Discussion:

Asset Embeddings

Xavier Gabaix, Ralph S.J. Koijen, Robert J. Richmond, Motohiro Yogo

Yinan Su
Johns Hopkins University Carey Business School

ABFR Webinar
on Sept. 26, 2022

Big picture

Big picture research directions:

- quantity data $\rightarrow$ asset pricing
- AI/ML $\rightarrow$ finance research

Big picture

Big picture research directions:

- quantity data $\rightarrow$ asset pricing
- AI/ML $\rightarrow$ finance research

Objective of the paper

- "asset embeddings" (numerical representation of assets) go beyond observable stock characteristics

Big picture

Big picture research directions:

- quantity data $\rightarrow$ asset pricing
- AI/ML $\rightarrow$ finance research

Objective of the paper

- "asset embeddings" (numerical representation of assets) go beyond observable stock characteristics

- Important question, valuable work
- Creative ideas, innovative tools
- Rich content, extensive analysis

Textual embedding

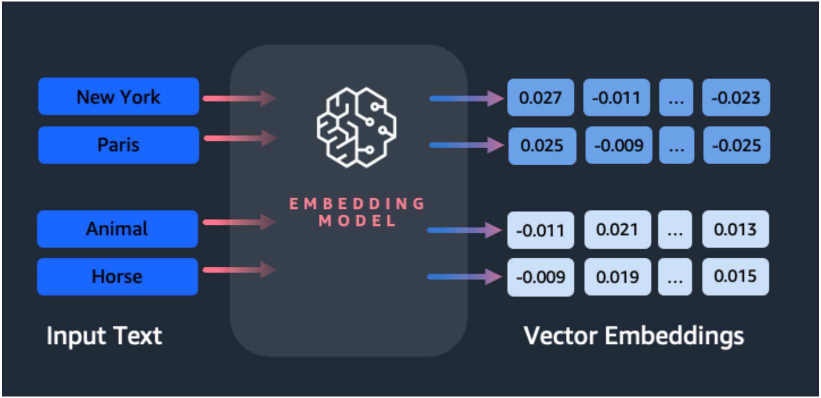


illustration source: AWS Machine Learning Blog.

https://aws.amazon.com/blogs/machine-learning/getting-started-with-amazon-titan-text-embeddings/

Semantics similarity

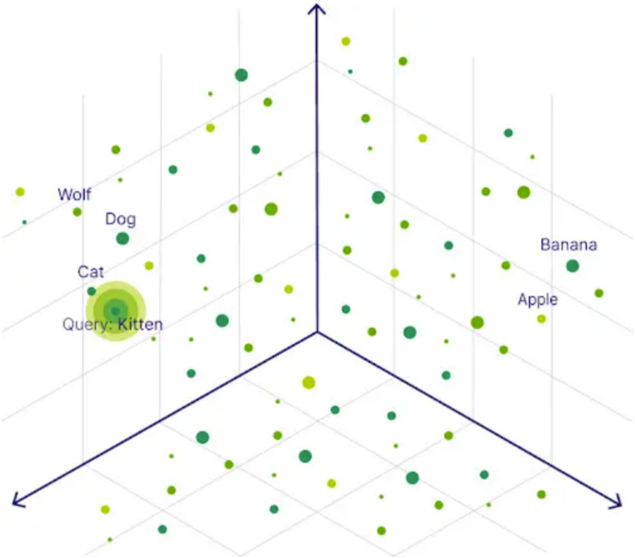


illustration source: From prototype to production: Vector databases in generative AI applications.
https://stackoverflow.blog/2023/10/09/from-prototype-to-production-vector-databases-in-generative-ai-applications/

Key idea, how is NLP useful for our task?

- Treat portfolio holdings as "sentences"

  ARK$_t$: "*Zoom, IBM, Tesla, Walmart, ...*"

  SPY$_t$: "*Apple, Microsoft, Nvidia, ...*"

  ...

  create a new "roll call" language

  very simple "grammar": order by holdings rank

Key idea, how is NLP useful for our task?

- Treat portfolio holdings as "sentences"
  ARK$_t$: "*Zoom, IBM, Tesla, Walmart, ...*"
  SPY$_t$: "*Apple, Microsoft, Nvidia, ...*"
  ...
  create a new "roll call" language
  very simple "grammar": order by holdings rank
- If AI is so smart to understand human languages, computer languages, etc., probably also this "roll call" language!

Key idea, how is NLP useful for our task?

- Treat portfolio holdings as "sentences"
  ARK$_t$: "*Zoom, IBM, Tesla, Walmart, ...*"
  SPY$_t$: "*Apple, Microsoft, Nvidia, ...*"
  ...
  create a new "roll call" language
  very simple "grammar": order by holdings rank
- If AI is so smart to understand human languages, computer languages, etc., probably also this "roll call" language!
- "understand" in the statistical sense, tasks like:
  predict the next word
  fill in the blanks

Key idea, how is NLP useful for our task?

- Treat portfolio holdings as "sentences"
  ARK$_t$: "*Zoom, IBM, Tesla, Walmart, ...*"
  SPY$_t$: "*Apple, Microsoft, Nvidia, ...*"
  ...
  create a new "roll call" language
  very simple "grammar": order by holdings rank
- If AI is so smart to understand human languages, computer languages, etc., probably also this "roll call" language!
- "understand" in the statistical sense, tasks like:
  predict the next word
  fill in the blanks
- That is what we want for assets as well!
  we want to find stocks that are similar to each other
  "similar" in the sense of
  1) being held by the same investors, and
  2) with similar weights
- So let's train NLP nn on this language corpus and get the embeddings (neuron activations)

Key idea

- Creative idea!
- Holdings reflect asset characteristics
  "similar" firms should often appear together
  different investors have different "styles" (size, value, ...), so different aspects of the firm can be captured

Key idea

- Creative idea!
- Holdings reflect asset characteristics
  "similar" firms should often appear together
  different investors have different "styles" (size, value, ...), so different aspects of the firm can be captured
- My comments are mostly technical
  Thinking about the methodological connection between nlp methods and firm characteristics and asset pricing research
  My message: a transfer from ml/nlp to finance is not necessarily straightforward, requires careful consideration

Comment: input, contextualized, sentence embeddings

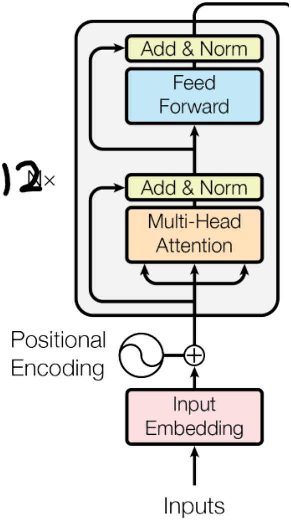


illustration source: “Attention is All You Need” by Vaswani et al. (2017)

Comment: input, contextualized, sentence embeddings

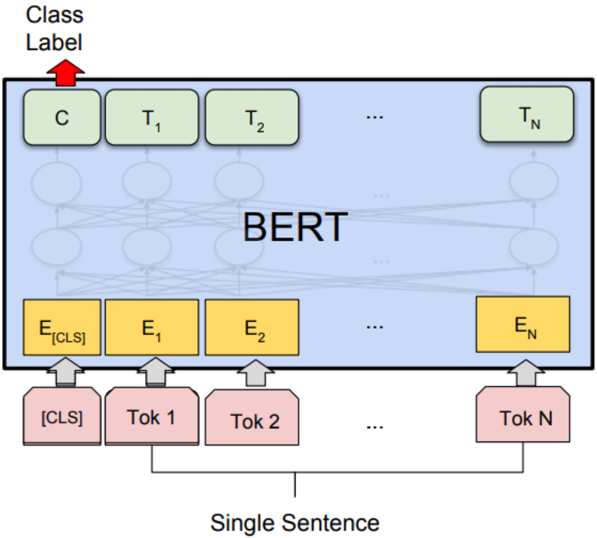


illustration source: https://yashuseth.wordpress.com/2019/06/12/bert-explained-faqs-understand-bert-working/

Comment: input, contextualized, and sentence embeddings

Summary of terminology:

- input embedding:
  token-level (indexed by firm)
  (context free, before 12-layer transformer)
- contextualized embedding:
  token-sentence-level (indexed by firm, investor)
  (considers a word's meaning vis-a-vis the entire sentence)
- *sentence embedding*:
  contextualized embedding of the special token [CLS]
  represents the semantic content of the whole sentence
  key output of BERT

Comment: contextualize in investor or time?

Unit of observation of the embeddings:

- We want: a, t (firm, month/quarter)
  as in [size$_{a,t}$, value$_{a,t}$, momentum$_{a,t}$, ...]

Comment: contextualize in investor or time?

Unit of observation of the embeddings:

- We want: a, t (firm, month/quarter)
  as in [size$_{a,t}$, value$_{a,t}$, momentum$_{a,t}$, ...]
- Method section: a, i (firm, investor)
  contextualized embeddings

Comment: contextualize in investor or time?

Unit of observation of the embeddings:

- We want: a, t (firm, month/quarter)
  as in [size$_{a,t}$, value$_{a,t}$, momentum$_{a,t}$, ...]
- Method section: a, i (firm, investor)
  contextualized embeddings
- Results section: a (firm)
  input embeddings

Comment: contextualize in investor or time?

Unit of observation of the embeddings:

- We want: a, t (firm, month/quarter)
  as in [size$_{a,t}$, value$_{a,t}$, momentum$_{a,t}$, ...]
- Method section: a, i (firm, investor)
  contextualized embeddings
- Results section: a (firm)
  input embeddings

Problems:

1. No “time” in characterizing a firm
   static? each period is isolated?
   can we still do simple tasks like characteristics-sorted portfolios?
2. Attention mechanism seems like an overkill
   input embedding is not the key output of BERT, sentence embedding is, imo

Comment: contextualize in investor or time?

Unit of observation of the embeddings:

- We want: a, t (firm, month/quarter)
  as in [size$_{a,t}$, value$_{a,t}$, momentum$_{a,t}$, ...]
- Method section: a, i (firm, investor)
  contextualized embeddings
- Results section: a (firm)
  input embeddings

Problems:

1. No "time" in characterizing a firm
   static? each period is isolated?
   can we still do simple tasks like characteristics-sorted portfolios?
2. Attention mechanism seems like an overkill
   input embedding is not the key output of BERT, sentence embedding is, imo

My proposal:

- each firm-quarter as a sentence
  do sentence embeddings

My proposal

- each firm-quarter as a sentence
  AAPL_{202409}: "*SPY, QQQ, ARKK, ...*"
  MSFT_{202409}: "*VOO, Buffet, SPY, ...*"
- do sentence embeddings:
  AAPL_{202409}: $[0.1, -0.2, +0.3, ...]$
  MSFT_{202409}: $[...]$

What is good about this?

- firm-<u>time</u> panel structure is back
- **characterizes firms by who holds them**
  BERT can learn investor types (token level)
  (suppose two hedge funds are "synonyms," then ...)
- supports OOS in time
  train BERT IS, feed new sentence to pre-trained model
  (underlying assumption: investor properties are stable)

My proposal

- each firm-quarter as a sentence
  AAPL_{202409}: "*SPY, QQQ, ARKK, ...*"
  MSFT_{202409}: "*VOO, Buffet, SPY, ...*"
- do sentence embeddings:
  AAPL_{202409}: $[0.1, -0.2, +0.3, ...]$
  MSFT_{202409}: $[...]$

What is good about this?

- firm-time panel structure is back
- **characterizes firms by who holds them**
  BERT can learn investor types (token level)
  (suppose two hedge funds are "synonyms," then ...)
- supports OOS in time
  train BERT IS, feed new sentence to pre-trained model
  (underlying assumption: investor properties are stable)

[This is related to InvestorBERT, but the paper views it as a way to embed investors, not firms (still token-level embeddings). My proposal emphasizes sentence-level embeddings.]

Additionally

I think it is possible to encode structured sentences like
AAPL_{202409}:
[SPY, holding=\$2b, flow=+\$30m],
[ARK, holding=\$1b, flow=−\$10m], ...
with text-numerical mixed inputs.

This is very valuable for applying nlp tools for finance, which have more structured data.

Back to the big picture

- quantity data
- AI/ML

Valuable tools for economics research

Back to the big picture

- quantity data
- AI/ML

Valuable tools for economics research

- *Trading Volume Alpha*, with Ruslan Goyenko, Bryan Kelly, Tobias Moskowitz, Chao Zhang
  - trading volume prediction for after-cost portfolio optimization
  - neural networks and transfer learning

Back to the big picture

- quantity data
- AI/ML

Valuable tools for economics research

- *Trading Volume Alpha*, with Ruslan Goyenko, Bryan Kelly, Tobias Moskowitz, Chao Zhang
  - trading volume prediction for after-cost portfolio optimization
  - neural networks and transfer learning

- *Quantity, Risk, and Return*, with Yu An and Chen Wang
  - factor exposure (β) and flow-induced factor quantity (q) together explain the cross-section of expected returns
  - BTQ model (β times quantity)

Back to the big picture

- quantity data
- AI/ML

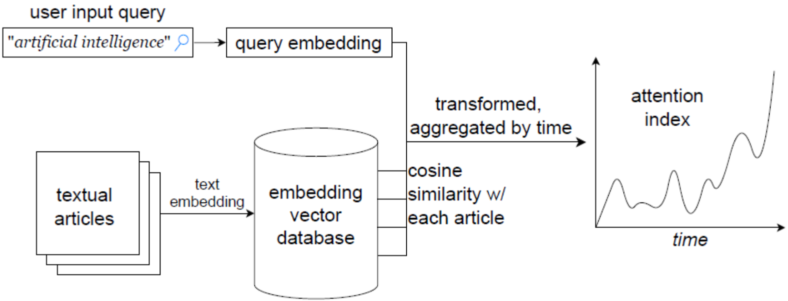
Valuable tools for economics research

- *Trading Volume Alpha*, with Ruslan Goyenko, Bryan Kelly, Tobias Moskowitz, Chao Zhang
  - trading volume prediction for after-cost portfolio optimization
  - neural networks and transfer learning

- *Quantity, Risk, and Return*, with Yu An and Chen Wang
  - factor exposure (β) and flow-induced factor quantity (q) together explain the cross-section of expected returns
  - BTQ model (β times quantity)

- *Tracking Narratives with LLM Embeddings* (in progress), with Leland Bybee and Jonathan Fan
  - one-stop shop for taming the "narrative zoo"
  - any narrative based on textual query, OpenAI's textual embeddings

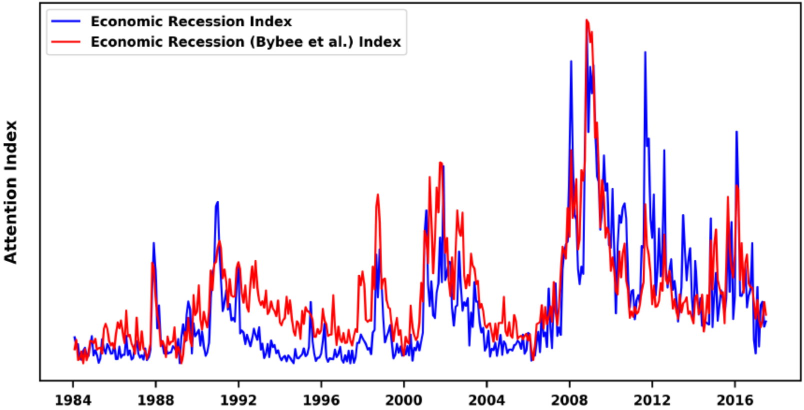
Tracking narratives with large language model embeddings (work in progress)



- any textual query
- web-based service open to all

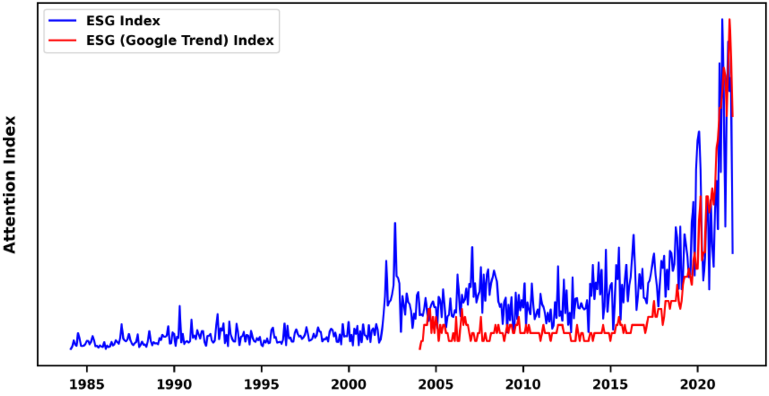
Example: recession

Figure 3: Replicating Economic Recession Index

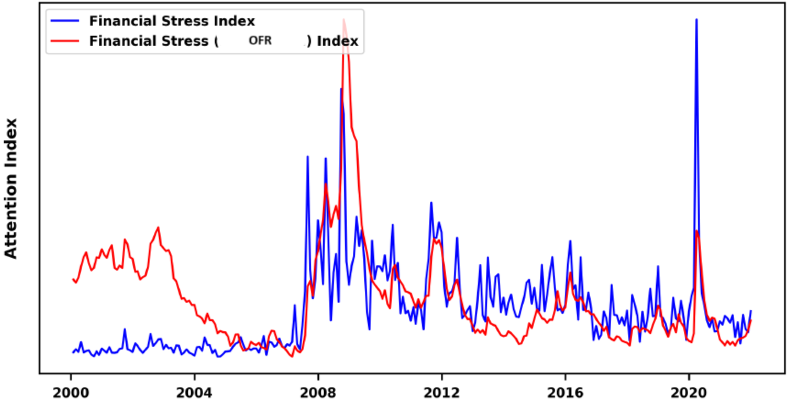


Example: ESG

Figure 5: ESG Index



Example: financial stress



Discussion:

Asset Embeddings

Xavier Gabaix, Ralph S.J. Koijen, Robert J. Richmond, Motohiro Yogo

Yinan Su
Johns Hopkins University Carey Business School

ABFR Webinar
on Sept. 26, 2022