Traduction de la Langue des Signes en Langue Naturelle par Modélisation Séquentielle Basée sur Transformers

Par Arnold Mokira et Manel O.

Ce rapport présente une approche complète pour la traduction automatique de la langue des signes en langue naturelle. Notre travail s'articule en deux phases : une première approche basée sur la classification statique d'alphabets gestuels utilisant MediaPipe et RandomForestClassifier, ayant atteint 100% de précision sur les images individuelles, suivie d'une analyse critique révélant les limitations de cette méthode pour la traduction contextuelle. Nous proposons ensuite une architecture avancée basée sur des Transformers pour la modélisation séquentielle de vidéos, permettant de capturer la dimension temporelle et contextuelle essentielle à la compréhension de la langue des signes. Cette évolution méthodologique reflète la transition d'une reconnaissance de caractères isolés vers une véritable traduction linguistique prenant en compte la nature séquentielle et syntaxique du langage gestuel.

Introduction: La langue des signes constitue le principal moyen de communication pour des millions de personnes sourdes et malentendantes à travers le monde. Cependant, la barrière linguistique entre les utilisateurs de la langue des signes et la population générale reste un défi majeur pour l'inclusion sociale et professionnelle. Les avancées récentes en vision par ordinateur et en apprentissage profond offrent des opportunités prometteuses pour développer des systèmes automatiques de traduction.

Les premiers travaux dans ce domaine se sont concentrés sur la reconnaissance de gestes isolés et d'alphabets manuels. Bien que ces approches aient démontré des performances élevées dans des environnements contrôlés, elles ne capturent pas la complexité intrinsèque de la langue des signes qui, comme toute langue naturelle, repose sur des structures grammaticales, des séquences temporelles et des contextes sémantiques.

Notre recherche s'inscrit dans cette évolution méthodologique, passant d'une reconnaissance statique d'alphabets à une modélisation séquentielle sophistiquée. Nous présentons d'abord notre système initial basé sur MediaPipe et RandomForestClassifier, puis analysons ses limitations avant de proposer une architecture Transformer adaptée à la traduction séquentielle de la langue des signes.

Travaux connexes: La reconnaissance automatique de la langue des signes a connu plusieurs générations de développements technologiques. Les premières approches utilisaient des gants équipés de capteurs pour capturer les mouvements de la main, mais ces dispositifs invasifs limitaient l'adoption pratique. L'avènement de la vision par ordinateur a permis des solutions non invasives basées sur l'analyse d'images et de vidéos.

Les méthodes récentes exploitent l'apprentissage profond pour extraire des caractéristiques visuelles pertinentes. Les réseaux de neurones convolutifs (CNN) ont démontré leur efficacité pour la classification d'images de gestes statiques, tandis que les architectures récurrentes (RNN, LSTM) ont permis de modéliser des séquences temporelles. L'émergence des Transformers a révolutionné le traitement séquentiel en permettant une meilleure capture des dépendances à long terme grâce aux mécanismes d'attention.

MediaPipe, développé par Google, représente une avancée majeure en fournissant des outils robustes pour la détection et le suivi en temps réel des mains, du corps et du visage. Son utilisation dans la reconnaissance de la langue des signes a été largement adoptée pour l'extraction de caractéristiques anatomiques précises.

Méthodologie initiale:

Architecture du système de classification statique

Notre première approche s'est concentrée sur la reconnaissance d'alphabets gestuels individuels. Le système se compose de trois modules principaux : l'extraction de caractéristiques, la classification et le déploiement en temps réel.

Extraction de caractéristiques avec MediaPipe

MediaPipe fournit un modèle pré-entraîné capable de détecter jusqu'à 21 points de repère pour chaque main dans une image. Pour chaque frame vidéo, nous appliquons la détection sur les deux mains simultanément. Les coordonnées tridimensionnelles (x, y, z) de ces 42 points (21 points par main) sont extraites et normalisées par rapport à la taille de l'image.

Ces coordonnées sont ensuite transformées en un vecteur aplati de dimension 126 (42 points × 3 coordonnées), formant ainsi la représentation numérique de chaque geste. Cette vectorisation préserve les informations spatiales relatives entre les différents points anatomiques de la main tout en réduisant la complexité dimensionnelle.

Classification par RandomForest

Le vecteur de caractéristiques extrait est utilisé comme entrée pour un classifieur RandomForestClassifier. Ce choix se justifie par plusieurs avantages : robustesse au surapprentissage grâce à l'agrégation d'arbres de décision, capacité à gérer des données non linéaires, et interprétabilité relative des décisions.

L'entraînement a été effectué sur un ensemble de données comprenant des exemples de chaque lettre de l'alphabet de la langue des signes. Chaque exemple consiste en une image statique où le geste est clairement visible. Le modèle a été optimisé avec 100 estimateurs et une profondeur maximale ajustée pour éviter le surapprentissage.

Déploiement en temps réel

Le système déployé capture le flux vidéo depuis une webcam standard. Pour chaque frame capturée, le pipeline suivant est exécuté : détection des mains via MediaPipe, extraction du vecteur de caractéristiques, prédiction de la lettre correspondante par le RandomForest, et affichage du résultat à l'écran.

La fréquence de traitement atteint approximativement 30 frames par seconde sur du matériel standard, permettant une expérience utilisateur fluide. Les prédictions sont affichées avec un niveau de confiance associé.

Résultats de l'approche initiale:

Performance de classification

Sur notre ensemble de test comprenant des images statiques de gestes bien formés, le modèle RandomForest a atteint une précision de 100%. Cette performance exceptionnelle s'explique par la nature contrôlée des données : éclairage uniforme, arrière-plan neutre, gestes statiques et bien définis.

La matrice de confusion révèle une absence totale de confusion entre les différentes lettres, indiquant que les caractéristiques extraites par MediaPipe sont suffisamment discriminantes pour distinguer les gestes alphabétiques dans des conditions idéales.

Limitations observées

Malgré ces résultats encourageants, le déploiement en conditions réelles a révélé plusieurs limitations critiques. Premièrement, la méthode traite chaque frame de manière indépendante, ignorant complètement la dimension temporelle de la communication gestuelle.

En pratique, la langue des signes ne consiste pas simplement en une succession de lettres isolées. Les signes complets combinent mouvements, positions, orientations et expressions faciales pour former des mots et des phrases. Un signe peut s'étendre sur plusieurs frames et sa signification dépend de la trajectoire complète du mouvement, et non d'une position statique unique.

Deuxièmement, le système génère une prédiction pour chaque frame, résultant en une séquence de lettres souvent redondantes et bruitées. Par exemple, maintenir un geste pendant une seconde génère 30 prédictions identiques, rendant la reconstruction du message difficile.

Troisièmement, les transitions entre gestes produisent des prédictions erronées car les configurations intermédiaires ne correspondent à aucune lettre valide. Sans modélisation de la structure temporelle, le système ne peut distinguer un geste intentionnel d'un mouvement transitoire.

Analyse critique et reformulation du problème:

Nature séquentielle de la langue des signes

La langue des signes est fondamentalement une langue visuo-gestuelle avec sa propre grammaire, syntaxe et sémantique. Contrairement à l'épellation digitale (fingerspelling) qui encode lettre par lettre, la plupart des signes sont des unités lexicales complètes représentant des mots ou des concepts entiers.

La compréhension d'un message nécessite donc l'analyse de séquences vidéo complètes, où chaque signe s'étend sur plusieurs frames consécutives. La signification émerge de la dynamique temporelle : vitesse, accélération, trajectoire et synchronisation avec les expressions faciales.

Reformulation comme problème de traduction séquence-à-séquence

Nous reformulons le problème comme une tâche de traduction séquence-à-séquence : étant donnée une séquence vidéo d'entrée représentant un message en langue des signes, produire une séquence textuelle en langue naturelle correspondant à la traduction fidèle du message.

Cette formulation capture mieux la nature du problème et s'aligne sur les architectures modernes de traitement du langage naturel, notamment les Transformers qui ont révolutionné la traduction automatique textuelle.

Architecture proposée basée sur Transformers:

Motivations pour l'utilisation de Transformers

Les Transformers présentent plusieurs avantages décisifs pour notre problématique. Le mécanisme d'auto-attention permet de capturer les dépendances à long terme entre frames distantes, essentiel pour comprendre des signes complexes. L'architecture parallélisable accélère l'entraînement sur de longues séquences vidéo. Enfin, les Transformers ont démontré leur supériorité dans les tâches de traduction, suggérant leur pertinence pour la traduction gestuelle.

Architecture globale du système

Notre architecture proposée comprend quatre composants principaux : un encodeur de caractéristiques vidéo, un encodeur Transformer temporel, un décodeur Transformer pour la génération de texte, et un module d'attention croisée reliant encodeur et décodeur.

Encodeur de caractéristiques vidéo

Pour chaque frame de la séquence vidéo d'entrée, nous utilisons MediaPipe pour extraire les caractéristiques des deux mains (42 points × 3 coordonnées). Ces vecteurs de dimension 126 sont enrichis par des caractéristiques additionnelles : vitesse et accélération calculées par différences finies entre frames consécutives, orientations relatives des segments de doigts, et distance entre les deux mains.

Le vecteur enrichi, de dimension approximative 150, est ensuite projeté dans un espace de dimension $d_{model} = 512$ via une couche linéaire dense. Un encodage positionnel sinusoïdal est ajouté pour injecter l'information de position temporelle.

Encodeur Transformer temporel

L'encodeur consiste en une pile de $N = 6$ couches identiques. Chaque couche comprend un mécanisme d'auto-attention multi-têtes avec $h = 8$ têtes d'attention, permettant au modèle de se concentrer simultanément sur différents aspects de la séquence.

L'auto-attention calcule, pour chaque position temporelle, une représentation pondérée de toutes les autres positions. Mathématiquement, pour une séquence d'entrée $X = (x_1, ..., x_T)$, l'attention est définie par :

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{1}$$

où Q, K, V sont respectivement les matrices de requêtes, clés et valeurs.

Chaque couche inclut également un réseau feed-forward position-wise avec une dimension intermédiaire de 2048, des connexions résiduelles et une normalisation par couche.

Décodeur Transformer pour génération de texte

Le décodeur génère la traduction textuelle de manière autoregressive. Il comprend également $N = 6$ couches, chacune contenant trois sous-modules : auto-attention masquée sur la séquence de sortie générée, attention croisée vers les représentations de l'encodeur, et réseau feed-forward.

L'auto-attention masquée assure que la prédiction d'un mot ne dépend que des mots précédemment générés, préservant la propriété autoregressive nécessaire pour la génération séquentielle.

L'attention croisée permet au décodeur de se concentrer sur les parties pertinentes de la séquence vidéo d'entrée pour générer chaque mot de la traduction.

Fonction de perte et optimisation

Le modèle est entraîné par minimisation de la perte d'entropie croisée entre les séquences prédites et les traductions de référence. Formellement, pour une séquence cible $y = (y_1, ..., y_L)$, la perte est :

$$\mathcal{L} = -\sum_{i=1}^{L} \log P(y_i | y_{<i}, X) \tag{2}$$

L'optimisation utilise l'algorithme Adam avec un taux d'apprentissage adaptatif suivant un échauffement linéaire suivi d'une décroissance proportionnelle à l'inverse de la racine carrée du nombre d'étapes.

Considérations pratiques:

Prétraitement des données

La construction d'un ensemble de données approprié constitue un défi majeur. Les vidéos doivent être segmentées en clips correspondant à des phrases ou énoncés complets. Chaque clip est annoté avec sa traduction textuelle correcte.

Un prétraitement robuste inclut la normalisation des longueurs de séquence par padding ou troncature, l'augmentation de données par variations d'échelle, rotation, et décalage temporel, et la gestion des cas où MediaPipe échoue à détecter les mains.

Gestion des séquences de longueur variable

Les Transformers peuvent traiter des séquences de longueurs variables grâce aux masques d'attention. Pour des raisons d'efficacité computationnelle, nous définissons une longueur maximale de séquence (par exemple, 200 frames) et utilisons un padding masqué pour les séquences plus courtes.

Stratégies de décodage

Pendant l'inférence, plusieurs stratégies de décodage sont envisageables. La recherche gloutonne sélectionne à chaque étape le mot le plus probable, simple mais sous-optimale. La recherche par faisceau (beam search) maintient les k séquences les plus probables, améliorant la qualité au coût de calculs additionnels.

Évaluation et métriques:

Métriques de traduction

L'évaluation des performances utilisera des métriques standard de traduction automatique. Le score BLEU (Bilingual Evaluation Understudy) mesure la similarité n-gramme entre traductions prédites et références. Le score METEOR considère synonymes et variations morphologiques. La distance d'édition de caractères (CER) et de mots (WER) quantifient les erreurs de substitution, insertion et suppression.

Évaluation humaine

Au-delà des métriques automatiques, l'évaluation humaine par des locuteurs natifs de la langue des signes reste essentielle. Des critères d'adéquation (le sens est-il préservé ?), de fluidité (la traduction est-elle naturelle ?) et de complétude seront évalués.

Défis et perspectives:

Défis techniques

Plusieurs défis techniques subsistent. La collecte de larges ensembles de données annotées de haute qualité demeure coûteuse. La variabilité entre signataires (vitesse, amplitude, dialectes régionaux) complique la généralisation. L'ambiguïté contextuelle de certains signes nécessite une compréhension sémantique profonde.

Extensions futures

Des extensions prometteuses incluent l'intégration d'informations faciales et corporelles complémentaires, la traduction bidirectionnelle (texte vers langue des signes), l'adaptation à différentes langues des signes nationales, et l'optimisation pour le déploiement sur appareils mobiles.

Conclusion: Ce rapport documente l'évolution de notre approche de traduction de la langue des signes, depuis un système initial de classification statique d'alphabets vers une architecture sophistiquée de modélisation séquentielle. Notre analyse critique a révélé que la traduction fidèle de la langue des signes nécessite la capture de dynamiques temporelles complexes, impossibles avec des méthodes frame-par-frame.

L'architecture Transformer proposée, combinant extraction robuste de caractéristiques via MediaPipe et modélisation séquentielle avancée, offre un cadre prometteur pour relever ce défi. Les mécanismes d'attention permettent de capturer les dépendances à long terme essentielles à la compréhension gestuelle, tandis que l'architecture séquence-à-séquence s'aligne naturellement sur la nature traductive du problème.

Les travaux futurs se concentreront sur l'implémentation complète, la collecte de données, et l'évaluation empirique de cette architecture. Le succès de cette approche pourrait significativement améliorer l'accessibilité et l'inclusion des communautés sourdes et malentendantes.

Par Arnold Mokira et Manel OURIR

E-mail: dr.mokira@gmail.com, GRAVO Hand

References

1 Vaswani, A., et al.: 'Attention is all you need', *Advances in Neural Information Processing Systems*, 2017, **30**, pp. 5998-6008
2 Lugaresi, C., et al.: 'MediaPipe: A Framework for Building Perception Pipelines', *arXiv preprint arXiv:1906.08172*, 2019
3 Camgoz, N.C., et al.: 'Neural Sign Language Translation', *IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7784-7793
4 Sutskever, I., Vinyals, O., Le, Q.V.: 'Sequence to sequence learning with neural networks', *Advances in Neural Information Processing Systems*, 2014, **27**, pp. 3104-3112
5 Breiman, L.: 'Random Forests', *Machine Learning*, 2001, **45**(1), pp. 5-32