Computing For Medicine
CSE585/BIO546
Mid Semester Examination

Instructions

1. **Exam Duration: 1 hour**
2. **Maximum Marks: 100**
3. **Only one option in MCQs is correct**
4. **MCQs have to be submitted on the Answer Sheet, responses marked on the Question Paper will NOT be considered.**

Section A(60 marks)

Q1- Which of the following correctly represents the formula for scaled dot-product attention (ignoring bias terms)? (4 Marks)

a) Softmax($\frac{QK}{\sqrt{d_k}}$)V

b) Softmax($\frac{KQ^T}{\sqrt{d_k}}$)V

c) Softmax($\frac{d_k Q}{\sqrt{K}}$)V

d) Softmax($\frac{QK^T}{\sqrt{d_k}}$)V

Q2- Which of the following is **NOT true** about BERT? (4 Marks)

a) BERT representations can be transferred to downstream NLP tasks with minimal task-specific architecture changes.

b)BERT can be fine-tuned for tasks such as text classification, question answering, and named entity recognition.

c) **BERT processes tokens strictly in a left-to-right autoregressive manner during pretraining.**

d) BERT's architecture allows contextual embeddings where the same word can have different vector representations depending on surrounding context.

Q3- What in Transformers directly enables learning of weights? (4 Marks)

a) Skip connections

b) Feed-forward layer

c) Multi-head Attention

d) Positional Encoding

Q4- Which of the following is a disadvantage of one-hot encoding for medical text data? (4

Marks)

a) It introduces semantic ambiguity by mapping similar medical terms to the same vector

b) **It results in extremely high-dimensional sparse vectors, making computation and storage inefficient.**

c) It inherently preserves the temporal sequence of symptoms in clinical narratives.

d) It naturally captures synonym relationships between rare medical terms

Q5- Which of the following statements correctly distinguishes the Continuous Bag-of-Words (CBoW) model from the Skip-gram model in word2vec architecture? (4 Marks)

a) CBoW captures word order and contextual meaning, while Skip-gram ignores context.

b) Skip-gram captures word order and contextual meaning, while CBoW ignores context.

c) **CBoW is trained to predict surrounding the middle word given a context while Skip-gram predicts context given the middle word.**

d). CBoW has positional encoding, whereas skip-gram does not

Q6- In Transformers, what is the primary purpose of **positional encoding**? (4 Marks)

a) To introduce token-level embeddings that are trainable

b) To normalize the output of attention layers

c) To create skip connections for gradient flow

d) To provide information about the order of tokens in a sequence

Q7- In the self-attention mechanism, what is the purpose of the scaling factor $\frac{1}{\sqrt{d_k}}$ in the attention computation? (4 Marks)

a) To prevent vanishing gradients during backpropagation

b**) To normalize the dot products and prevent extremely large values before softmax**

c) To maintain positional information in the sequence

d) To enable parallel processing of multiple attention heads

Q8-The primary advantage of using residual connections (skip connections) in deep neural networks is to … (4 Marks)

a) prevent vanishing gradients during backpropagation

b) normalize the dot products and prevent extremely large values before softmax

c) maintain positional information in the sequence

d) enable parallel processing of multiple attention heads

Q9- Which of the following is a limitation of TF-IDF representation? (4 Marks)

a) It captures semantic relationships between words effectively

b) It normalizes the term frequencies by their frequency in the documents

c) It fails to capture semantic similarity between documents

d) It produces dense vector representations

Q 10- Based on the "Vector Space Paradigm: Distributional Hypothesis," what is the core idea behind modern word representations?(4 marks)

a. Words with similar meanings are represented by unrelated vectors.

b. The meaning of a word is determined by the words that surround it.

c. The syntactic structure of a sentence is more important than the words themselves.

d. Words can only be represented by a single integer value.

Q11 - Given the following vocabulary and sentence, convert the sentence into One-Hot Encoding and Bag of Words representations (4 marks)

Vocabulary: { "the": 0, "cat": 1, "sat": 2, "on": 3, "mat: 4, "dog": 5, "barked":6 }

Sentence: “dog barked on cat”.

Ans **One-Hot Encoding (per word):**

dog → [0,0,0,0,0,1,0]

barked → [0,0,0,0,0,0,1]

on → [0,0,0,1,0,0,0]

cat → [0,1,0,0,0,0,0]

Bag of Words vector for whole sentence:

[0,1,0,1,0,1,1]

Q12- Consider the sentence: "the quick brown fox jumps over the lazy dog". Using a Skip-Gram model with a context window size of 3, list all the training pairs where the target word is "fox". (4 Marks)

Ans Context: {quick, brown, the, jumps, over}

Training pairs (target → context):

- (fox, the)

- (fox, quick)
- (fox, brown)
- (fox, jumps)
- (fox, over)

5 pairs total.

Q 13- In SNOMED CT, which statement is correct?(4 marks)

a. Each clinical idea is represented only by its preferred term.
b. Concepts are language-dependent and vary across regions.
c. **Every concept has a unique, language-independent identifier and is linked to multiple descriptions.**
d. Descriptions, not concepts, form the foundation of SNOMED CT.

Q14) In SNOMED CT, relationships such as “finding site” or “causative agent” are used to: (4 marks)

a. Provide billing information for clinical procedures.
b. Link concepts into a computable ontology that supports clinical reasoning.
c. Store synonyms for terms in multiple languages.
d. Identify patient demographics such as age and gender.

Q15) Which of the following is an example of post-coordination in SNOMED CT? (4 marks)

a. Selecting a single concept ID for “Type 2 Diabetes Mellitus with Increased Urination Frequency.”
b. Linking two concepts such as Type 2 Diabetes Mellitus and Increased Urination Frequency, using a relationship to express a combined clinical idea.
c. Using only synonyms to capture different ways of saying “Diabetes.”
d. Mapping a SNOMED CT concept directly to an ICD-10 code.

SECTION B (40 Marks)

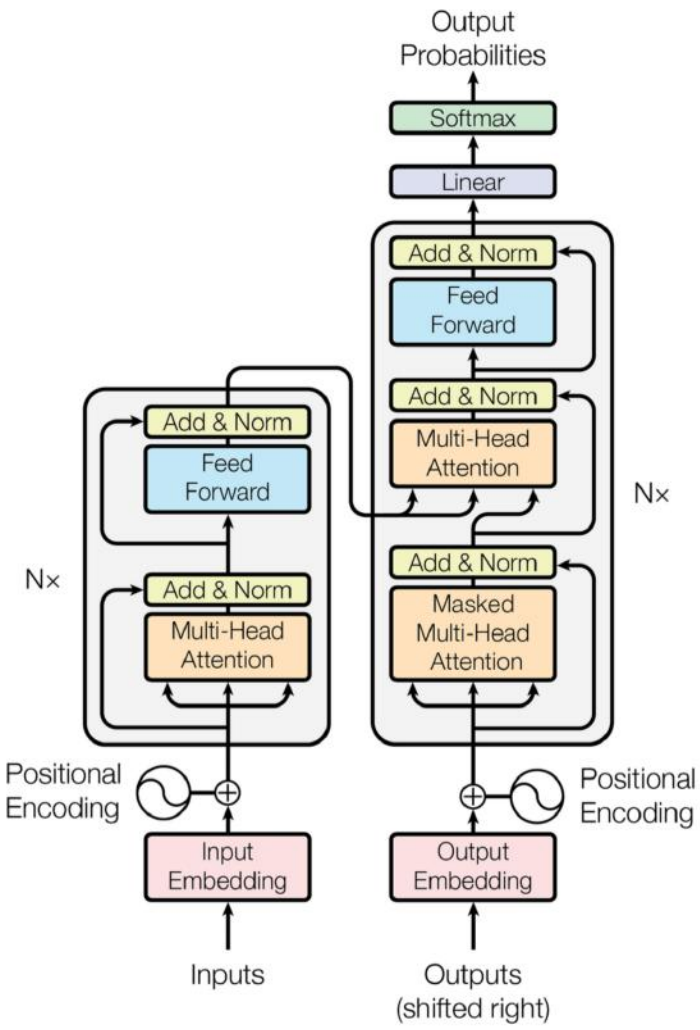
Q1- Discuss how integrating Transformer-based language models with UMLS thesaurus to enhance clinical decision support. Illustrate how UMLS information will be integrated into the block structure of transformer models. (10 Marks)

Ans Transformer-based language models, such as BERT or GPT at understanding contextual relationships in text. In clinical settings, they can process electronic health records (EHRs), clinical notes, and research articles. However, purely data-driven models may miss domain-specific medical knowledge, leading to incorrect or incomplete recommendations.

The UMLS (Unified Medical Language System) thesaurus is a **curated ontology** containing medical concepts (CUIs), synonyms, and relationships between concepts. Integrating UMLS with Transformers helps models understand and leverage structured clinical knowledge alongside unstructured text. This enhances Clinical Decision Support (CDS) by providing more accurate diagnoses, treatment suggestions, and risk predictions.

Integration for Clinical Decision Support (CDS):

By incorporating UMLS, transformers can better **map clinical terms to standardized medical concepts** (e.g., mapping "high blood sugar" to "Hyperglycemia"). This allows the model to reason not only on linguistic context but also on **domain knowledge**, leading to improved clinical entity recognition, relation extraction, and decision support tasks (e.g. diagnosis suggestions). The enriched model can reduce ambiguity, handle synonyms, and improve interpretability in CDS systems.



Clinical Note Text - "Patient has high blood sugar and frequent urination."

Mapped UMLS Concepts - Hyperglycemia (C0011849), Polyuria (C0034000)

- **Word embeddings:** ["Patient", "has", "high", "blood", "sugar", ...]
- **Concept embeddings:** [C0011849, C0034000]
- **Attention Layer:** Increase interaction between "high blood sugar" and "frequent urination" based on UMLS relation
- **Output:** Transformer predicts Type 2 Diabetes Mellitus with higher confidence, leveraging both textual context and UMLS knowledge

Methods of Integration

a) Concept Embedding Augmentation:

Clinical text is first mapped to UMLS concepts using a tool like MetaMap or QuickUMLS. Each concept is converted into an **embedding vector**, representing its meaning in a structured space. During Transformer training or inference, the model receives both word embeddings and concept embeddings, enriching its understanding of medical terminology.

b) Knowledge Graph Encoding:

UMLS contains relationships between concepts (e.g., "causes," "treats," "associated with"). These relationships can be encoded as a knowledge graph. Graph embeddings of connected concepts can be injected into the Transformer's attention layers, guiding the model to pay attention to medically relevant associations.

c) Attention Biasing using UMLS:

UMLS relationships can bias the Transformer's self-attention mechanism, increasing attention weights between words that correspond to related medical concepts. Example: If the note mentions "hyperglycemia" and "insulin," UMLS relationships inform the model that these are strongly related, allowing better context propagation in the Transformer layers.

A standard Transformer block has: Multi-Head Self-Attention → Add & Norm → Feed-Forward → Add & Norm. UMLS can be integrated in these ways:

Input Layer Integration: Combine word embeddings with UMLS concept embeddings as the initial input to the Transformer. This allows the model to understand both linguistic and medical knowledge from the first layer.

Attention Layer Integration: Modify the attention scores using UMLS knowledge. For tokens mapped to related concepts, increase attention weights proportionally to the strength of their relationship in UMLS. This ensures medically related terms interact more strongly during contextual encoding.

Intermediate Fusion Layers: Inject UMLS knowledge embeddings into the feed-forward layers as side inputs. Helps the model combine learned linguistic patterns with curated medical knowledge before producing final outputs.

Illustration in Block Structure:

Input = [Word Embedding + Position Encoding + UMLS Concept Embedding]
Transformer Encoder Layers = Self-Attention (uses both text + concept similarity) → Feed Forward → Normalization
Output = Knowledge-enriched representation mapped to CDS task (diagnosis, drug recommendation, etc.)

Benefits for Clinical Decision Support:

- Improves diagnosis suggestions by connecting symptoms to related conditions.
- Enhances treatment recommendations using concept relationships (drug-disease interactions).
- Reduces ambiguity in medical language, handling synonyms, abbreviations, and variations in terminology.
- Facilitates explainable AI, as UMLS relationships can justify why the model made certain recommendations.

In summary, integrating UMLS into transformers creates a **knowledge-enhanced model** that combines contextual understanding with domain-specific semantics, thereby making clinical decision support systems more accurate, interpretable, and reliable.

Q2- Explain the concept of positional encoding (5) and look-ahead masking (5) in a Transformer model. (10 Marks)

Ans Positional Encoding

Transformers do not have a built-in notion of word order because they process all tokens in a sequence simultaneously. **Positional encoding** is added to the input embeddings to provide information about the position of each token in the sequence.

It encodes position information using **sinusoidal functions** (sine and cosine) or learned embeddings.

Why is positional encoding necessary in Transformer models

Positional encoding is necessary because the Transformer model relies purely on self-attention mechanisms, which do not inherently capture the sequential order of tokens. Unlike RNNs and CNNs that naturally handle sequential information, attention-based models process all tokens in parallel without considering their positions. Positional encoding provides each token with information about its position in the sequence, enabling the model to capture the order of tokens and effectively learn relationships between them in a context-aware way.

How does the Transformer use sine and cosine functions for positional encoding?

In Transformers, positional encoding for each token position pos is calculated using sine and cosine functions at different frequencies. For each dimension i in the positional encoding vector, the encoding is given by:

$PE(pos,2i)=sin(pos/100002i/dmodel)$

$PE(pos,2i+1)=cos(pos/100002i/dmodel)$

Here, pos represents the position of the token, and i represents the dimension. The use of different frequencies (scaling the position with exponential powers of 10,000) allows the model to capture relative position information across different dimensions. The sine and cosine values at different frequencies help encode position information such that any offset between two positions can be computed linearly, aiding the model in learning relative distances between tokens.

Purpose

Helps the model understand word order, which is crucial in language tasks (e.g., “The cat chased the dog” vs. “The dog chased the cat”). For the sentence “Patient has fever”, the words are assigned positions 1, 2, 3. Positional encoding tells the model that “Patient” comes first, “has” second, and so on, even though all tokens are processed in parallel.

Look-ahead masking is used in the **decoder** during training to prevent the model from "seeing" future tokens when predicting the next token. Transformers use self-attention, which lets every token in a sequence "look at" all other tokens to gather context. This works great for understanding text all at once, but it's a problem for language generation, where we predict one token at a time.
Purpose: Prevents a token from "seeing" future tokens when predicting the next token in a sequence. When generating text, the model predicts the next word based on previous words. If the model could see future words during training, it would "cheat," learning from answers it shouldn't know yet Look-ahead masking prevents this by hiding future tokens in the sequence.

1. Why it's needed:

Transformers use attention, which lets every word in a sequence "look at" every other word. This is great for understanding context but can be a problem in sequence generation tasks (like translating a sentence or writing the next word in a sentence).
If the model could see future words, it would cheat. For example:
Input: I love to ___ pizza
Without masking, the model could see eat in the future and just copy it, instead of learning proper patterns.
Imagine a sentence:
["Patient", "has", "fever"]
During training, when predicting each word:

- To predict "Patient", it sees only itself.
- To predict "has", it can see "Patient" but not "fever".
- To predict "fever", it can see "Patient" and "has," but nothing beyond.

The mask is applied as a triangular matrix where entries above the diagonal are blocked. This ensures no future information leaks.

3. Where it fits in the Transformer:

It's applied in the decoder during training.
The attention layer in the decoder uses this mask to block information from future words.
Helps the model learn causal relationships: what comes next depends only on what has already happened.

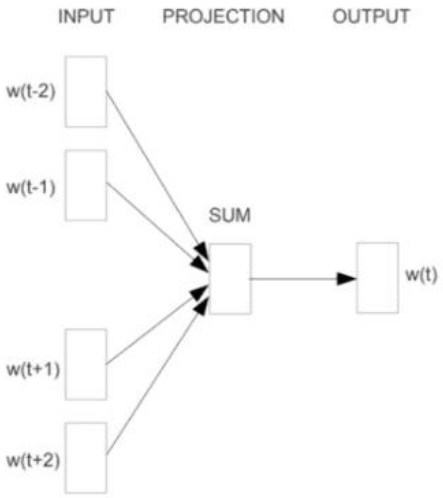
Q3. Compare and contrast the **S**kip-Gram and Continuous Bag of Words (CBOW) architectures in Word2Vec. Your answer should:

1. Illustrate the objective function of each model.

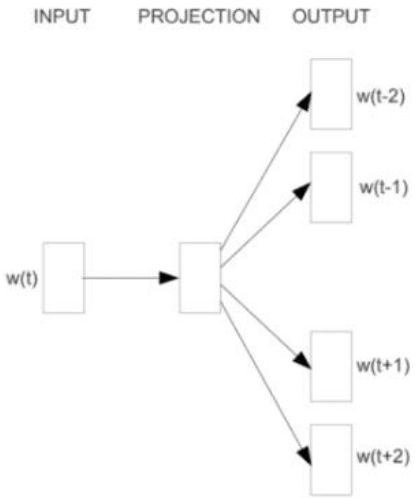
2. Illustrate with a toy example sentence from clinical text on how training pairs are generated differently.
   (10 Marks)

Ans Word2Vec creates vector representations of words, so words that appear in similar contexts get similar vectors.

Feature	CBOW	Skip-Gram
Goal	Predict the target word from surrounding context words	Predict surrounding context words from the target word
Input	Context words (words around the target)	Single target word
Output	Target word	Context words
Training	Faster	Slower
Used case	Capturing general semantic meaning	Capturing richer relationships, especially for rare terms



CBOW



Skip-gram

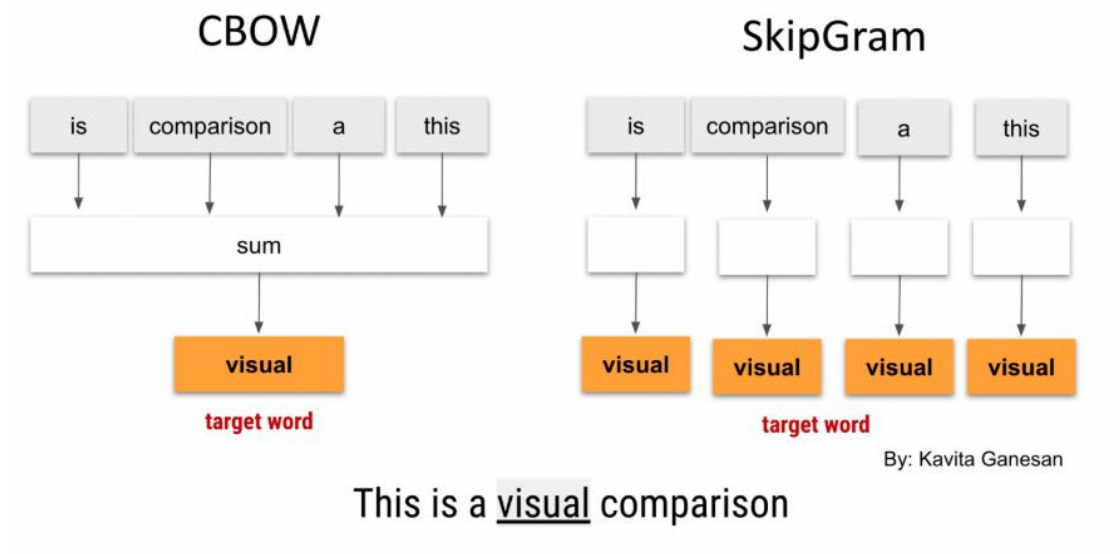


Figure 2: Difference between SkipGram and CBOW training architectures .

How they work (conceptually)

CBOW: Looks at the surrounding words and tries to guess the word in the middle.

Skip-Gram: Takes the middle word and tries to guess the words around it.

Toy Example: Clinical Sentence

Sentence: "Patient shows high blood pressure"

Context window: 2 words on each side that is a total of window size = 5

CBOW (context → target):

Use surrounding words to predict the target:

- Context: "Patient", "shows" → Target: "high"
- Context: "shows", "high" → Target: "blood"
- Context: "high", "blood" → Target: "pressure"

Skip-Gram (target → context):

Use the target word to predict surrounding words:

- Target: "high" → Context: "Patient", "shows", "blood", "pressure"
- Target: "blood" → Context: "shows", "high", "pressure"

Key Difference:

CBOW: “Guess the missing word from neighbors.”

Skip-Gram: “Guess neighbors from this word.”

Q4. A group of students suggests that with an already deidentified data such as MIMIC it should be possible to simply copy the dataset onto a pen drive or share it through Google Drive links for research collaboration. Write an essay on why this should not be done? What are challenges in the era of LLMs with larger context data being processed in this scenario? (10 marks)
Ans Why Sharing Deidentified Data Like MIMIC Freely is Risky
MIMIC (Medical Information Mart for Intensive Care) is a widely used clinical dataset containing detailed patient information from ICU stays. Even though it is deidentified to protect patient privacy, it still contains sensitive information about real patients’ diagnoses, treatments, and outcomes. Some students might think that because the data is anonymized, it is safe to copy onto pen drives or share via cloud platforms like Google Drive. However, this is not advisable, and doing so can lead to serious ethical, legal, and technical consequences.

1. Risks of Sharing Deidentified Data
   - Reidentification Risk: Deidentification removes direct identifiers (names, phone numbers, addresses), but quasi-identifiers like age, gender, admission date, or rare diagnoses can still be used to reidentify patients. Advanced techniques and cross-referencing with publicly available data can sometimes reveal patient identities.
   - Violation of Data Use Agreements (DUAs): Access to MIMIC data is granted only after signing a DUA, which prohibits unauthorized sharing. Sharing data on pen drives or Google Drive would breach these agreements, potentially resulting in loss of access, legal action, or institutional penalties.
   - Ethical Responsibility: Even anonymized data represent real patient experiences. Researchers have a duty to handle such data securely and ensure it is used only for approved research purposes.

2. Challenges in the Era of Large Language Models (LLMs)
   - Increased Reidentification Risk via LLMs: Modern LLMs can process and memorize large amounts of data. If sensitive clinical datasets are uploaded or shared insecurely, LLMs could inadvertently learn or leak patient-specific information. Even anonymized text may contain patterns that LLMs can exploit to reconstruct private information.
   - Large Context Windows: LLMs can process long sequences of text, meaning more information from the dataset can be exposed in one go. This increases the chance of privacy breaches if shared outside secure environments.

- Data Leakage Through Collaboration Tools: Cloud services, email, or shared drives may have weak security, making them vulnerable to unauthorized access or hacking. Once clinical data leaves a controlled environment, it is almost impossible to guarantee it remains secure.

3. Best Practices for Handling MIMIC and Similar Data

- Use secure institutional servers with controlled access for storing and processing data.
- Follow the MIMIC Data Use Agreement strictly.
- Share only derived results or aggregate statistics, not raw patient-level data.
- Avoid uploading sensitive data to external services, including Google Drive, unless encrypted and approved by the institution.

Conclusion

Even deidentified datasets like MIMIC cannot be treated as completely risk-free. Simply copying them to a pen drive or sharing via Google Drive is unsafe, unethical, and often illegal. The advent of LLMs, which can process large contexts and potentially memorize sensitive details, amplifies the risks of reidentification and data leakage. Responsible handling, secure storage, and adherence to legal and ethical guidelines are essential for protecting patient privacy while enabling research.