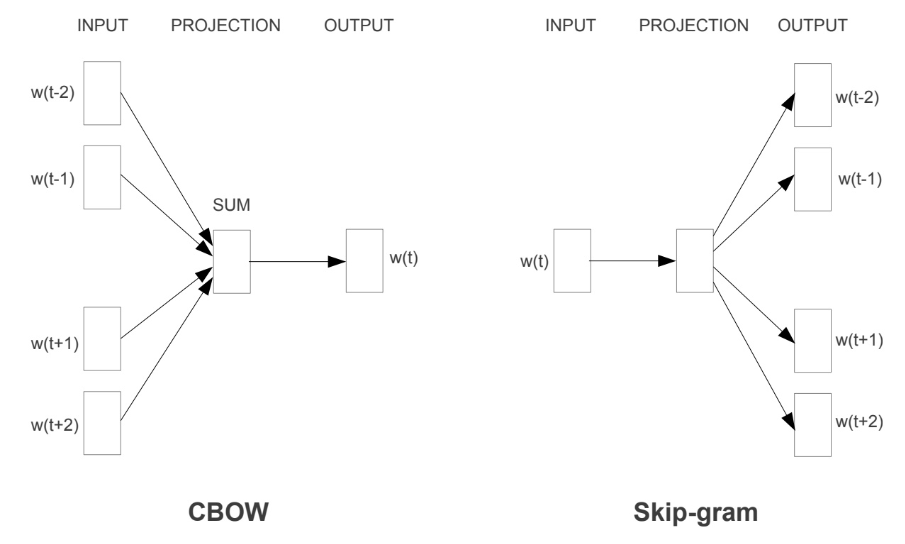
**TÌM HIỂU WORD2VEC**

Trong natural language processing thì để xử lý dữ liệu text chúng ta cũng phải chuyển dữ liệu từ text sang numeric, tức là đưa nó vào một không gian mới người ta thường gọi là embedding. Trước đây người ta mã hóa theo kiểu one hot encoding tức là tạo một vocabulary cho dữ liệu và mã hóa các word trong document thành những vecto, nếu word đó có trong document thì mã hóa là 1 còn không có sẽ là 0. Kết quả tạo ra một sparse matrix, tức là matrix hầu hết là 0. Các mã hóa này có nhiều nhược điểm đó là thứ nhất là số chiều của nó rất lớn (NxM, N là số document còn M là số vocabulary), thứ 2 các word không có quan hệ với nhau. Điều đó dẫn đến người ta nghĩ ra một model mới có tên là **Word embedding**, ở đó các word sẽ có quan hệ với nhau về semantic tức là ví dụ như paris-tokyo, man-women, boy-girl những cặp từ này sẽ có khoảng cách gần nhau hơn trong Word embedding space. Word embedding có 2 model nổi tiếng là word2vec và Glove.

Có 2 mô hình Word2vec được áp dụng: Skip-gram, Continuous Bag of Words (CBOW)



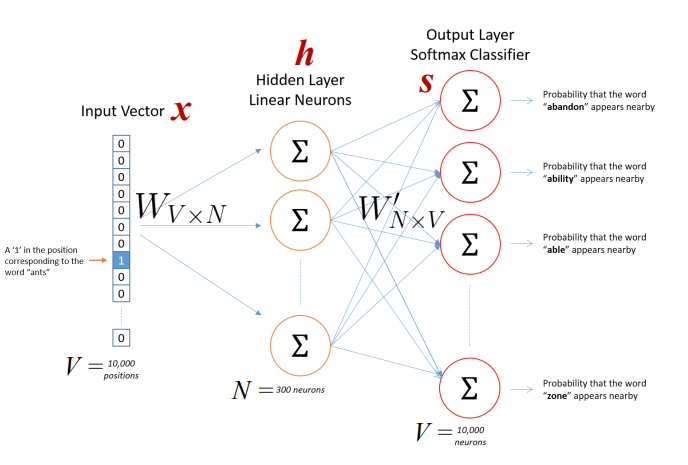
**CBoW**:

* Cho các từ ngữ cảnh
* Đoán xác suất của một từ đích

**Skip-gram**:

* Cho từ đích
* Đoán xác suất của các từ ngữ cảnh

**Skip-gram:**

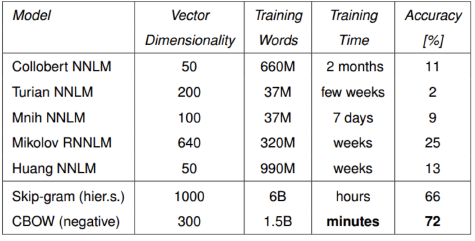


* Input là one-hot-vector mỗi word sẽ có dạng x­1, x2,…xv. Trong đó V là số vocabulary, là một vector trong đó mỗi word sẽ có giá trị 1 tương đương với index trong vocabulary và còn lại sẽ là 0.
* Weight matrix giữa input và hidden layer là matrix W (có dimention VxN) có active function là linear, weight giữa hidden và out put là W′ (có dimention là NxV) active function của out put là soft max.
* Mỗi row của W là vector N chiều đại diện cho vw là mỗi word trong input layer.Mỗi row của W là  . Lưu ý là input là 1 one hot vector (sẽ có dạng 000100) chỉ có 1 phần tử bằng 1 nên.

Từ hidden layer đến out put là matrix . Ta tính score  cho mỗi word trong vocabulary.

**Ưu điểm của word2vec:**

**Tốc độ:**

****