# word2vecの高速化

4.1 – 4.3

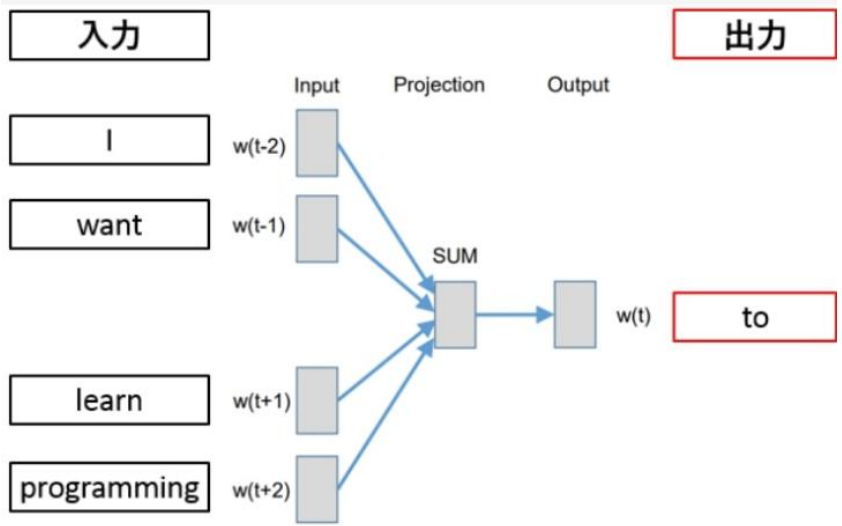
# 【復習】word2vecについて

現在の目的：単語の分散表現を得る事

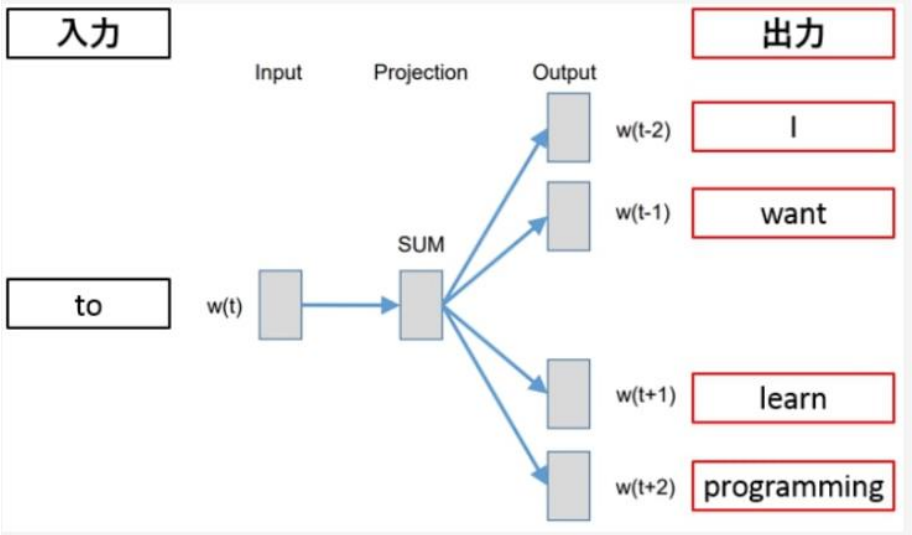
分散表現：単語を計算出来るように、単語が持つ意味を含んだ数値情報

word2vecは単語の分散表現を得る手法で大きく次の二つ

・CBOW(Continuous Bag-of-Words Model)
周辺の単語から中心語を予測する。

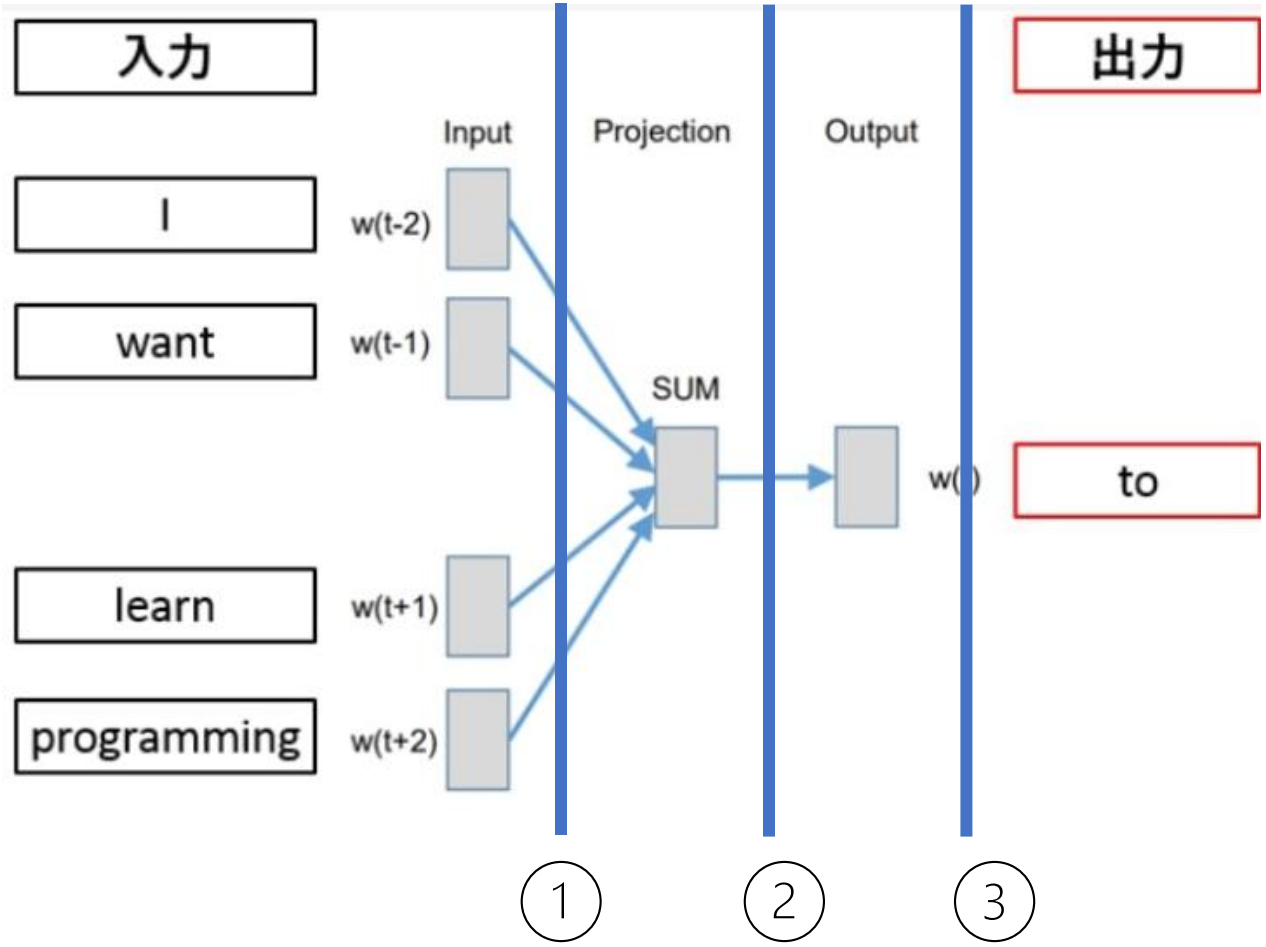



・skip-gram法（Continuous Skip-Gram Model)
中心の単語から周辺の単語を予測する。



# 現在の問題点

入力で扱う単語数が増加すると、中間層の前後の重みとの乗算とSoftmaxでの確立への変換の負荷が重くなる



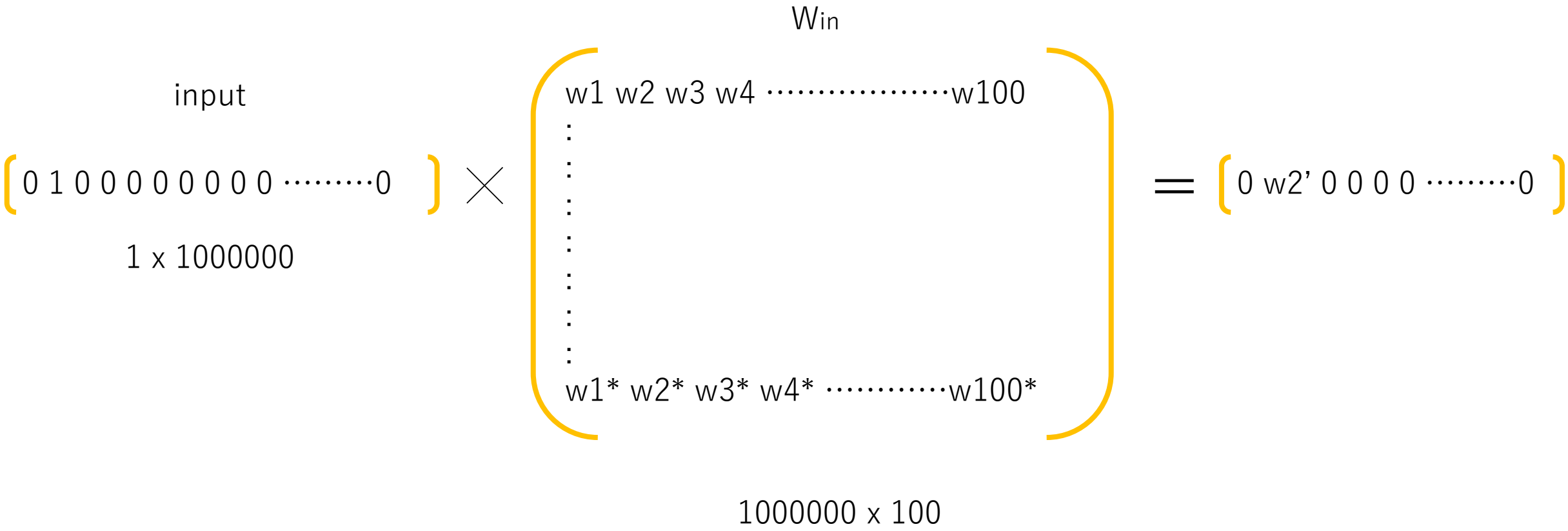

①入力から中間層までの計算

②中間層から出力層までの計算

③出力値を確立へ変換するときの計算

## ② 入力層から中間層までの問題　③

入力層からの処理をMatMUlレイヤ層ではone-hotベクトルと重み($W_{in}$)の積を計算している。
入力の語彙数が100万の様に巨大になってしますと巨大なベクトルと重み行列の積の計算する必要がある。
入力と重みで処理しているのは、入力のone-hotで該当する重みを抽出しているに過ぎない



# ① 中間層から出力層までの問題　③出力値を確率へ変換するときの計算

今までは、入力の単語数の100万を出力するので
中間層からWoutを100万回かけて１００万個出力する。
その結果をSoftmax関数にぶつけて計算している

$$f_i(x) = \frac{e^{x_i}}{\sum_{k=1}^{n} e^{x_k}}$$

分子は、eを“単語の値“で乗する
分母は、eをすべて(１００万個)で乗する
→　100万回×100万回なので1兆回の計算が必要になる。

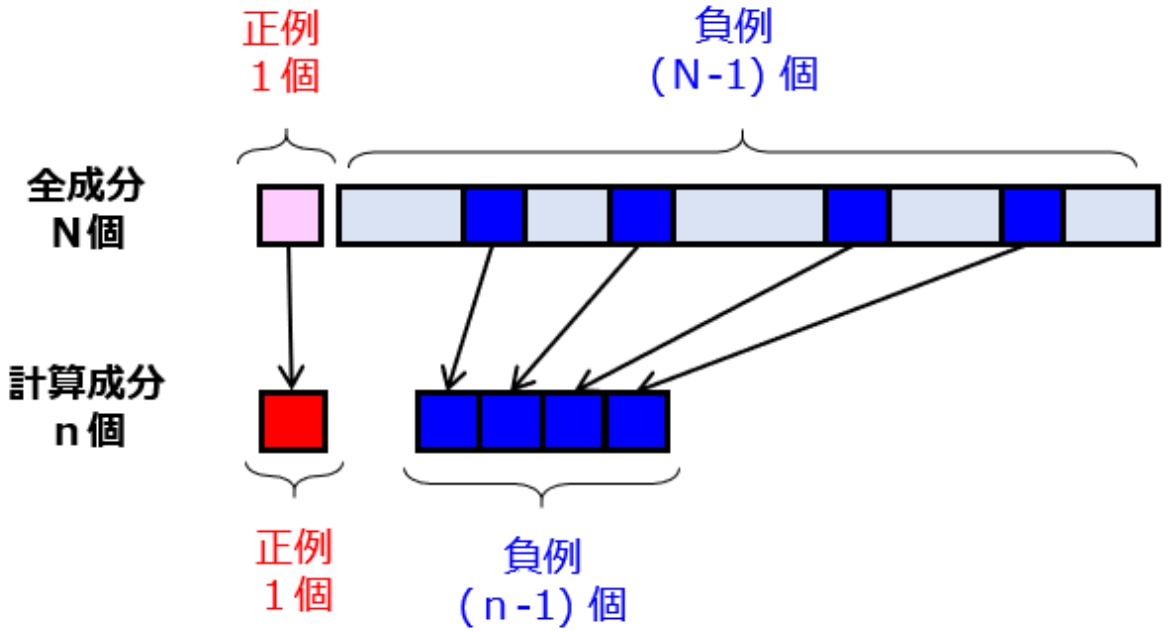
# ① 中間層から出力層までの問題　③出力値を確率へ変換するときの計算

Negative sampling

100万単語から１つの正解を見つけることが難しい。

正解を１つ誤答を100の二分類に変更してロジスティク回帰で学習する方法。

抽出方法は、入力に使うコーパスの中での各単語の出現率(確率)によって見つける。



① 中間層から出力層までの問題　③出力値を確率へ変換するときの計算

二値分類のスコアを確率に変換するのでシグモイド関数で変換。
その後、損失関数を交差エントロピーをぶつける。
正解とした単語でLossが0に近づけば重みづけは正しいとなる。