ISSN 1870-4069

Reconocimiento de las señas estáticas del LSM con características basadas en aprendizaje profundo

Rafael Fernández Rodríguez, Francisco Javier Peralta Rosas, Luis Ángel Zuñiga-Madrid, Pedro Arguijo

Tecnológico Nacional de México,
Campus Misantla,
México

{rafael.fernandez.rgz, francisco.peraltarosas, angel.zunigamadrid}@gmail.com, pedro_arguijo@excite.com

Resumen. El lenguaje de señas es el método de comunicación entre personas que sufren de problemas del habla y de la audición. Los métodos de extracción de características juegan un papel importante en la obtención de una alta tasa de reconocimiento. En este trabajo se evalúan las características profundas extraídas por la arquitectura VGG16 tanto en la capa fc6 y fc7, así como la concatenación de ambas para identificar las señas estáticas del lenguaje de señas mexicano. Para el reconocimiento de las señas se utilizaron los clasificadores de RF y SVM. Los resultados generales, en la clasificación de las veintiuna señas estáticas del lenguaje de señas mexicano, fueron del 92.5% y 95.32% para RF y SVM, respectivamente, con las características extraídas en la capa fc6.

Palabras clave: Reconocimiento de señas, Transferencia de aprendizaje, bosques aleatorios, máquinas de soporte vectorial.

LSM Static Sign Recognition with Deep Learning-Based Features

Abstract. Sign language is the method of communication between people who suffer from speech and hearing problems. Feature extraction methods play an important role in obtaining a high recognition rate. In this paper we evaluate the deep features extracted by the VGG16 architecture in both fc6 and fc7 layers, as well as the concatenation of both to identify the static signs of Mexican sign language. RF and SVM classifiers were used for sign recognition. The overall results, in the classification of the twenty-one static Mexican sign language signs, were 92.5% and 95.32% for RF and SVM, respectively, with the features extracted in layer fc6.

Keywords: Sign recognition, transfer learning, random forest, support vector machine.

1. Introducción

El reconocimiento de gestos es por demás significante en la interacción humano-computadora tales como la realidad virtual, juegos interactivos, procesamiento de imágenes o el reconocimiento del lenguaje de señas. El lenguaje de señas es una forma de comunicación, no verbal, utilizada por personas con problemas de audición y habla para expresar sus pensamientos y emociones. Los principales componentes del lenguaje de señas son los signos manuales y no manuales. Los signos manuales son: posición, orientación, forma y trayectoria de la mano. Mientras que los no manuales representan el movimiento del cuerpo y las expresiones faciales.

Aunque la información más importante se transmite a través de las manos, los signos no manuales permiten aclarar y enfatizar el significado de los signos manuales. Existen dos enfoques que se utilizan comúnmente para interpretar los gestos o señas en la interacción humano-computadora. El enfoque basado en guantes de datos [1, 2] se basa en dispositivos electromecánicos conectados a un guante para digitalizar los movimientos de las manos y los dedos en datos multi-paramétricos.

El principal problema de esta implementación es que requiere llevar dispositivos, los cuales son costosos y provoca comportamientos menos naturales. El segundo basado en visión artificial se divide a su vez en: enfoque basado en el modelo 3D de la mano y el basado en la apariencia. El enfoque del modelo 3D de la mano se basa en el modelo cinemático 3D de la mano e intenta estimar los parámetros de ésta mediante la comparación entre las imágenes de entrada y la posible apariencia 2D proyectada por el modelo de la mano 3D.

El enfoque basado en la apariencia utiliza características extraídas de la imagen RGB para modelar la apariencia visual de la mano y comparar estos parámetros. Como se mencionó, la mayoría de los modelos basados en apariencia se basan en extraer características que representan el contenido de las imágenes. Para hacer frente a la dependencia del punto de vista, estos métodos deben tener propiedades de invariancia a los cambios de traslación, rotación y escala.

Estas características pueden estar basadas en regiones como los momentos de Hu, momentos de Zernike o de Jacobi-Fourier [3–9], o basadas en contornos como los descriptores de Fourier [10, 11] o Histograma de Gradientes Orientados [7]. También se han utilizado características tipo Haar 3D extraídas de imágenes de profundidad [12]. Dichas características se han clasificado con redes neuronales artificiales o máquinas de soporte vectorial.

Sin embargo, se debe mencionar que sin importar el método de extracción de características se debe resolver el problema de segmentación de la mano que es un proceso complejo. Recientemente, el aprendizaje profundo mediante redes neuronales convolucionales (CNN) ha obtenido mucho éxito en tareas de reconocimiento visual como la clasificación de imágenes y la detección de objetos [13-15].

Dado que las características adquiridas a partir de estas redes neuronales son bastante potentes, es popular tratar estas CNN, entrenadas en grandes conjuntos de datos de imágenes naturales, como extractores genéricos de características; a este proceso se le conoce como transferencia de aprendizaje. Al reutilizar el conocimiento adquirido en tareas relacionadas, se vuelve más fácil abordar tareas más exigentes, como la recuperación de imágenes, la segmentación semántica y el reconocimiento de emociones, por mencionar algunas.

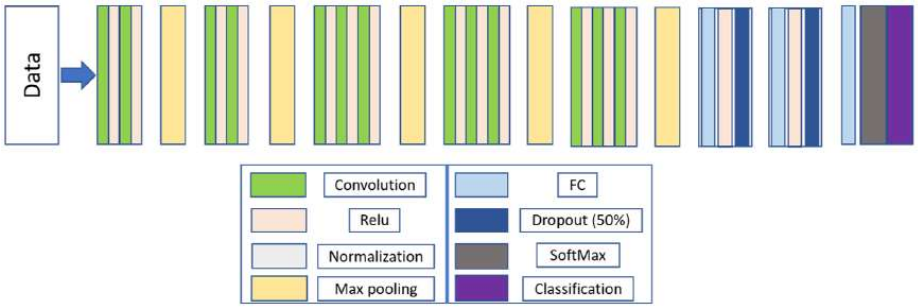

Fig. 1. Arquitectura de VGG16.

De acuerdo con lo mencionado por el Consejo Nacional para el Desarrollo y la Inclusión de las Personas con Discapacidad, la Lengua de Señas Mexicana (LSM), es la lengua que utilizan las personas sordas en México. Como toda lengua, posee su propia sintaxis, gramática y léxico [16]. Se compone de signos visuales con estructura lingüística propia y consta de 27 señas, de las cuales 21 son estáticas y las restantes dinámicas.

Este trabajo se centra en la clasificación de las señas estáticas del LSM. Proponemos extraer características profundas en dos capas completamente conectadas, fc6 y fc7, de la arquitectura VGG16 y realizar la clasificación de estas con Bosques Aleatorios (Random Forest, RF) y Máquinas de Soporte Vectorial (Support Vector Machine, SVM) tanto de manera individual como concatenando ambos conjuntos de características.

El estudio está organizado de la siguiente manera. Si bien la transferencia de aprendizaje y la arquitectura VGG16 se aborda en el Sección 2, la descripción de los clasificadores RF y SVM se dan en la sección 3. El conjunto de datos utilizado se describe en la Sección 4. En la Sección 5 se discuten los resultados. Finalmente, en la Sección 6, se presentan algunas observaciones finales.

2. Transferencia de aprendizaje

La transferencia de aprendizaje (transfer learning) es un ingrediente esencial para la inteligencia artificial, donde el conocimiento aprendido en un dominio para alguna tarea puede transferirse a otro dominio para una tarea diferente [17]. En el contexto del aprendizaje profundo, la transferencia de aprendizaje comúnmente se implementa en una red neuronal convolucional profunda, CNN, entrenada previamente en un gran conjunto de datos etiquetados.

Este enfoque se ha aplicado con éxito en muchas áreas de la visión artificial. Las capas iniciales de una CNN extraen características genéricas simples (bordes, esquinas, curvas, manchas de color, etc.), que son aplicables a todo tipo de imágenes, mientras que las capas finales representan características muy abstractas y específicas de los datos.

Por lo tanto, se espera que el uso de un modelo entrenado de forma óptima en un gran conjunto de datos y su posterior ajuste en un conjunto de datos diferente dé como

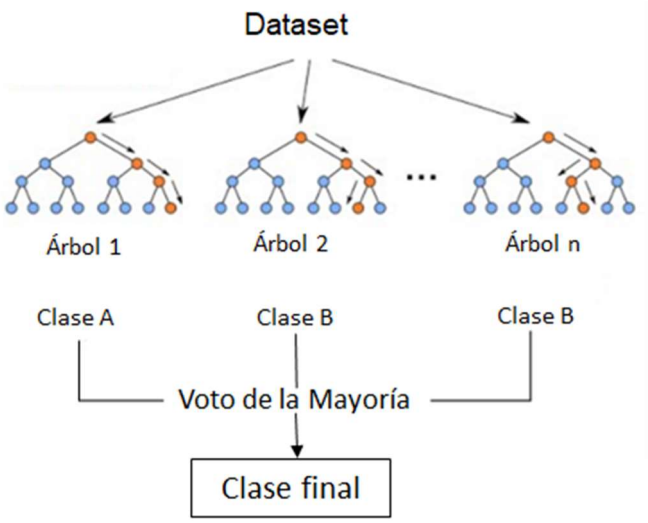

Fig. 2. Estructura de Random Forest.

resultado una mayor precisión y un proceso de entrenamiento más rápido, en comparación con el entrenamiento de las mismas CNN desde cero, debido a las características compartidas presentes en las capas iniciales. Existen dos enfoques que se pueden utilizar en la transferencia de aprendizaje:

– Extracción de características: Consiste en utilizar las características de una red previamente entrenada para representar las imágenes de nuevos conjuntos de datos. Estas características se utilizan para entrenar un nuevo clasificador. Al utilizar la CNN como extractor de características, se elimina la última capa totalmente conectada, que es la capa de salida.

Los valores de las características se pueden extraer como valores sin procesar o después de haber sido transformados por la función ReLU, donde una salida x se asigna a $\max(0, x)$. Luego, se utilizan estas características profundas para entrenamiento y clasificación.

– Ajuste fino: El ajuste fino de la red pre-entrenada es relevante cuando el conjunto de datos objetivo es muy grande. Consiste en descongelar algunas de las capas superiores de un modelo congelado que se utiliza para la extracción de características. Además, los parámetros de todas las capas (excepto las últimas) de la red pre-entrenada se inicializan, el entrenamiento se realizará más rápidamente que si la inicialización hubiera sido aleatoria.

VGG16 [18] es una arquitectura simple y ampliamente utilizada para ImageNet. Toma como entrada una imagen de 224 × 224 pixeles y devuelve un vector de tamaño 1000 con las probabilidades de pertenecer a cada clase. VGG16 contiene 13 capas de convolución, 3 capas completamente conectadas (fc6, fc7 y fc8) y 5 capas de agrupación (como se muestra en la Fig. 1).

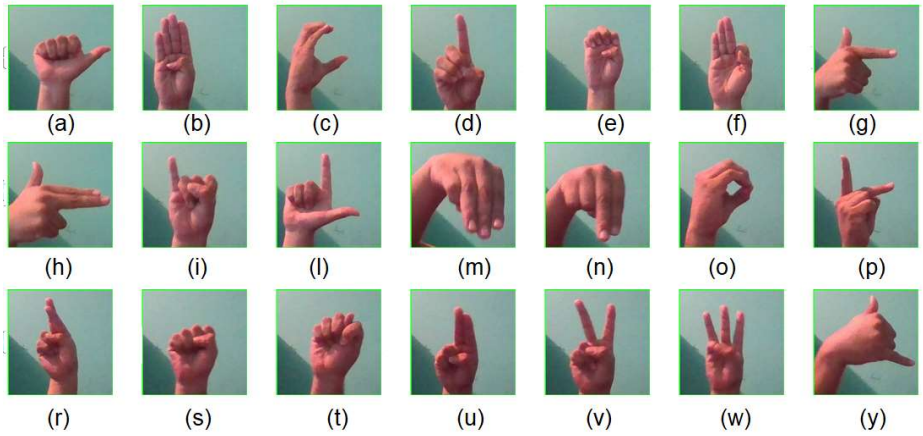

Fig. 3. Muestra representativa de las señas estáticas del LSM utilizado. Debajo de cada seña se indica a que letra corresponde.

Las 16 capas convolucionales se utilizan para extraer características de la imagen. En cada capa de convolución hay un filtro múltiple de 3x3, con 1 píxel como paso. La última capa softmax se utiliza para la clasificación. En cada capa de convolución, ReLU se aplica como función de activación. En este trabajo utilizamos la arquitectura VGG16 para extraer vectores de características profundas en las capas de activación fc6 y fc7, en ambos casos son características pre ReLU, las cuales individualmente generan un vector de 4096 características.

Seleccionamos VGG16 porque se ha aplicado ampliamente en diversas tareas de clasificación. Se trata de una de las primeras arquitecturas que exploran la profundidad de la red al ampliarla a 16 capas y utilizar filtros de convolución muy pequeños (3×3). Además, al haber sido entrenada en el conjunto de datos de Imagenet es un extractor de características capaz de generalizar muy bien las imágenes y puede aplicarse a una amplia variedad de tareas como la clasificación de objetos, la detección, la localización, etc.

3. Clasificadores

La clasificación es una tarea que requiere el uso de algoritmos de aprendizaje automático para asignar una clase a ejemplos no vistos previamente del problema. Identificar la categoría o clase a la que pertenece un nuevo dato es el objetivo de un problema de clasificación. A continuación, describimos brevemente los clasificadores utilizados en este trabajo.

3.1. Bosques aleatorios (Random Forest)

Los algoritmos de bosque aleatorio (RF) forman una familia de métodos de clasificación que se basan en la combinación de varios árboles de decisión no podados (ver Fig. 2). La particularidad de tales ensambles de clasificadores es que sus componentes basados en árboles se obtienen a partir de un cierto grado de aleatoriedad.

Tabla 1. Exactitud obtenida con RF en el conjunto de entrenamiento. Las razones de entrenamiento y prueba son las indicadas.

	fc6	fc7	fc6 fc7
10/90	0.8772	0.8026	0.87
20/80	0.925	0.9272	0.7835
30/70	0.9249	0.919	0.9259

En base a esta idea, RF se define como un principio genérico de conjuntos aleatorios de árboles de decisión [19]. La unidad básica de RF (la denominada aprendiz base) es un árbol binario construido utilizando particiones recursivas (RPART). RF Ofrece un rendimiento excelente en una serie de problemas prácticos, ya que no es sensible al ruido en el conjunto de datos y no está sujeto a un sobreajuste.

Los árboles de clasificación en el Bosque Aleatorio se construyen recursivamente utilizando el criterio de impureza Gini que se utiliza para determinar las divisiones en la variable predictiva. La división de un nodo del árbol se realiza sobre la variable de manera que reduce la incertidumbre presente en los datos y por lo tanto la probabilidad de una clasificación errónea. La división ideal de un nodo del árbol ocurre cuando es cero el valor de Gini.

El proceso de división continúa hasta que se crea un "bosque", formado por múltiples árboles. La clasificación se produce cuando cada árbol del bosque emite un voto para la clase más popular. El bosque aleatorio entonces elige la clasificación que tiene más votos sobre todos los árboles del bosque. La poda no es necesaria ya que cada clasificación se produce por un bosque final que consiste en árboles generados independientemente creados a través de un subconjunto aleatorio de datos, evitando el sobreajuste.

Las tasas de error de generalización dependen de la fuerza de los árboles individuales en el bosque y la correlación entre ellos. Esta tasa de error converge a un límite a medida que aumenta el número de árboles en el bosque. Otra ventaja de RF es que no hay validación cruzada o un conjunto de pruebas separado para obtener una estimación imparcial del error del conjunto de pruebas.

La precisión del conjunto de prueba se estima internamente en RF al ejecutar muestras fuera de bolsa (OOB). Por cada árbol que crece en RF, aproximadamente un tercio de los casos están fuera de bolsa (fuera de la muestra de bootstrap). Las muestras fuera de bolsa (OOB) pueden servir como un conjunto de prueba para el árbol cultivado en los datos que no son OOB.

3.2. Máquinas de soporte vectorial (Support Vector Machine)

Las máquinas de soporte vectorial (SVM) son una técnica de reconocimiento supervisado, desarrollada originalmente para clasificar clases de objetos linealmente separables. En la SVM, el hiperplano óptimo o frontera de decisión se determina en función de un pequeño subconjunto de todos los ejemplos de entrenamiento que maximizan la separación entre conjuntos; o sea, la capacidad de clasificación entre clases, denominados vectores de soporte.

Tabla 2. Exactitud obtenida con SVM en el conjunto de entrenamiento. Las razones de entrenamiento y prueba son las indicadas.

	fc6	**fc7**	**fc6 fc7**
10/90	0.9192	0.916	0.9333
20/80	0.9532	0.9482	0.9556
30/70	0.9535	0.9608	0.951

Si los datos de entrenamiento no se pueden separar linealmente, es decir, si no es posible construir un hiperplano en el espacio original, la separación se realiza en un espacio dimensional superior.

Para fines prácticos, es habitual utilizar una función de mapeo no lineal llamada kernel, seleccionada entre funciones polinómicas, funciones de base radial, funciones de base radial gaussiana y funciones sigmoidales. SVM se diferencia de otras técnicas como las redes neuronales artificiales, ya que el problema de mínimos locales no afecta a SVM, porque su entrenamiento se basa en problemas de optimización convexa.

4. Conjunto de datos

El conjunto de datos utilizado en este trabajo es de acceso libre [7] y contiene imágenes de las veintiuna señas estáticas del LSM. Consta de 300 imágenes por seña, además cada imagen representa la señal de la letra con una distinta variación de acuerdo con rotación, traslación y escalamiento. En la Fig. 3 se muestra una imagen representativa de cada una de las señas. En la captura de las imágenes se utilizó un fondo verde para facilitar su segmentación.

5. Resultados

El método propuesto consiste en dos etapas. Primero, todas las imágenes del LSM se redimensionaron a 224×224 píxeles de acuerdo con el requisito de entrada de la del modelo pre-entrenado VGG16. Las características profundas de extrajeron de las capas fc6 y fc7 y por capa se obtuvo un vector de 4096 dimensiones para cada imagen.

También se realizó la concatenación de ambos conjuntos de características para formar una única representación de características profundas (8192 dimensiones por imagen). Como segunda etapa se realizó el entrenamiento de los clasificadores RF y SVM para determinar la etiqueta correspondiente a las señas de los conjuntos de características extraídos en fc6, en fc7 y la concatenación de ambas.

La metodología indicada se realizó con un procesador i7-CPU, 8GB de RAM y sistema operativo de 64 bits. La implementación de un proceso de aprendizaje puede generar procedimientos que requieren mucho tiempo y recursos, dependiendo de la calidad de los datos disponibles. Debido a la cantidad total de imágenes en el dataset, para los experimentos se seleccionaron aleatoriamente tres distintas proporciones para la parte de entrenamiento y pruebas de ambos clasificadores, dichas proporciones fueron: 10/90, 20/80 y 30/70.

Consideramos estas proporciones para evaluar su influencia en la exactitud teniendo en cuenta las características estudiadas y, además, con la finalidad de comparar

resultados previamente reportados con este dataset [7]. Es importante señalar que un muestreo aleatorio de la partición de entrenamiento y prueba puede producir resultados diferentes en diferentes ejecuciones.

Las Tablas 1 y 2 muestran la exactitud obtenida con el conjunto de características de pruebas **seleccionado** para los clasificadores RF y SVM, respectivamente. En ambas tablas se indica la razón de entrenamiento/prueba, así como la capa de extracción considerada: fc6, fc7 y la concatenación de ambas fc6 fc7.

Como se puede observar en ambas tablas la exactitud en la clasificación con las características concatenadas son mayores que los obtenidos con las características de la capa fc7; con la única excepción que se muestra en la Tabla 1 para la razón 20/80.

Sin embargo, la exactitud de las características de la capa fc6 y las características concatenadas son similares. Lo cual implica que se obtiene una buena clasificación únicamente con las características de la capa fc6. Aunque los resultados presentados en las Tablas 1 y 2 muestran que SVM tiene el mejor desempeño se debe mencionar que requiere un mayor tiempo de entrenamiento que RF.

De acuerdo con los resultados publicados previamente que consideran el mismo conjunto de imágenes para el reconocimiento del LSM, nuestros resultados están dentro del rango a pesar de la gran dimensionalidad de las características consideradas.

6. Conclusiones y trabajo a futuro

Se propuso un sistema de reconocimiento del LSM basado en la extracción de características profundas. Las características extraídas con la arquitectura VGG16 se clasificaron con RF y SVM. Se utilizaron las capas fc6 y fc7 para la extracción de características, obteniendo un vector con 4096 datos por imagen.

En los tres experimentos considerados para ambos clasificadores se observó que la mejor clasificación se obtiene con las características de la capa fc6. Como se esperaba, las razones de entrenamiento/pruebas juegan un rol importante para derivar un sistema robusto de clasificación.

Los mejores resultados de clasificación se obtuvieron con SVM. Como trabajo futuro se propone realizar una comparación del desempeño de las características extraídas por diversas arquitecturas profundas previamente entrenadas. De igual manera, debido a la cantidad de imágenes con las que cuenta el dataset, implementar la transferencia de aprendizaje por ajuste fino.

Referencias

1. Saldaña-González, G., Cerezo-Sánchez, J., Bustillo-Díaz, M. M., Ata-Pérez, A.: Recognition and classification of sign language for spanish. Computación y Sistemas, vol. 22, no. 1 (2018) doi: 10.13053/cys-22-1-2780
2. Maitre, J., Rendu, C., Bouchard, K., Bouchard, B., Gaboury, S.: Basic daily activity recognition with a data glove. Procedia Computer Science, vol. 151, pp. 108–115 (2019) doi: 10.1016/j.procs.2019.04.018
3. Otiniano-Rodríguez, K., Camara-Chavez, G., Menotti, D.: Hu and Zernike moments for sign language recognition. In: Proceedings of the International Conference on Image Processing, Computer Vision, and Pattern Recognition, pp. 918–922 (2012)

4. Martinez-Perales, J. C., Flores-Carapia, R., Luna-Benoso, B.: An automated method for identification of vowels on the sign language. Contemporary Engineering Sciences, vol. 8, pp. 1499–1508 (2015) doi: 10.12988/ces.2015.59278
5. Solís, F., Hernández, M., Pérez, A., Toxqui, C.: Static digits recognition using rotational signatures and Hu moments with a multilayer perceptron. Engineering, Scientific Research Publishing, Inc, vol. 6, no. 11, pp. 692–698 (2014) doi: 10.4236/eng.2014.611068
6. Perez, L. M., Rosales, A. J., Gallegos, F. J., Barba, A. V.: LSM static signs recognition using image processing. In: Proceedings of the 14th International Conference on Electrical Engineering, Computing Science and Automatic Control, IEEE (2017) doi: 10.1109/iceee.2017.8108885
7. Mancilla-Morales, E., Vázquez-Aparicio, O., Arguijo, P., Meléndez-Armenta, R. Á., Vázquez-López, A. H.: Traducción del lenguaje de señas usando visión por computadora. Research in Computing Science, vol. 148, no. 8, pp. 79–89 (2019)
8. Joshi, G., Vig, R., Singh, S.: Analysis of Zernike moment-based features for sign language recognition. Advances in Intelligent Systems and Computing, Springer Singapore, pp. 1335–1343 (2018) doi: 10.1007/978-981-10-5903-2_140
9. Solís, F., Toxqui, C., Martínez, D.: Mexican sign language recognition using Jacobi-Fourier moments. Engineering, Scientific Research Publishing, Inc, vol. 7, no. 10, pp. 700–705 (2015). doi: 10.4236/eng.2015.710061
10. Nur Fauzan, M. H., Rakun, E., Hardianto, D.: Feature extraction from smartphone images by using elliptical Fourier descriptor, centroid and area for recognizing indonesian sign language SIBI (Sistem Isyarat Bahasa Indonesia). In: Proceedings of the 2nd International Conference on Intelligent Autonomous Systems (ICoIAS), IEEE (2019) doi: 10.1109/icoias.2019.00008
11. Kishore, P. V. V., Prasad, M. V. D., Prasad, C. R., Rahul, R.: 4-Camera model for sign language recognition using elliptical Fourier descriptors and ANN. In: Proceedings of the International Conference on Signal Processing and Communication Engineering Systems, IEEE (2015) doi: 10.1109/spaces.2015.7058288
12. Jimenez, J., Martin, A., Uc, V., Espinosa, A.: Mexican sign language alphanumerical gestures recognition using 3D Haar-like features. IEEE Latin America Transactions, vol. 15, no. 10, pp. 2000–2005 (2017) doi: 10.1109/tla.2017.8071247
13. Krizhevsky, A., Sutskever, I., Hinton, G. E.: Imagenet classification with deep convolutional neural networks. In: Proceedings of Advances in Neural Information Processing Systems 25, pp. 1097–1105 (2012)
14. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition (2014) doi: 10.48550/ARXIV.1409.1556
15. Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A.: Going deeper with convolutions. pp. 1–9 (2014) doi: 10.48550/ARXIV.1409.4842
16. Gobierno de México. Lengua de Señas Mexicana (LSM) (2022) www.gob.mx/conadis/articulos/lengua-de-senas-mexicana-lsm?idiom=es
17. Yosinski, J., Clune, J., Bengio, Y., Lipson, H.: How transferable are features in deep neural networks? In: Proceedings of the Advances in Neural Information Processing Systems 27, pp. 3320–3328 doi: 10.48550/ARXIV.1411.1792
18. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. Computer Science, Computer Vision and Pattern Recognition (2015) doi: 10.48550/arXiv.1409.1556
19. Breiman, L.: Random forests. Machine Learning, Springer Science and Business Media LLC, vol. 45, no. 1, pp. 5–32 (2001) doi: 10.1023/a:1010933404324