
IRON
HACK
EMBEDDINGS

conceptFundamental

"You shall know a word by the company it keeps"

Firth, 1957

(Firth, J. R. 1957:11)

Why learn vector space models?

Where are you heading?

Where are you from?

↓

Different meaning

What is your age?

How old are you?

↓

Same Meaning

WHAT ARE EMBEDDINGS?

- Embeddings are mathematical representations of unstructured information (text, sound, images) that allows objects with similar semantics to have similar representations
- e.g. in a good text embedding, words with similar meanings are embedded in similar vectors (contrast this with BoW)
- e.g. in a good image embedding, pictures with similar objects are embedded in similar vectors

Embeddings

from words to numbers

	MARY	IS	HUNGRY	HAPPY	FOR	APPLES	NOT	JOHN	HE	
Mary is hungry for apples." →	1	1	1	0	1	1	0	0	0	→ [1, 1, 1, 0, 1, 1, 0, 0, 0
'John is happy he is not hungry for apples." →	0	2	1	1	1	1	1	1	1	→ [0, 2, 1, 1, 1, 1, 1, 1, 1

Embeddings

from words to numbers

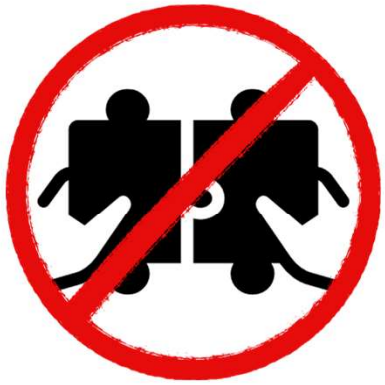
What if we could represent the words via its “meaning” and not its letters?

Using only “letters”

Student

Lab

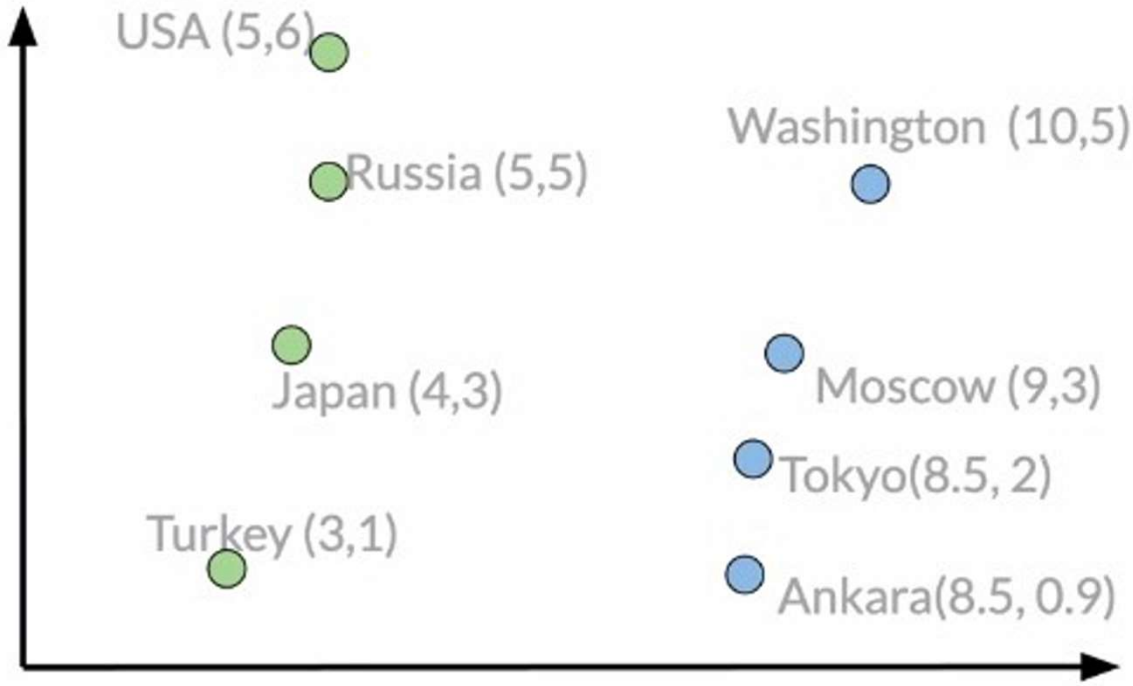
Exercise



User query: “Someone in my class had a question!

Embeddings

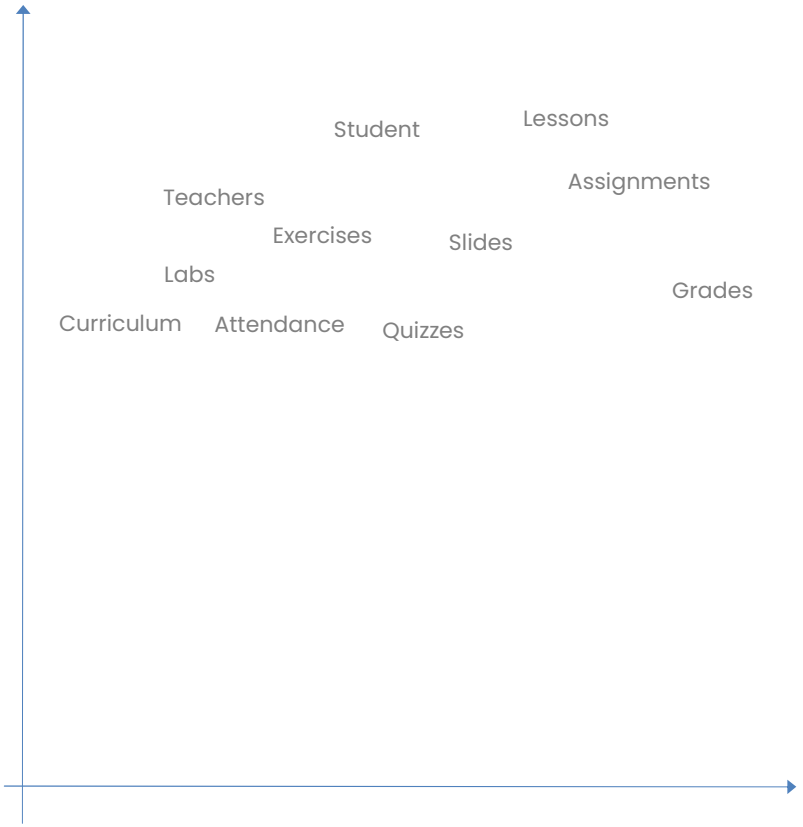
from words to numbers



Embeddings

from words to numbers

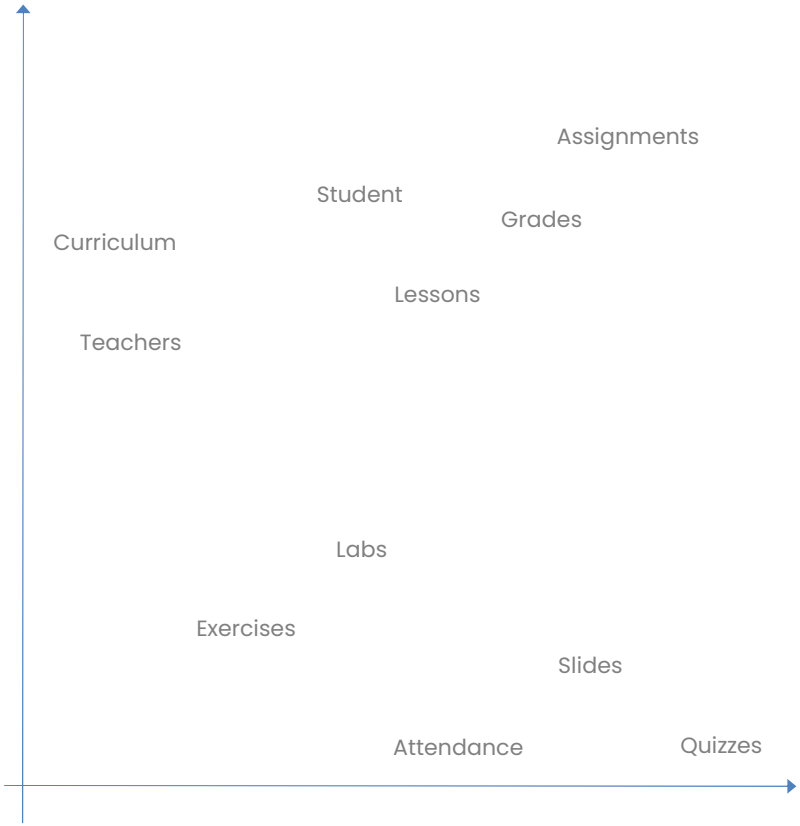
The classroom dashboard integrates data from Students and Teachers to display attendance and grades. Teachers use the platform to assign Exercises and Lessons to Students. Labs provide practical exercises that complement the theoretical Lessons. Slides are used by Teachers during Lessons to illustrate key points. Assignments are tracked and graded by Teachers, with results available to Students. Quizzes are periodically given to assess Students' understanding of the material. The curriculum outlines the sequence of Lessons and Labs for the term. Attendance is monitored daily and logged by Teachers. Grades are calculated based on performance in Assignments, Quizzes, and Labs. Students can view their progress in real-time on their dashboard. Teachers use data from Lessons and Quizzes to tailor future Exercises and Assignments.
"Where can I find the slides for today's lesson?" "How do I submit my assignment?" "What exercises should I complete for the lab session?" "Can I review my grades for the last quiz?" "How is attendance being tracked in our classroom?" "Is there a way to see the curriculum for this semester?" "Can the teacher upload the slides from yesterday's lesson?" "How do I know which labs I need to complete?" "What is the due date for the current assignment?" "How are grades calculated from quizzes and assignments?" "Can the teacher see which students have completed the exercises?"



Embeddings

from words to numbers

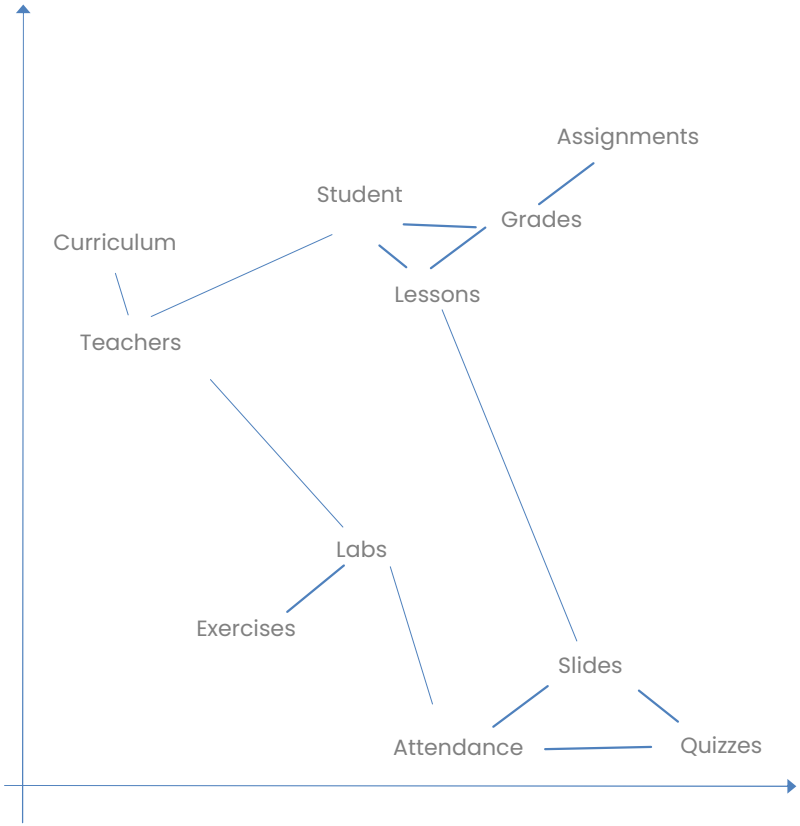
The classroom dashboard integrates data from Students and Teachers to display attendance and grades. Teachers use the platform to assign Exercises and Lessons to Students. Labs provide practical exercises that complement the theoretical Lessons. Slides are used by Teachers during Lessons to illustrate key points. Assignments are tracked and graded by Teachers, with results available to Students. Quizzes are periodically given to assess Students' understanding of the material. The curriculum outlines the sequence of Lessons and Labs for the term. Attendance is monitored daily and logged by Teachers. Grades are calculated based on performance in Assignments, Quizzes, and Labs. Students can view their progress in real-time on their dashboard. Teachers use data from Lessons and Quizzes to tailor future Exercises and Assignments.
"Where can I find the slides for today's lesson?" "How do I submit my assignment?" "What exercises should I complete for the lab session?" "Can I review my grades for the last quiz?" "How is attendance being tracked in our classroom?" "Is there a way to see the curriculum for this semester?" "Can the teacher upload the slides from yesterday's lesson?" "How do I know which labs I need to complete?" "What is the due date for the current assignment?" "How are grades calculated from quizzes and assignments?" "Can the teacher see which students have completed the exercises?"



Embeddings

from words to numbers

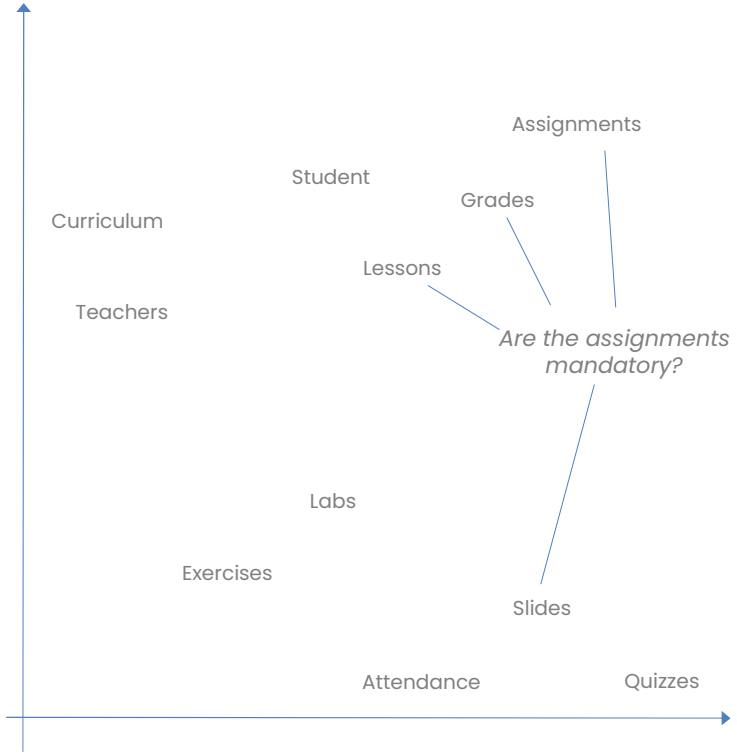
The classroom dashboard integrates data from Students and Teachers to display attendance and grades. Teachers use the platform to assign Exercises and Lessons to Students. Labs provide practical exercises that complement the theoretical Lessons. Slides are used by Teachers during Lessons to illustrate key points. Assignments are tracked and graded by Teachers, with results available to Students. Quizzes are periodically given to assess Students' understanding of the material. The curriculum outlines the sequence of Lessons and Labs for the term. Attendance is monitored daily and logged by Teachers. Grades are calculated based on performance in Assignments, Quizzes, and Labs. Students can view their progress in real-time on their dashboard. Teachers use data from Lessons and Quizzes to tailor future Exercises and Assignments.
"Where can I find the slides for today's lesson?" "How do I submit my assignment?" "What exercises should I complete for the lab session?" "Can I review my grades for the last quiz?" "How is attendance being tracked in our classroom?" "Is there a way to see the curriculum for this semester?" "Can the teacher upload the slides from yesterday's lesson?" "How do I know which labs I need to complete?" "What is the due date for the current assignment?" "How are grades calculated from quizzes and assignments?" "Can the teacher see which students have completed the exercises?"



Embeddings

from words to numbers

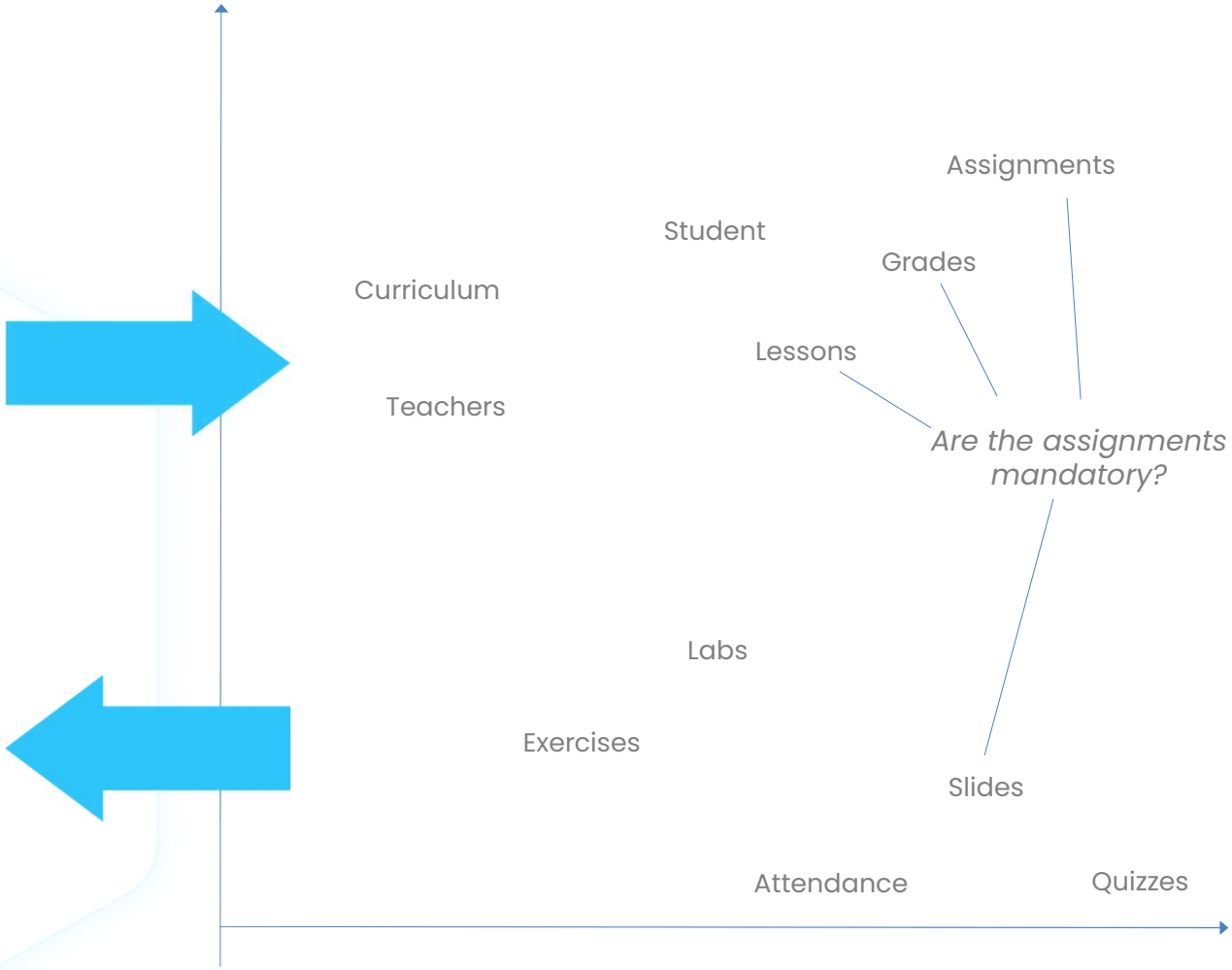
The classroom dashboard integrates data from Students and Teachers to display attendance and grades. Teachers use the platform to assign Exercises and Lessons to Students. Labs provide practical exercises that complement the theoretical Lessons. Slides are used by Teachers during Lessons to illustrate key points. Assignments are tracked and graded by Teachers, with results available to Students. Quizzes are periodically given to assess Students' understanding of the material. The curriculum outlines the sequence of Lessons and Labs for the term. Attendance is monitored daily and logged by Teachers. Grades are calculated based on performance in Assignments, Quizzes, and Labs. Students can view their progress in real-time on their dashboard. Teachers use data from Lessons and Quizzes to tailor future Exercises and Assignments.
"Where can I find the slides for today's lesson?" "How do I submit my assignment?" "What exercises should I complete for the lab session?" "Can I review my grades for the last quiz?" "How is attendance being tracked in our classroom?" "Is there a way to see the curriculum for this semester?" "Can the teacher upload the slides from yesterday's lesson?" "How do I know which labs I need to complete?" "What is the due date for the current assignment?" "How are grades calculated from quizzes and assignments?" "Can the teacher see which students have completed the exercises?"



Embeddings

from words to numbers

Are the assignments mandatory?



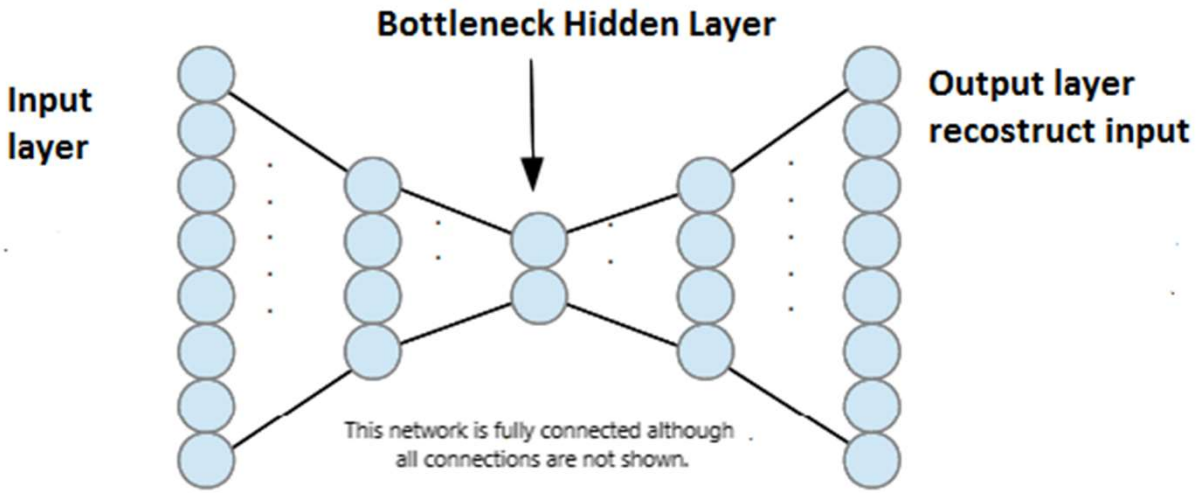
Docname; source
Docname;source
Docname;source
Docname;source
Docname;source
Docname;source

AUTOENCODERS

Unsupervised dimensionality reduction technique

Forces bottleneck layer to encode all information

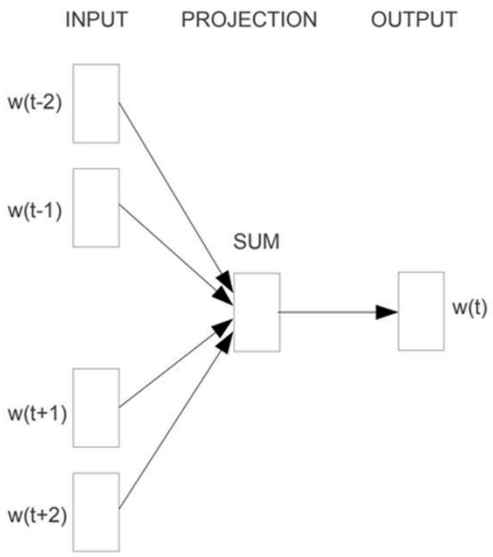
Good shortcut to encode information is via semantics (meaning)



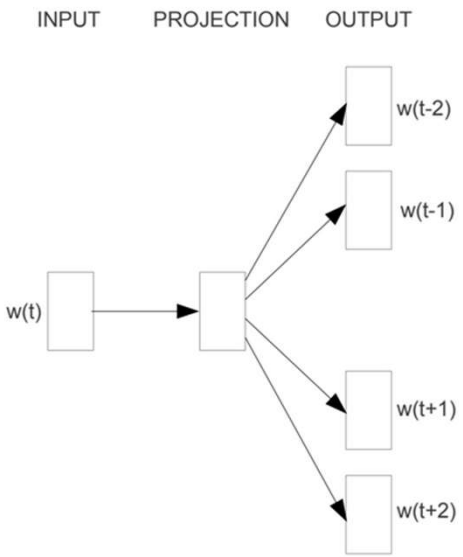
WORD EMBEDDINGS - WORD2VEC

Enriches the notion of autoencoder by considering not only the word but also the words that show up around it

Two versions, depending if it's more important to guess the word based on the context or predict the context based on the word



CBOW



Skip-gram

We have huge corpuses for training, essentially all of literature created so far

WORD EMBEDDINGS - BERT FAMILY

Word2Vec still associates the same vector to the same word

BERT - Bidirectional Encoder Representations from Transformers consider not only the word but also it's context *during encoding*, which means it can extract different meanings from the same word

"the man robbed a bank" vs "the bank of the river was flooded"

EMBEDDINGS

Pro : can be used pre-trained if your application is generic. For the vast majority of cases the word embeddings trained on Wikipedia are good enough

Pro: can be customized from a pre-learned model at a fraction of the cost. If you have a relatively modest corpus, you can piggyback on a pre trained model that already knows most of the smeantics of english

Con: can be quite difficult to train from scratch. Pretty much only google and a few other large data holders can create these models (queue GPT3)

Con: they are too cool. Food and shelter are also needed but pale in comparison.