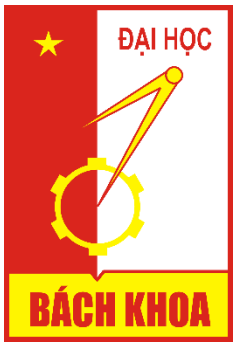
ĐẠI HỌC BÁCH KHOA HÀ NỘI
Trường Công Nghệ Thông Tin & Truyền Thông

----- □ 🕮 □ -----



Introduction to Data Science

Survey: **The developments of text data processing methods: From learning Bag-Of-Word to Transformer and Based-on Transformer**

Instructor: Dr. Nguyễn Đức Anh

Student:

Lê Hoàng Long	20232099M

Hà Nội, April 2024

Catalog

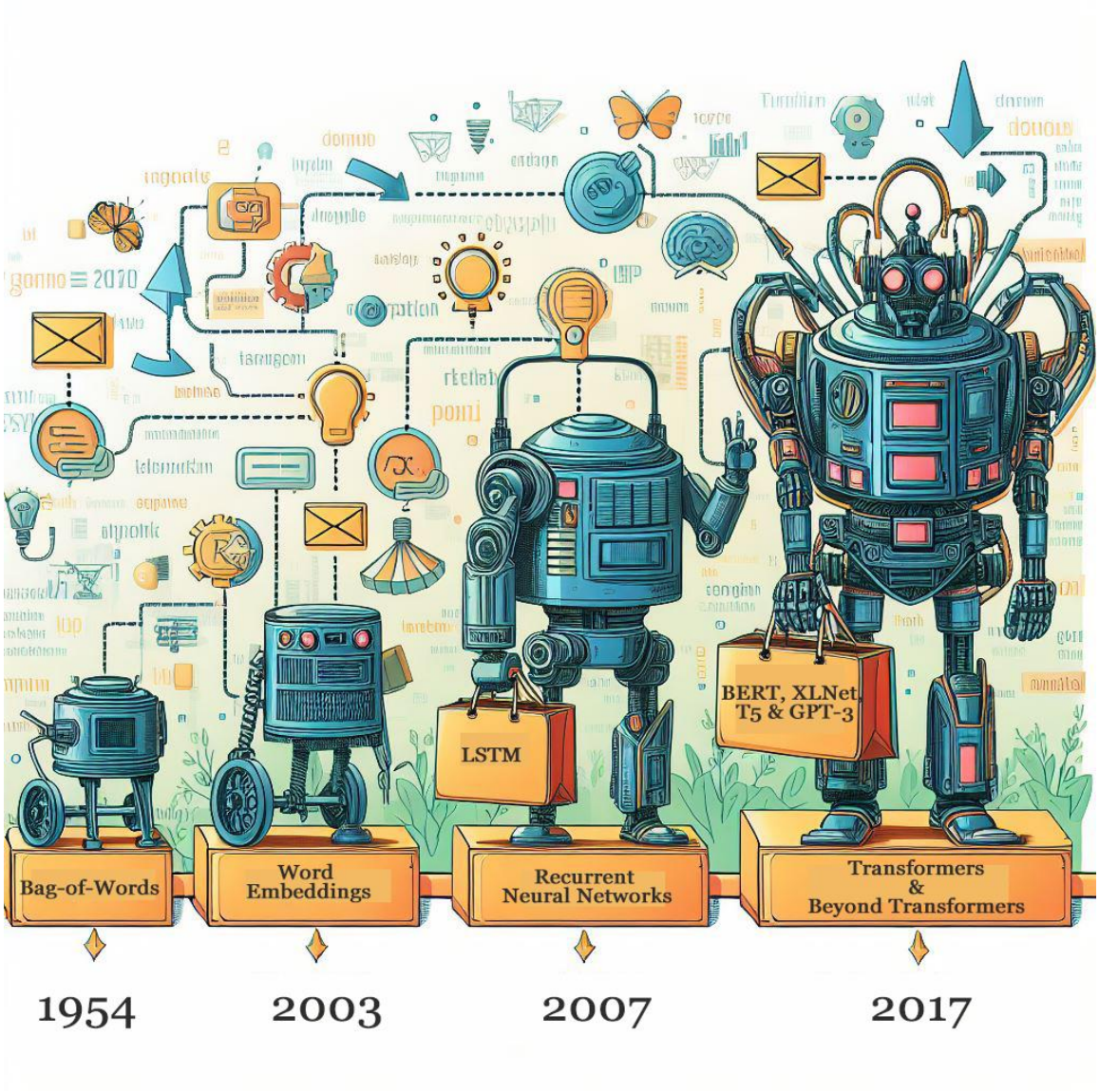
1. Introduction

The ability to understand and process human language is a fundamental challenge in computer science.Text data processing methods have undergone a remarkable transformation, progressing from rudimentary techniques to powerful deep learning models.

This survey explores this exciting journey, tracing the evolution from basic word counting to the cutting-edge world of Transformers and beyond. We will explore the strengths and limitations of each approach, highlighting the significant advancements that have propelled us towards a deeper understanding of textual data.

2. Discussion

a) A picture is worth a thousand words



b) The evolution from simple techniques to powerful deep learning models.

i. **Bag-of-Words (BoW)**: This early approach treated text as a "bag" of individual words, ignoring grammar and order. Words were simply counted to represent the document. While simple, BoW struggled to capture the context and relationships between words.

1. A picture is worth a thousand words

2. Image descriptions:

a) Imagine a bag filled with colorful balls. Each ball represents a unique word in a document, and the color can represent different categories (e.g., nouns, verbs).

b) The number of balls of each color reflects how often those words appear in the document. While the order of the balls doesn't matter (like the order of words in the BoW model), the overall color distribution captures the essence of the document.

3. How it works

a) BoW models treat text as an unstructured assortment of words, ignoring word order and context.

b) Each unique word becomes a separate dimension in a vector space.

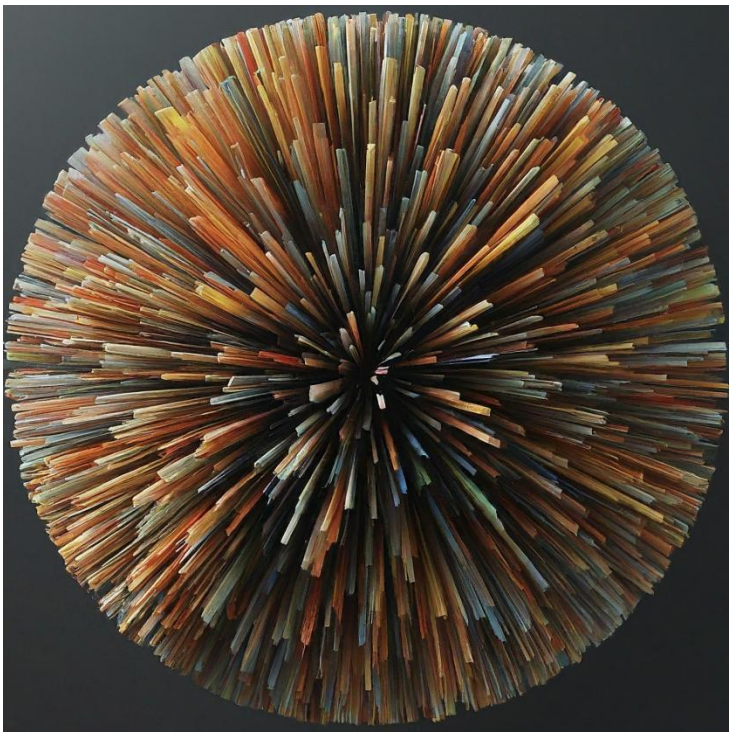
c) Text documents are represented as points in this vector space, with each dimension corresponding to the word’s frequency in the document.

d) Example: If we have two documents (“A rose is red, a violet is blue” and “My love is like a red, red rose”), their BoW vectors might look like (1, 1, 1) and (2, 1, 0), respectively.

4. Use Cases: BoW is commonly used in text classification, information retrieval, and sentiment analysis

ii. **Word Embeddings**: This breakthrough introduced a way to represent words as vectors in a high-dimensional space. Words with similar meanings ended up closer in this space, capturing semantic relationships.

1. A picture is worth a thousand words



2. Image descriptions:

a) Imagine a map where words are like cities. Words with similar meanings are located closer together on the map, reflecting their semantic relationships.

b) Word embeddings use mathematical techniques to create these maps, where distances between words capture how similar they are in meaning.

3. Popular Models:

a) Word2Vec: Learns word embeddings by predicting context words based on a target word.

b) GloVe (Global Vectors for Word Representation): Constructs word vectors using global co-occurrence statistics.

4. Advantages:
    a) Captures semantic relationships between words.
    b) Useful for downstream NLP tasks.
5. Limitations:
    a) Ignores word order.
    b) Fixed-size vectors for each word
6. Example: Word2Vec embeddings can represent “king - man + woman” as a vector close to “queen.”

iii. **Recurrent Neural Networks (RNNs)**: These models could process sequences of words, accounting for order and context. RNNs like LSTMs were particularly adept at handling long sequences of text, useful for tasks like machine translation and sentiment analysis.

1. A picture is worth a thousand words

2. Image descriptions
    a) Think of a conveyor belt with information packets traveling on it. Each packet represents a piece of data (like a word in a sentence). The RNN is like a processing station along the belt.

b) It analyzes each packet considering the information from previous packets (like the context of words in a sentence) before passing it on, potentially modified, to the next station.

3. Structure:

a) RNNs consist of interconnected layers similar to other neural networks.

b) However, they have a loop within the hidden layer that allows information to persist across time steps. This loop enables the network to maintain a state that captures the history of the sequence it's processing.

4. Types of RNNs: There are various RNN architectures, each addressing limitations of the basic RNN:

a) Long Short-Term Memory (LSTM): LSTMs are a popular RNN variant specifically designed to overcome the vanishing gradient problem. LSTMs have internal gating mechanisms that control the flow of information, allowing them to learn long-range dependencies in sequences. This makes them well-suited for tasks like machine translation and speech recognition.

b) Gated Recurrent Unit (GRU): GRUs are another variation that address the vanishing gradient problem with a simpler gating mechanism compared to LSTMs. They achieve good performance while requiring less computational power.

iv. **Transformers**: This architectural innovation revolutionized NLP. Unlike RNNs, Transformers can analyze all words in a sentence simultaneously, capturing long-range dependencies more effectively. Transformers led to significant advancements in various tasks, including machine translation, text summarization, and question answering.

1. Key Features:

a) Self-attention mechanism: Instead of relying on recurrent or convolutional layers, the Transformer uses self-attention to capture contextual information across the entire input sequence.

b) Positional encoding: To account for the order of words in a sequence, positional encodings are added to the input embeddings.

c) Multi-head attention: The model computes multiple attention heads in parallel, allowing it to attend to different parts of the input.

d) Encoder and decoder stacks: Transformers consist of an encoder stack (for input representation) and a decoder stack (for output generation in tasks like machine translation).

2. Models:

a) BERT (Bidirectional Encoder Representations from Transformers):

i. A picture is worth a thousand words

ii. Image descriptions

1. Imagine a series of stacked boxes which is similar to the encoder in the Transformer model. However, in BERT, these boxes process text data differently

2. Bidirectional Processing: Unlike the standard Transformer encoder that processes information sequentially (left to right), BERT considers both the preceding words (context before) and following words (context after) for each word in the sequence. Arrows in the image would illustrate this bidirectional flow of information. This allows BERT to capture richer meaning and context compared to traditional sequential models.

3. Stacked Layers: Each box represents a layer in the BERT model. As information flows through these layers, the model progressively refines its understanding of the text's overall meaning.

b) GPT (Generative Pre-trained Transformer):

i. A picture is worth a thousand words

ii. Image descriptions

1. These nodes represent the artificial neurons in the GPT's many layers.
2. Neural Network Structure: The interconnected layers symbolize the core of GPT's architecture. Information flows through these layers, allowing the model to process and analyze text data.
3. Information Flow: Visual elements within the image would likely depict arrows tracing the pathway information takes as it travels through the network. This highlights how GPT analyzes the input text, considers the relationships between words, and builds an understanding of the context.
4. Pre-training on Massive Text Data: The image wouldn't directly show text data, but it could imply a vast amount of information by including elements representing books, articles, or code. This signifies GPT's training on a massive dataset of text and code, allowing it to learn the intricacies of language.
5. Generative Capabilities: The image might incorporate elements like different writing styles or languages. This signifies GPT's ability to generate new text formats, translate languages, write different creative content, and answer your questions in an informative way, all based on the knowledge it has acquired from its pre-training

v. **Beyond Transformers**: Research is constantly pushing the boundaries. Transformer-based models like BERT, XLNet, and T5 are being fine-tuned for specific tasks, achieving even

greater accuracy. Additionally, new architectures like GPT-3 are exploring generative capabilities, allowing for tasks like creative text writing and code generation.

1. Examples:
    a) RoBERTa (A Robustly Optimized BERT Pretraining Approach): Improves BERT by using more training data and removing sentence order prediction.
    b) T5 (Text-to-Text Transfer Transformer): Casts all NLP tasks as text-to-text problems.
    c) ELECTRA (Efficiently Learning an Encoder that Classifies Token Replacements Accurately): Pre-trains a generator and discriminator for denoising tasks.
2. Advantages:
    a) State-of-the-art performance.
    b) Fine-tuning for specific tasks.
3. Limitations:
    a) Requires large-scale pre-training.
    b) Computationally expensive.

3. Conclusion

a) The evolution of text data processing methods paints a remarkable picture of progress. From the rudimentary bag-of-words approach to the sophisticated world of Transformers, we've witnessed a dramatic leap in our ability to extract meaning from textual data

b) Transformers and their dominance are undeniable. Their ability to analyze entire sentences simultaneously has revolutionized tasks like machine translation and question answering. However, the quest for even more sophisticated methods continues.