Reducción de las variaciones intraclase y las similitudes entre clases en el reconocimiento del alfabeto ASL mediante el aprendizaje de similitud semántica

Atoany Nazareth Fierro-Radilla1*, Gibran Benitez-Garcia2† and Karina Ruby Perez-Daniel3†

1*Engineering and Sciences School, Tecnologico de Monterrey Cuernavaca campus, México 95D km 104, Xochitepec, 62790 , Morelos, Mexico.

2Graduate School of Informatics and Engineering, The University of Electro-Communications, Chofu, Tokyo, 182-8585, Tokyo, Japan.

3Engineering Faculty, Universidad Panamericana, 498 Augusto Rodin St., Mexico City, 03920 , CDMX, Mexico.

*Corresponding author(s). E-mail(s): afierror@tec.mx; Contributing authors: gibran@ieee.org; kperezd@up.mx; †These authors contributed equally to this work.

Resumen. El lenguaje de señas es un método importante que se utiliza para poder comunicarse entre la comunidad que padece alguna enfermedad auditiva, especialmente por personas con problemas del escucha y del habla. En los Estados Unidos, aproximadamente dos millones de personas que viven con discapacidad auditiva utilizan ASL (por sus siglas en ingles American Sign Language). Por lo tanto, el objetivo de este estudio es investigar y desarrollar un sistema de reconocimiento para el alfabeto ASL, haciendo uso de dos redes convolucionales (CNN), VGG16 y Mobilenet; para poder mejorar la comunicación y las relaciones interpersonales con las personas que viven con sordera. La propuesta se basa en implementar aprendizaje de semántica similar para reducir la variación intraclase y la alta similitud interclase en un espacio euclidiano de las imágenes con signos. Los resultados muestran que el sistema propuesto tiene una precisión promedio de 99 % y 90.1 % en el conjunto de datos MINIST y ASL, respectivamente. Al aumentar la codificación de imágenes de la misma clase y reducirla en diferentes clases y usando la técnica de análisis estadístico t-SNE, el reconocimiento del alfabeto ASL mejora considerablemente, lo cual se demuestra en este trabajo.

Abstract. Sign language is an important method used to communicate among deaf community, especially by people with hearing and speech impairments. In the United States, approximately two million people living with hearing impairment use ASL (American Sign Language). Therefore, the objective of this study is to investigate and develop a system for ASL alphabet recognition using two Convolutional Neural Networks (CNN); VGG16 and Mobilenet to improve communication and interpersonal relationships with people living with deafness. The proposal is based on implementing similar semantic learning to reduce intra-class variation and high inter-class similarity in a Euclidean space of signed images. The results show that the proposed system has an average accuracy of 99 % and 90.1 % in the MINIST and ASL dataset, respectively. By increasing the encoding of images of the same class and reducing it in different classes and using the t-SNE statistical analysis technique, ASL alphabet recognition is considerably improved, which is demonstrated in this work.

Introducción

La forma en la que nos conectamos físicamente con el mundo es a través de las manos; realizamos la mayoría de las tareas diarias con ellas, por otro lado, utilizamos dispositivos periféricos como mouse y el joystick para trabajar con una computadora (Interacción humana-máquina) [?], dicha interacción es un área de investigación multidisciplinaria con diversas aplicaciones en control de robótica, estudio de comportamiento psicológico, realidad virtual, reconocimiento de lengua de señas, visualización científica y, en el Metaverso [? ?], está constantemente definiendo nuevas modalidades de comunicación. El Re conocimiento de Lengua de Señas (RLS) se realiza por medio de la Interacción Humana-Computadora convirtiendo los gestos y movimientos en comandos de texto y/o de voz, permitiendo la comunicación entre los seres humanos a través de una computadora entre las personas que viven una discapacidad auditiva y las personas oyentes [? ?].

En los Estados Unidos (US) hay aproximadamente dos millones de personas sordas, algunos de ellos nacen con pérdida auditiva en ambos oídos, mientras que otros pierden la audición debido a factores como la rubéola y la meningitis [?]. El lenguaje de señas americano (ASL) es el segundo idioma distinto del inglés más utilizado en los EE. UU. después del español, tiene 36 formas de manos, 26 letras y alrededor de 6000 palabras, que consisten en movimientos corporales complejos. Las señas se crean usando la mano derecha, la mano izquierda, ambas manos y expresiones faciales y/o corporales [? ? ?].

A pesar de que ASL es el principal modo de comunicación para la mayoría de las personas sordas en EE. UU., sigue existiendo problemas de comunicación con las personas oyentes ya que no comprende el lenguaje ASL. Si la ASL se pudiera traducir automáticamente en texto o voz en inglés y/o español, será mucho más fácil para las personas sordas sentirse incluidos y tener mayor comunicación. [?].

Sistema de reconocimiento ASL

Hace más de dos décadas se han iniciado diversos trabajos de investigación para el reconocimiento del lenguaje de señas, con grandes avances, en especial para los [? ? ?] estadounidenses, [?] australianos, [?] los indios, los coreanos [?], el chino [?], el polaco [?] y el árabe [? ? ?], sin embargo, existen grandes dificultades debido a la complejidad de los movimientos de manos y cuerpo en expresiones en lengua de señas (LS) [?]. Los enfoques utilizados para resolver los problemas de SLR se pueden clasificar en dos métodos principales, los cuales están divididos en sensores e identificación de imagen [? ? ?]. En los enfoques basados en sensores, se suele usar un guante o sensor especial para rastrear la orientación, la posición, la rotación y los movimientos de la mano[? ? ?], que puede proporcionar información bastante precisa de la mano [? ?]. ¿Sin embargo, son demasiado pesados e incómodos para el uso diario [?]. Por otro lado, los métodos basados en la identificación de imagen se basan en técnicas de procesamiento de imágenes y aprendizaje automático para capturar y clasificar el movimiento del cuerpo y la forma de la mano usando imágenes en color sin necesidad de sensores conectados al ser humano [? ? ?]. Este documento se centra únicamente en el enfoque basado en la visión.

Deletreo

El deletreo con los dedos es la representación de cada letra del alfabeto mediante un signo. Los usuarios de ASL usan el alfabeto deletreado con los dedos y manos (AFA), se pueden tener diversas variaciones del alfabeto, así como diferentes señas que representen un significado distinto. El AFA consta de 22 formas de manos que cuando se mantienen en ciertas posiciones y/o se producen con ciertos movimientos, representan las 26 letras del alfabeto inglés [? ?]. Las investigaciones coinciden en que la ortografía digital está integrada en ASL de manera muy sistemática [? ?]. Uno de los usos es principalmente para representar nombres propios o palabras en inglés sin equivalentes en lenguaje de señas [? ? ?]. Además, la ortografía digital es una parte importante del lenguaje de señas para los nuevos usuarios y ayuda a las personas a abreviar señas más largas, a comunicar dos palabras compuestas [?] y a cerrar la brecha entre el léxico ASL a través de la geografía y las culturas [? ?]. La incorporación de la ortografía de los dedos en ASL resulta más conveniente en escenarios críticos, como es en ámbito de la medicina.

Retos en el Reconocimiento del alfabeto ASL

La clasificación de los signos depende de las configuraciones de las manos, las cuales son captadas por una cámara de color y/o profundidad. Las complejidades de los signos hacen que el reconocimiento del alfabeto ASL sea una tarea difícil debido a dos factores principales, la similitud entre clases y las variaciones dentro de las clases. La similitud entre clases significa que algunas letras están muy estrechamente relacionadas con otras y difieren muy poco en la ubicación de los dedos. Por ejemplo, las letras M y N solo se diferencian entre sí por si el pulgar está entre el primer y el segundo o entre el segundo y el tercer dedo. Las grandes variaciones intraclase significan que, dentro de una clase, existen diferencias entre las muestras, como la iluminación, el color de la piel, las variaciones del fondo y la posición relativa del firmante con respecto a la cámara.

Trabajos relacionados

El reconocimiento del alfabeto ASL se basa principalmente en dos subtareas: extracción de características y clasificación multiclase [?]. En el primer proceso se realiza la detección y extracción de características locales. Por otro lado, en este último proceso, estas características extraídas son comprendidas y caracterizadas para clasificar las muestras [?]. Para el reconocimiento del alfabeto ASL, en la literatura se pueden encontrar dos enfoques que se han utilizado: métodos basados en artesanía y basados en CNN.

Métodos hechos a mano

Hecho a mano significa que el algoritmo de extracción de características ha sido manualmente construido [31]. Para los enfoques tradicionales es necesario diseñar el mejor algoritmo de extracción de características que se adapte; además, una vez que la característica se selecciona, se debe elegir un clasificador de tal manera que se ajuste con la etapa de extracción de características.

Los primeros sistemas de reconocimiento del alfabeto ASL se proponen utilizando filtros de Gabor para la selección de características y random forest para la clasificación [27], el autor obtuvo un 49 % de precisión.

Métodos basados en redes neuronales CNN

En 2012, las redes neuronales convolucionales (CNN) se hicieron muy populares entre la comunidad de visión artificial debido al éxito de Alexnet en el desafío de reconocimiento visual ImageNet a gran escala (ILSVRC). Una de las ventajas de CNN sobre métodos artesanales es que las CNN extraen y clasifican características en el mismo algoritmo y no hay necesidad de intervención humana en ninguna de sus etapas. En otras palabras, la red aprende qué características son las mejores para ser extraídas para un mejor resultado de la clasificación. En general, una CNN se compone de capas convolucionales (compuesto por convolución y operaciones de agrupación máxima), totalmente conectado en capas. Los primeros obtienen representaciones no lineales de imágenes (característica extracción), y este último genera la incrustación y realiza la clasificación de las características de la imagen extraída.

Arquitectura Siamesa CNN

CNN se basa en analizar una gran cantidad de datos para funcionar bien. Sin embargo, es complicado para algunas aplicaciones obtener esa gran cantidad de muestras de entrenamiento. La generación del conjunto de entrenamiento para una red siamesa se basa en la selección aleatoria de un par de imágenes; este par de imágenes pueden ser positivas (ambas imágenes pertenecen a la misma clase) y negativas (muestras pertenecientes a diferentes clases).

Esto hace que las redes siamesas sean más robustas al desequilibrio de clases. Una red siamesa se basa en dos redes neuronales convolucionales idénticas que comparten sus parámetros y se utilizan para aprender similitudes semánticas y encontrar similitudes de los resultados comparando la incrustación que genera cada red [?]. La hipótesis de este artículo es que, usando una arquitectura siamesa, se puede reducir la alta similitud interclase y las altas variaciones intraclase.

Sistema propuesto para el Reconocimiento del Alfabeto ASL

Los experimentos se realizaron utilizando dos arquitecturas CNN diferentes, VGG16 y Mobilenet, así como sus versiones siamesas, en dos conjuntos de datos diferentes, del alfabeto y lenguaje de señas MNIST. El entrenamiento se hizo utilizando Keras y Tensorflow como marcos en una máquina Intel Core i7 con un GPU NVIDIA GeForce RTX 2070 SUPER.

Datasets

Para los experimentos, se utilizaron los conjuntos de datos del lenguaje de señas MNIST y ASL Alphabet. El conjunto de datos MNIST de lenguaje de señas [?] está compuesto por 34,627 imágenes de tamaño 28x28 píxeles divididas en 24 clases (de la A a la Z y que no contienen muestras para J y Z debido a movimientos de gestos). Por otro lado, el conjunto de datos de ASL Alphabet [?] está compuesto por 87,000 imágenes de 200x200x3 divididas en 29 clases (de la A a la Z) y 3 clases más etiquetadas como ”SPACE”, ”DEL” y ”NOTHING”; en este papel, "J 2 "Z" se consideran signos estáticos.

VGG16

VGG16 se propuso por primera vez en [?] y logró una precisión de prueba del 92,7 % entre los cinco primeros en ImageNet [?], que es un conjunto de datos compuesto por más de 14 millones de imágenes de 1000 clases diferentes.

Este modelo se envió a ILSVRC-2014, mejorando el rendimiento de Alexnet al reemplazar el tamaño de kernel grande con kernels 3x3. La función de grupo utilizada por la red VGG16 es una capa de grupo máximo de 2x2 con un paso de 2. Esta arquitectura tiene 3 capas densas seguidas de una capa softmax como salida. La principal contribución de VGG16 es que, en lugar de tener una gran cantidad de hiperparámetros, los autores se enfocaron en tener capas convolucionales de 3x3 con zancada (s = 1) y usar el mismo relleno.

En el entrenamiento de ambos conjuntos de datos, las imágenes se redimensionaron a 224x224x3 debido a los requisitos de forma de entrada de la red. La red se entrenó primero en el conjunto de datos del lenguaje de señas del MNIST durante 100 épocas.

Se utilizaron 24.720 imágenes como conjunto de entrenamiento y 2.735 para el conjunto de validación, logrando una precisión de entrenamiento de 0,9231, pérdida de entrenamiento de 5,3653, precisión de validación de 0,9872 y pérdida de validación de 0,3970.

Por otro lado, para el entrenamiento con el conjunto de datos ASL Alphabet, el número de épocas fue de 50 debido a la capacidad de la memoria. En este experimento se utilizaron 78.300 imágenes de 224x224x3 como conjunto de entrenamiento y 8.700 imágenes de 224x224x3 como conjunto de validación. Los resultados del entrenamiento fueron: precisión de entrenamiento de 0,9285, pérdida de entrenamiento de 5,5736, precisión de validación de 0,8477 y pérdida de validación de 16,2988.

Mobilenet

Mobilenet fue propuesto en [?]. Este modelo ligero utiliza convoluciones separables en profundidad, lo que reduce el número de parámetros en compa ración con las redes con convoluciones regulares con la misma profundidad.

Además, en Mobilenet, se realiza una sola convolución en cada canal de color en lugar de combinarlos y aplanarlos. Como aplica su nombre, Mobilenet está diseñado para ser utilizado en aplicaciones móviles. La forma de entrada de las imágenes fue de 224x224x3 para realizar los experimentos en las mismas condiciones que en VGG16. Para el conjunto de datos de lenguaje de señas del MNIST, se utilizaron 24.720 imágenes como conjunto de entrenamiento y 2.735 imágenes como conjunto de validación, obteniendo un valor de pérdida de entrenamiento de 0,0042, precisión de entrenamiento de 0,9989, pérdida de validación de 0,0003 y precisión de validación de 1,00. Para el entrenamiento del Alfabeto ASL, el conjunto de entrenamiento y el conjunto de validación están compuestos por 78.300 y 8.700 imágenes, respectivamente, obteniendo un valor de precisión de entrenamiento de 0,9947 y un valor de pérdida de entrenamiento de 0,0158; para validación, 0.9291 de precisión y 0.2754 para pérdida. Para el entrenamiento del Alfabeto ASL, el conjunto de entrenamiento y el conjunto de validación están

compuestos por 78.300 y 8.700 imágenes, respectivamente, obteniendo un valor de precisión de entrenamiento de 0,9947 y un valor de pérdida de entrenamiento de 0,0158; para validación, 0.9291 de precisión y 0.2754 para pérdida.

Siamese VGG16

Para el aprendizaje de similitud semántica, se utilizaron dos VGG16 tipo D idénticos; estas dos redes comparten sus parámetros. La etiqueta binaria indica la similitud del par de imágenes. El número de neuronas en la última capa es 1000 para ambos conjuntos de datos. Aquí, en el entrenamiento siamés, la última capa no contiene la probabilidad de que una imagen pertenezca a una determinada clase, sino una codificación para representar el contenido de la imagen en un espacio euclidiano. Cuanto mayor sea la dimensión de la codificación de la imagen, mejor será la representación de la imagen; sin embargo, las incrustaciones de imágenes grandes representan una mayor demanda de recursos informáticos. Después de varios experimentos, 1000 neuronas en la última capa mostraron el mejor rendimiento teniendo en cuenta la compensación entre la complejidad del cálculo, la precisión y las limitaciones del hardware.

En el caso del conjunto de datos de lenguaje de señas del MNIST, el tamaño de la imagen original es 28x28x1; sin embargo, el tamaño de entrada mínimo para VGG16 es 32x32x1 y, debido a que VGG16 es una red profunda, se decidió usar 64x64x1 como tamaño de entrada de imagen para para no perder información importante a través de las capas convolucionales. Los conjuntos de entrenamiento y validación están compuestos por 24.720 y 2.735 imágenes.

La versión siamesa de VGG16 entrenada en el conjunto de datos MNIST Sing Language logró 0,9861 de precisión de entrenamiento, 0,0199 de pérdida de entrenamiento, 0,9834 de precisión de validación y 0,0237 de pérdida de validación.

Siamese Mobilenet

Se utilizaron dos redes Mobilenet idénticas con la misma arquitectura utilizada en la sección 4.3. La última capa contiene 1000 neuronas. Debido a limitaciones de hardware, las imágenes se redimensionaron a 64x64 para ambos conjuntos de datos; los hiperparámetros son

básicamente los mismos que se usaron para el siamés VGG16, excepto por la cantidad de épocas. Para el conjunto de datos de lenguaje de señas del MNIST, el conjunto de entrenamiento y validación estuvo compuesto por 24 709 y 2746 imágenes, respectivamente. Los resultados del entrenamiento son los siguientes: pérdida de entrenamiento de 4.42E-4, precisión de entrenamiento de 1.0, pérdida de validación de 0.0061 y precisión de validación de 1.0.

Por otro lado, para el conjunto de datos de ASL Alphabet, solo se usó el 10% del conjunto de datos completo debido a limitaciones de hardware; el conjunto de entrenamiento estuvo compuesto por 7.830 imágenes mientras que el conjunto de validación de 870 imágenes. Los resultados del entrenamiento son los siguientes: pérdida de entrenamiento de 0,0064, precisión de entrenamiento de 0,9925, pérdida de validación de 0,0102 y precisión de validación de 0,9888.

Resultados experimentales

El reconocimiento del alfabeto ASL se realiza mediante una tarea de clasificación, y para ello se utilizaron los modelos generados luego del entrenamiento de las redes mencionadas anteriormente. Para los modelos VGG16 y Mobile net, la tarea de clasificación tiene como base alimentar una imagen del conjunto de prueba (muestras nunca vistas por la red) y dejar que la red prediga el alfabeto. Las predicciones se compararon con los ejemplos reales. El rendimiento de la clasificación se mide utilizando la matriz de confusión y las métricas más comunes utilizadas para la evaluación de la clasificación, como la exactitud, la precisión, la recuperación y la puntuación F1.

Por otro lado, para las arquitecturas siamesas, la tarea de clasificación se realiza por incrustaciones de imágenes cuyos elementos dependen de la similitud de la imagen de entrada de entrenamiento. Para la evaluación se alimentó al modelo con las imágenes de entrenamiento se para obtener un centroide de clase calculando el promedio de todos estos vectores, posteriormente para la predicción de clase, se utiliza una imagen para una red para obtener su imagen incrustada. Finalmente, se hizo el cálculo de la distancia entre esta imagen incrustada y cada centroide de clase utilizando la misma distancia métrica del entrenamiento.

Con el fin de mejorar la hipótesis de que el aprendizaje de similitud semántica podría mejorar el reconocimiento del alfabeto ASL, se realizó un análisis estadístico utilizando t-SNE (incrustación de vecinos estocásticos distribuidos en t). La técnica t-SNE es un método de reducción de dimensionalidad no lineal muy adecuado para incrustar datos de alta dimensión para su visualización en un espacio de baja dimensión (en este caso, en un espacio euclidiano 2D).

Conclusiones

El lenguaje de señas es la forma en que las personas sordas se comunican con los demás, sin embargo, la mayoría de las veces, las personas oyentes no conocen este idioma. Por lo tanto, existe una brecha de comunicación que impacta negativamente a la comunidad sorda. Por lo tanto, en este artículo, se presenta un método novedoso para el reconocimiento del alfabeto ASL.

Uno de los mayores desafíos en el reconocimiento del alfabeto ASL es la gran variación intraclase y la gran similitud entre clases entre las imágenes. Los métodos tradicionales luchan por encontrar patrones para la clasificación alfabética de ASL. Para solucionar esto, se planteó la hipótesis de que si era posible generar patrones según la similitud de las imágenes, se podría mejorar el reconocimiento del alfabeto ASL. Para ello, se implementó un aprendizaje de similitud semántica utilizando redes siamesas, que en pocas palabras, son dos redes idénticas que comparten sus parámetros. Los experimentos muestran que los vectores de características generados por las arquitecturas siamesas son mejores representaciones de imágenes al tener en cuenta las similitudes y diferencias entre ellas. Otro hallazgo en los experimentos fue que a pesar de que el conjunto de entrenamiento para la CNN siamesa fue mucho menor en comparación con el conjunto de entrenamiento utilizado en una sola CNN, las redes siamesas no presentaron sobreajuste, esto se debe a la ayuda del aprendizaje de un disparo. El aprendizaje de una sola vez permite que la red aprenda de solo unas pocas imágenes por clase. Además, las redes siamesas son resistentes al desequilibrio de clases debido a que, al final del día, la red intenta aprender solo dos clases, pares de imágenes similares y no similares. El tiempo de entrenamiento y los requisitos de memoria de hardware de las redes siamesas son los mayores inconvenientes porque involucra pares cuadráticos de muestras para aprender.