Visual Cognate Analysis Using Convolutional Neural Networks

James Lydon
Department of Computer Science
Drexel University
Philadelphia, PA, USA
JPL337@drexel.edu.com

ABSTRACT

This research analyzes English, German, Spanish, and Italian vocabulary in order to establish patterns of etymological history between languages. The analysis is performed using a convolutional neural network to examine the visual representation of selected cognates. Results demonstrate a correlation between cognate relationships and the neural network matching, suggesting the possibility that neural networks could be used to find presently unknown root origins of some words.

1 Introduction

Languages evolve with minds of their own and there is significant linguist interest in tracing the family relationships of languages. The phylogenetic trees describing modern, widely-spoken languages are fairly well understood. However in some specific cases the relationships are not so clear.

The field of etymology is particularly interested in cognates – words of related origins which often have similar phonetic sounds or visual appearances. Cognates possess the ability to seamlessly pass between any language to any other language, even ignoring the general family tree of those languages. There are some words which have no known cognates with any word in any language. These words offer linguists a scavenger hunt in researching possible origins.

In recent years convolutional neural networks (CNNs) have been used to solve a wide variety of problems. This is particularly true in visual analysis where patterns between objects can be perceived with a much higher accuracy than the human eye.

The following research attempts to use convolutional neural networks to find visual patterns in cognates. First, we will build a CNN which given any word should determine with a high degree of accuracy what language that word belongs to. Second, we will input words with known cognate relationships into the CNN and analyze whether the CNN outputs a high value toward the language of the known cognate parent-languages. Third, we will attempt to input words with no known etymological origin and discuss the most likely parent-language linguists might find a cognate origin.

2 Related Work

The primary inspiration for this work is the convolutional neural nettwork analysis performed by Daggumati and Revesz on ancient syllabaries, another field of significant linguistic interest. The goal of their research was to further evidence known relationships of ancient languages and postulate on the possible relationships of those languages with debated origins.[1]

The implementation of the convolutional neural network within this paper uses CNN hyperparameters as described and recommended by Daggumati and Revesz. Further, the general idea of inputting a language's symbols/letters into a CNN and observing which languages result in the highest softmax value is similarly implemented in this research with general vocabulary of a language as input instead.

3 Implementation

The first goal is to create a convolutional neural network which can determine with a high degree of accuracy what language a given word belongs to. For this research we have chosen four languages to train the CNN to recognize: English, German, Spanish, and Italian. English is the primary point of interest and possesses an extraordinary number of cognates with all three others. German and English are both "Germanic languages" and so we can expect a large number of cognates between the two. Spanish and Italian, which are both "Romantic languages" are expected to share a large number of cognates between themselves, but share very little with German. Thus, the four languages together should provide an interesting nexus of cognate relationships to analyze.

3.1 Data. To train the model we first need a large dataset of words in each of the four languages. These words should represent a wide vocabulary in order to fully capture the most significant patterns of a language. As we are analyzing the visual representation of cognates rather than the phonetic representations, we need each word to be represented in a consistent image form. We would also like to try not to over-represent one language's vocabulary more than any other language's vocabulary.

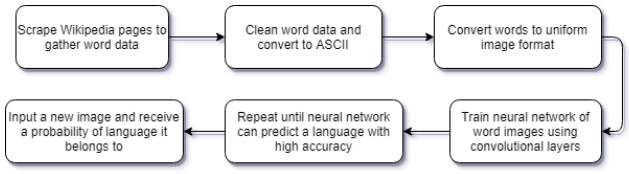

Fig 1. Flow of the Visual Cognate Analysis Application

To gather the dataset, we use the Wikipedia Python module to scrape words from Wikipedia pages of similar content and length length across each language. A random sampling of Wikipedia pages will not be sufficient. In some languages a given Wikipedia page may be extremely lengthy and descriptive and in other languages the same page may be small or non-existent. We choose some pages such as Philosophy, Religion, etc. which are adequately-written in each of the four language. Then we choose pages for large cities or which primarily speak the given language as well as current politicians and significant historical figures who are of interest in the given language.

We scrub the results we gather so that all words are converted to lowercase ASCII. This removes anomalies such as diacritics, umlauts, etc. which might immediately reveal the language origin of a word without any analysis on the letter arrangement. We also cut out words which are too long as they will also provide us with certain anomalies, especially in German. Doing this we obtain a total of 10,000 – 15,000 words for each language.

We convert each of the words one-by-one to images using the Pillow package in Python. Each word is represented as an 84x84 image, which appears to be just about enough width to represent every word and is symmetrical in height for more convenient analysis in our convolutional neural network.

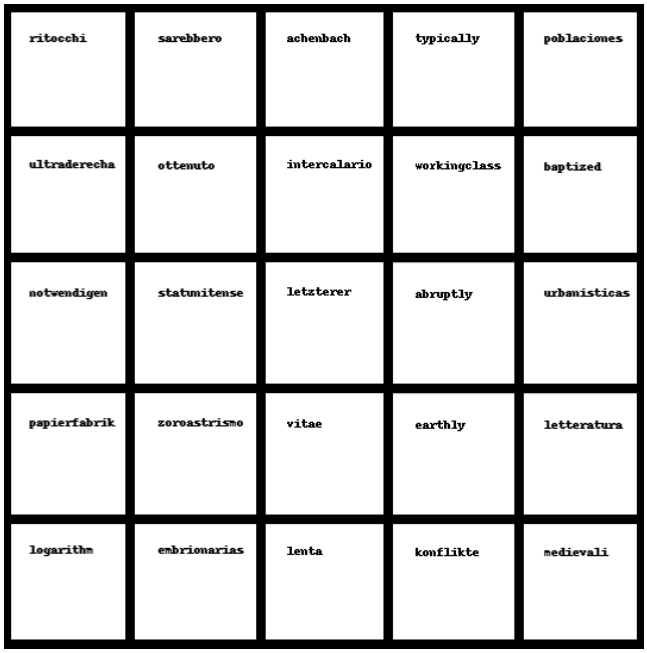

Fig 2. A random sampling of word image data from the four languages

3.2 Training

We created our convolutional neural network using Python with the TensorFlow and Keras frameworks. There are four classes an image can be assigned to, one for each of the languages we have gathered data for. Starting with our 84x84 image, we apply multiple convolution filters using the relu activation function, as recommended by Daggumati and Revesz. [1] In the final layer of the convolutional neural network we use the softmax activation function. This will give us a probabilistic result between 0 and 1 for each class.

The CNN compiles using the Adam optimizer with a learning rate of 0.001. Adam works well in training deep neural networks, and 0.001 is used as recommended once more by Daggumati and Revesz.

The TensorFlow model fits the roughly 45,000 – 55,000 shuffled images to the four classes over the course of ten epochs. 8000 of the images were selected to be used as validation data while the neural network ran. This was computed on a Windows machine using an Nvidia GeForce RTX 2070 Super GPU. Each epoch of the process computed for approximately one hour.

After the training, the resulting model is about 80% accurate at determining the correct language of a given word of one of the four languages. Of the 20% of words that might be classified incorrectly, there are many words which might be extremely similar or identical to a valid word in another language. This is especially true between Spanish and Italian. Therefore we believe these are acceptable results and the model has an adequately high-degree of accuracy at classifying patterns in visual representations of words such that we can use it to analyze cognates.

3.3 Analysis of Known Cognates. With the model described above we can begin inputting cognates which have known relationships between the four languages. The model should return a probability for which the word exists in each of the four languages and we can compare the probabilities to its known origins. We are careful to choose words which are not simply a member of the training data so that the model must necessarily analyze the composition of the letter structure of the word.

One example we try is the English/Spanish word “Matrimonial”. The word is identical between the two languages but the English version of the word gets its origin from Spanish. Italian has a similar cognate in “Matrimoniale” and German has no cognate of the word. We should therefore expect that the model predicts a high degree of accuracy for Spanish and a very low degree of accuracy toward German, and English and possibly Italian between.

Fig 3. Analysis on the word “Matrimonial”. Columns are English, Spanish, German, Italian

The model outputs an extremely high value for Spanish, with a small value for English. This result captures the history of the word, demonstrating with very high confidence its origin in Spanish versus the other four languages.

Another example of a cognate we try is the English word “Banana”. The word itself has its roots in the identical Spanish word Banana, which is identically a cognate in Italian, and near-identically a cognate with the German word Banane. We should therefore expect a high softmax value for Spanish, its origin language.

Fig 4. Analysis on the word “Banana”. Columns are English, Spanish, German, Italian

Interestingly, we see English and Spanish both perform highly with roughly the same probabilities. It is interesting to observe how low the word scores in Italian despite the close language relation between Spanish and Italian and despite the Italian word Banana being visually identical to the Spanish word Banana . If we had not known the origins of this word already, we could correctly make a very safe assumption with the data above that the word “Banana” has its origins in either English or Spanish.

We try the Italian word “Favoloso” which has its roots in Latin, which is itself very closely related to modern-day Italian and to a lesser extent modern-day Spanish. Favoloso is a cognate with the Spanish word Fabuloso and the English word Fabulous. We choose this word because it demonstrates another interesting linguistic quirk in etymology. The cognates Favoloso / Fabuloso / Fabulous have a false cognate with the German word Fabelhaft. A false cognate is a word which ostensibly appears to be related but is only similar in appearance or phonetics by coincidence.

Fig 5. Analysis on the word “Favoloso”. Columns are English, Spanish, German, Italian

We see the model correctly determine it is an Italian word and apply a low confidence to Spanish, its most similar cognate. German and English are completely ruled out as origins of the word and the model therefore correctly disregarded the false cognate possibility.

We conclude that the model implemented possesses the ability to predict with an adequate degree of accuracy the historical roots of a word based only on its visual representation.

3.3 Analysis of Cognates of Unknown Origin. Given our conclusions above, we try inputting words which have no known roots and analyze the results.

First we try the English word “Curse”. This word appears only in English and has no known origin or relation in any other language. It is possible there is an overlooked or otherwise lost cognate in another language to be found.

Fig 6. Analysis on the word “Curse”. Columns are English, Spanish, German, Italian

We see the model correctly determines Curse is an English word. Of interest is that it has a small degree of confidence in Spanish. It could be the case that the composition of the word, the way with which the letters are arranged, has some similarity to other Spanish words. This could be an indication of some Spanish false cognates. Or if there exists a true cognate at all, we might conclude that Spanish may be the best language to search for one.

We try another word “Toad” which is found in English but no other languages as far as we presently know.

Fig 7. Analysis on the word “Toad”. Columns are English, Spanish, German, Italian

Similar to before, we see the word is instantly recognized as English and Spanish once again appears as a possible alternative. We may postulate that certain old English words of unknown origin may be similar in letter structure to existent Spanish words.

5 Conclusions and Future Work

In this research we have exhibited a convolutional neural network which can predict with a sufficiently high degree of accuracy which language a word belongs to given only an image of the word.

The success of the convolutional neural network in detecting letter patterns allowed us to consider whether such an application would have success in detecting the correct origins of a word. This appeared to be moderately successful. And this by extension allows us to consider if such an application might be able to predict cognates of words which are presently unknown to linguists.

We have not proven the existence of a new cognate using this tool. However we have provided evidence that this type of implementation may be able to do so, given a more robust dataset of vocabulary and using more languages than just English, Spanish, German, and Italian. Specifically including French would improve English cognate analysis drastically, as French appears to share many more cognates with English than Spanish or Italian do.

We theorize that there would be significant increases in accuracy if, rather than using simple ASCII representations of the word as done here, words were instead converted to some phonetic representation. For example. The International Phonetic Alphabet (IPA) is widely used by linguists to represent the sounds we make when we pronounce a word in any given language.

Doing the above would provide two advantages. First, we would remove misleading letter patterns. A word may be identically arranged between two languages but the letters themselves may have different pronunciations between the two languages. This is a nuance that would be missed by a computer vision model and it might drastically affect the determination of whether a word is a cognate or not. Secondly, this approach would more accurately

represent the transfer of cognates between languages. Historically words are not so often mixed between written language as much as they are mixed between spoken language. This is less true today than in the past, but a phonetic approach would likely capture the total number of cognates more accurately in any case.

Indeed, a more accurate approach to finding cognates may also be found in analyzing audio data of words rather than this type of computer vision analysis altogether. Gathering the necessary data for an audio analysis would be more difficult than gathering image data, but it may be adequately achievable today using modern text-to-speech software. This, we suspect, would be the largest area of further research in this topic.

REFERENCES

[1] Daggumati, Shruti, and Peter Z. Revesz. "Data mining ancient scripts to investigate their relationships and origins." Proceedings of the 23rd International Database Applications & Engineering Symposium. 2019.