similarité entre un texte généré et une vérité terrain mais leur corrélation avec les jugements humain n'est pas toujours vraie. Une méthode proposée pour évaluer le sous-titrage d'image est la réponse-question visuelle qui consiste à questionner le contenue d'une image et le système doit y répondre, mais encore la pertinence et non ambiguïté de ces questions peut poser problèmes. C'est pourquoi les métriques d'évaluation seront un point crucial pour améliorer les actuels systèmes de translation.

3. Alignement

L'alignement multimodal est le fait de trouver des relations entre des composants de différentes modalités, par exemple on cherche les régions d'une image qui correspondent à certains mots de sa légende. Deux types d'alignement peuvent être distinguer : explicite et implicite. Dans l'alignement **explicite**, la tâche d'alignement est une étape claire et nécessaire d'une autre tâche plus globale. Une partie importante de ce travail repose sur la métrique de similarité, on cherche à mesurer une similarité entre les différents sous-composants des modalités. Cette mesure de similarité peut être définie manuellement et de manière non-supervisé avec notamment les approches de programmation dynamique comme DTW qui cherche une correspondance optimale entre 2 séquences ou encore CCA qui utilise un espace coordonné. Cependant ces approches non supervisé ont des limites de modélisation, c'est pourquoi des approches (comme du deep learning) s'appuyant sur des données étiquetées peuvent être plus efficaces et encore plus grâce à la récente disponibilité de jeu de données alignées fiables dans les communautés de la vision et du langage. D'autre part, l'alignement **implicite** est réalisé indirectement pendant une autre action. Par exemple, quand la translation est accomplie par un réseau de neurone on utilise le mécanisme d'attention pour se concentrer sur certains sous composants de la source. Pour le cas d'un problème de sous-titrage d'image, un mécanisme d'attention sera appris pour se concentrer sur une partie particulière de l'image lors de la génération successive des mots.

La tâche explicite d'alignement rencontre quelques difficultés : peu de données annotées, trouver une métrique de similarité n'est pas évident, plusieurs alignements peuvent exister sans pour autant avoir une correspondance dans les 2 modalités. Les travaux récents montrent que les techniques non supervisées permettent de s'affranchir de ces difficultés tout en résolvant une autre tâche.

4. Fusion

La fusion multimodale est le concept d'intégration de l'information de plusieurs modalités dans le but de faire de la classification ou de la régression. Ici, on essaie de ne pas confondre la représentation et la fusion multimodale qui sont souvent confondus dans les modèles et les descriptions actuelles et on décrit 2 principales catégories d'approche. La première est celle des approches dites **modèle-agnostique** qui sont celles le plus utilisées historiquement car simple à implémenter et peut être utiliser avec n'importe quel classifieur/régresseur unimodal. Dans ce type de technique, on intègre les différentes features indépendamment du modèle (avant ou après) par exemple en les concaténant, les moyennant, les pondérant … L'autre approche est celle **basé-sur-modèle** où le modèle de classification/régression est désigné pour répondre à ce problème de multimodalité, on distingue les méthodes multi-kernel, graphiques et réseaux de neurone. Les méthodes Multiple kernel learning sont une extension des SVM mais qui utilisent différents noyaux pour les différentes modalités, ils permettent une certaine flexibilité du fait de la sélection des noyaux mais surtout de trouver une solution globale au problème (fonction loss convexe). D'autres méthodes populaires sont les modèles graphiques qui modélise soit la probabilité jointe (génératif) soit la probabilité conditionnelle (discriminatif). Les modèles CRF sont particulièrement populaires et exploitent les structures spatiales et temporelles des données tout en permettant d'injecter des connaissances expertes dans le modèle. Finalement, les méthodes basées sur des réseaux de neurone sont utilisés fréquemment ce problème en fusionnant l'information dans les couches cachées. Les réseaux récurrents ont notamment permis d'incorporer des images et des phrases avec une architecture end-to-end qui facilite l'apprentissage d'à la fois la représentation et la fusion des composantes. L'obstacle principal de ces modèles reste le manque d'interprétabilité des prédictions et la nécessité d'un très grand jeu de données.

La fusion multimodale est un sujet qui a été largement traitée dans les sujets de recherche avec beaucoup d'approches différentes, chacune ayant des avantages et des inconvénients selon l'environnement et les données du problème. Malgré les avancées qui ont pu être réalisées, les challenges suivants sont toujours présents : les signaux peuvent ne pas être temporellement alignés, il est difficile d'avoir un modèle qui exploite une information supplémentaire (qui n'apparaît pas dans les modalités individuelle) et les modalités peuvent avoir différents types et niveaux de bruit à différents pas de temps.

5. Co-Apprentissage

Le dernier challenge multimodal se concentre sur l'aide à la modélisation d'une modalité pauvrement représentée dans les données en exploitant les connaissances apportées par une autre modalité plus riche. La plupart du temps ce processus est réalisé seulement lors de la phase d'apprentissage et pas pendant le test. Trois types d'approche sont identifiés. La 1ère est **parallèle** et requiert des données d'entrainement où les observations d'une modalité sont directement liées aux observations d'une autre modalité par exemple d'un discours audio-visuel où la vidéo et le discours proviennent du même interlocuteur. Dans ce cas, on peut utiliser du co-entrainement pour générer plus de label que disponibles en bootstrappant des classifieurs de chaque modalités. Notez que cette technique peut conduire à du surentrainement. On peut également faire du transfert d'apprentissage avec des modèles de réseau de neurone pour transférer l'information d'une représentation à une autre, il a par exemple été utiliser pour créer un modèle qui lit sur les lèvres. La 2ème approche est **non-parallèle** et repose sur des modalités qui ne partagent pas d'instance mais seulement des concepts. Dans ce cas, le *transfert d'apprentissage* est aussi possible et peut permettre d'améliorer la représentation et de réaliser de meilleurs performances. Le *conceptual grounding* est également une méthode applicable dans ce cas et fait référence à apprendre des concepts sémantique sans se reposer seulement sur le langage mais aussi sur la vision, le son ou autre. On peut aussi mettre en avant le *zero-shot learning* dont le but est de reconnaître un concept sans avoir eu à voir aucun exemple de celui-ci, de manière multimodal il permet de distinguer un objet dans une modalité à travers l'aide d'une seconde modalité. La 3ème approche combine les deux premières dans un processus **hybride** où 2 modalités non-parallèles sont reliées par une modalité partagée.

Le co-apprentissage multimodal permet à une modalité d'influencer l'entrainement d'une autre, en exploitant les informations complémentaires entre les modalités. Il est important de noter que le co-apprentissage est indépendant de la tâche et pourrait être utilisé pour créer de meilleurs modèles de fusion, de traduction et d'alignement. Ce défi est illustré par des algorithmes tels que le co-entrainement, l'apprentissage de la représentation multimodale, le

conceptual grounding et l'apprentissage zéro-shot (ZSL) et a trouvé de nombreuses applications dans la classification visuelle, la reconnaissance d'action, la reconnaissance vocale audiovisuelle et l'estimation de la similitude sémantique.

6. Conclusion

Cet article, introduit une taxonomie de l'apprentissage automatique multimodal : représentation, translation, fusion, alignement et co-apprentissage. Certains d'entre eux tels que la fusion sont étudiés depuis longtemps, mais un intérêt plus récent pour la représentation et la traduction a conduit à un grand nombre de nouveaux algorithmes multimodaux et d'applications multimodales passionnantes. Les auteurs espèrent que cette taxonomie aidera à cataloguer les futurs articles de recherche et aussi à mieux comprendre les problèmes non résolus restants auxquels fait face l'apprentissage automatique multimodal.

Dans ce résumé, j'ai essayé de présenter principalement l'idée générale qui est celui de lister et classifier les différentes catégories de problèmes multimodaux en m'intéressant aux types d'approches et algorithmes possibles. Cependant l'article original est bien plus détaillé et cite plus de 250 références bibliographique sur lesquels il s'appuie tout au long de ses explications. Un point important et intéressant du développement des auteurs est qu'ils n'oublient pas de donner les avantages et inconvénients des différents méthodes ce qui peut vraiment aider à choisir l'approche à choisir selon un problème donné. De plus, les quelques tableaux et schémas récapitulatifs sont, de mon point de vue, des éléments appréciable pour le lecteur car ils permettent d'avoir un regard bref et résumé sur les explications des auteurs. Cependant, l'article reste quand même très dense avec un condensé d'informations expertes qui se suivent les unes à la suite des autres sous forme d'énumérations et qui redirige vers les travaux cités dans la bibliographie. Le lecteur peut facilement se perdre et abandonner l'article, je conseillerais aux auteurs d'aérer un plus le discours en prenant le temps d'introduire et d'amener les concepts et les exemples en distinguant clairement une partie application pour que le lecteur puisse plus facilement savoir où il en est dans sa lecture. On pourrait également imaginer une annexe ou un paragraphe particulier qui donnerait un aperçu du fonctionnement général des différentes familles de modèle cités tout au long de l'article comme les CNN, LSTM, CRF, RBM pour les lecteurs.