Clustering multi-vues de documents selon un mécanisme de cross-attention pour la fusion de caractéristiques

Khadidja Wissal Baki*,**, Parisa Rastin*
Guénaël Cabanes*, Mohamed Elbahri*** David Calvo** Yannick Toussaint*

*LORIA, 615 Rue du Jardin-Botanique, 54506 Vandœuvre-lès-Nancy
prenom.nom@loria.fr,
**OSIDOC, 2 All. des Barbanniers, 92230 Gennevilliers
prenom.nom@osidoc.com,
***Université Djillali Liabès, Sidi Bel Abbès, Algérie
prenom.nom@univ-sba.dz

1 Introduction

L'augmentation croissante de la quantité de documents complexes renforce le besoin de modèles capables de traiter efficacement ce type de données multivues, où plusieurs types d'informations coexistent. Les approches classiques, traitant soit les caractéristiques textuelles et visuelles séparément, soit proposant une fusion simple des différents espaces de représentation, ne tirent pas pleinement parti de leur complémentarité. Cet article propose une méthode innovante pour fusionner ces caractéristiques via un mécanisme de cross-attention(Vaswani et al., 2017), utilisant LayoutLM (Xu et al., 2020) pour le texte et Vision Transformer (ViT) (Dosovitskiy et al., 2021) pour les informations visuelles (images, graphes,...). Cette intégration enrichit significativement la compréhension des documents et la performance de l'extraction d'informations, en particulier pour des documents complexes tels que les contrats, les formulaires ou les rapports financiers. Nous démontrons expérimentalement l'efficacité de cette approche pour le clustering de documents, avec des améliorations notables selon diverses métriques d'évaluation de clustering, en comparaison avec les modèles unimodaux et multimodaux existants.

2 Méthode

Extraction des caractéristiques : La méthode repose sur une extraction séparée des caractéristiques textuelles et visuelles. Le texte est analysé via LayoutLM, qui préserve le contexte spatial et sémantique. Cela inclut la tokenisation, l'intégration des informations de mise en page, et la création d'embeddings textuels contextualisés. Les caractéristiques visuelles sont extraites avec le Vision Transformer. L'image est divisée en patchs linéarisés, auxquels sont ajoutées des intégrations positionnelles. Ces patchs sont ensuite traités pour générer des embeddings visuels reflétant les informations spatiales et structurelles de l'image.

Mécanisme de cross-attention : Le mécanisme de cross-attention combine les embeddings textuels et visuels pour capturer leurs interactions. Le texte agit comme requête, et les images comme clé et valeur. Les scores sont calculés via un produit scalaire, normalisés avec softmax, puis pondérés pour produire des embeddings enrichis. Ces derniers alimentent des algorithmes de clustering comme KMeans et HDBSCAN, regroupant les documents selon leurs similarités multimodales.

3 Résultat

Notre méthode a été évaluée sur les jeux de données SROIE, RVL-CDIP et DocLayNet. Sur DocLayNet, elle atteint un score de silhouette de 0.92 et une pureté des clusters de 0.5154, surpassant les approches basées uniquement sur LayoutLMv3 ou BERT-Large.

Pour RVL-CDIP, le score Calinski-Harabasz est de 1213.59, tandis que pour SROIE, il atteint 6150.97, confirmant l'adaptabilité de notre méthode à divers types de documents.

Comparée à LayoutLMv3, qui se concentre davantage sur les images, et à BERT-Large, qui manque de capacités multimodales, notre approche exploite efficacement la synergie entre texte et image, générant des clusters plus précis.

En termes de coûts computationnels, la combinaison de LayoutLMv1 et Vision Transformer (199M paramètres) offre une solution plus légère que LayoutLMv3 (345M paramètres), tout en maintenant des performances optimales.

4 Conclusion

Cette étude propose un mécanisme de cross-attention performant pour la fusion multimodale des documents complexes. Les résultats montrent une nette amélioration du clustering, tant en termes de qualité des regroupements que d'efficacité computationnelle.

Notre approche est particulièrement adaptée aux scénarios où les ressources sont limitées. De plus, elle offre une flexibilité pour s'adapter à différents types de documents et contextes applicatifs. Dans les travaux futurs, nous envisageons d'étendre cette méthode à des domaines tels que la classification supervisée et l'analyse de grandes collections documentaires.

Références

Dosovitskiy, A., L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, et N. Houlsby (2021). An image is worth 16x16 words : Transformers for image recognition at scale.

Vaswani, A., N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, et I. Polosukhin (2017). Attention is all you need. Volume abs/1706.03762.

Xu, Y., M. Li, L. Cui, S. Huang, F. Wei, et M. Zhou (2020). Layoutlm : Pre-training of text and layout for document image understanding. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '20. ACM, doi: 10.1145/3394486.3403172.