Autoencoder with Word2Vec

Gonsoo Moon
Post Master Student of Computer Science
University of Texas at Dallas,
gonsoomoon@gmail.com

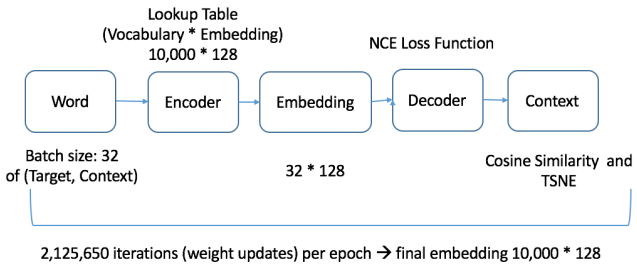
1. Abstract

Using the "skipgram" model, Autoencoder is implemented with layers of an encoder, embeddings and decoder. For visualization, not only do nearest words are generated with cosine similarity, but also t-distributed stochastic nearest embedding("tsne") is used

2. Introduction

The <Figure1> is shown for our Autoencoder with: 1) as input 32 batch size consisting of a pair of (target, context), an encoder using a lookup table, embedding, the NCE (Noise-Contrastive Estimation) loss function as a decoder, producing final embedding 10,000 * 128 matrix and finally, a result using the cosine similarity and TSNE (t-Distributed Stochastic Nearest Embedding).

<Figure1>



3. Data Preprocessing

The input data is downloaded from http://mattmahoney.net/dc/ and it is a number of 17,005,207 words as a corpus. The vocabulary size is set to 10,000 that occurred most frequently in the corpus. Thus, most common words are [['UNK', 1,737,307], ('the', 1,061,396), ('of', 593,677), ('and', 416,629), ('one', 411,764)] (UNK includes words not in the vocabulary). For the skipgram model, 5 of a skip window that is how many words to consider left and right from target word, 4 of number of skips that is how many times to reuse an input to generate. Based on the above parameters, as an example of a pair of input and output as a label, the following pairs are generated: originated -> anarchism, originated -> as, as -> a, as -> originated, a -> as, a -> term, term -> a, term -> of. The batch size is 32 so that a number of iterations per an epoch is used with 2,125,650 = 17,005,207 * 4 / 32 (data size * number of skips / batch size). Five epochs are used and embedding size is 128.

4. Training

The NCE (Noise-Contrastive Estimation) loss function is used. I use five epochs so that 10,628,000 iterations are repeated for updating weights that is a size of embedding, 10,000 * 128. The final iteration is shown as follows:
Elapsed: 10628000 batches, Cost = 4.739091682

5. Result

The <Figure2> shows one of the results that is using the cosine similarity. The word 'later' 's nearest words are 'eventually', 'then', 'before', 'subsequently', 'several', 'after', 'first' and 'he'. The word 'July' shows accurately month's word such as 'august', 'april', 'january', 'december', 'june', 'november', 'october'.

<Figure 2>

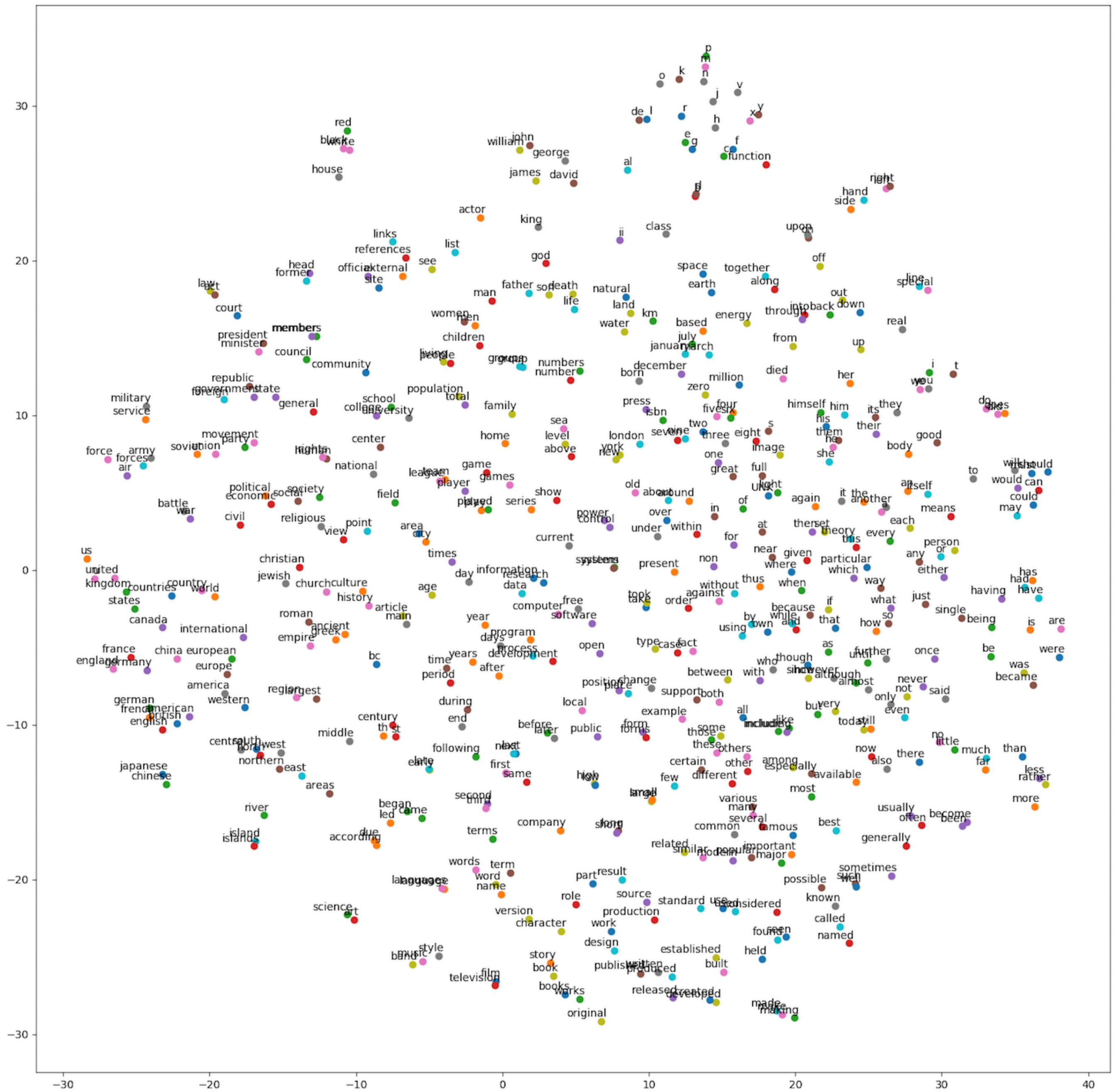
```
Nearest neighbor of numbers: number, arithmetic, points, elements, together, correspondence, chaucer, astrology
Nearest neighbor of war: battle, ii, during, bombing, wars, period, victory, foster
Nearest neighbor of later: eventually, then, before, subsequently, several, after, first, he
Nearest neighbor of july: august, april, january, december, june, november, september, october
Nearest neighbor of terms: second, words, third, every, often, both, UNK, value
```

The <Figure3> shows 2-dimensional vector space of words. From the final embeddings of 10,000 * 128, TSNE reduces to 10,000 * 2 so that each word is represented by (x, y) coordinate. As an example in the <Figure3> showing 500 words out of 10,000, the word 'data' is displayed roughly in (0, 0). Its nearest word is 'information', 'research' and 'computer'.

6. References

[1] Nikhil Buduma (2017). Fundamentals of deep learning. Sebastopol, CA: O'Reilly Media

< Figure 3>


30
20
10
0
−10
−20
−30
−30
−20
−10
0
10
20
30
40
red
house
actor
king
god
list
links
references
see
official
external
site
head
former
law
court
president
minister
members
council
community
republic
state
general
military
service
force
army
forces
air
movement
party
union
center
national
school
college
university
family
population
total
children
women
man
father
death
life
natural
land
water
km
space
earth
together
along
based
energy
through
back
down
out
off
upon
class
al
hand
side
right
function
george
david
james
william
john
line
special
real
up
from
her
died
million
zero
born
december
january
march
july
numbers
number
press
isbn
two
three
four
five
eight
one
image
great
full
himself
him
his
them
she
their
its
they
body
good
to
would
can
could
may
means
each
every
person
any
either
has
had
have
having
being
is
are
were
was
became
be
single
just
what
how
so
if
that
once
further
until
never
not
said
only
even
much
than
far
less
rather
more
usually
often
become
been
generally
sometimes
such
known
called
named
held
built
made
making
established
found
seen
considered
standard
source
production
work
design
story
book
books
original
released
developed
created
character
version
role
part
result
television
film
style
band
music
science
art
words
term
word
name
language
languages
terms
company
second
third
began
led
came
according
due
island
islands
river
areas
japanese
chinese
northern
east
central
south
west
north
middle
german
french
american
british
english
france
england
germany
china
european
europe
america
western
international
canada
united
kingdom
states
countries
country
world
us
christian
jewish
church
culture
history
article
main
roman
ancient
greek
empire
bc
region
largest
century
following
next
last
first
same
early
late
time
years
period
after
during
end
before
later
public
local
form
forms
year
program
days
process
development
computer
free
software
open
type
case
fact
between
with
though
however
although
almost
support
both
example
some
those
these
others
other
among
especially
certain
different
few
available
most
various
many
several
famous
common
best
related
similar
modern
popular
important
major
possible
position
place
change
all
including
included
like
buy
very
today
still
now
also
there
no
little
political
economic
social
society
religious
civil
battle
war
field
point
view
area
city
times
day
information
research
data
age
league
team
player
game
games
played
series
show
power
control
current
systems
system
present
non
near
given
where
when
without
against
thus
order
take
took
by
own
while
and
because
using
as
particular
which
way
other
another
theory
therefore
this
for
within
under
over
around
about
old
in
of
at
light
again
it
london
new
york
home
sea
level
above
small
large
high
low
short
long