## Week 1 – Word Embeddings

1. What is true about one-hot encoding?
   1. One-hot encoding vector is a multi-dimensional vectors of all 0s except for one dimension of 1.
   2. Distance between any two one-hot encoding vectors is always the same.
   3. One-hot encoding is commonly used to encode categorical data.
2. What is true about Word2vec?
   1. There are two different algorithms for Word2Vec: Skip gram and Continuous bag of words (CBOW).
   2. Larger context window means words that are farther apart will be considered similar in the skip gram calculation.
   3. Negative sampling treat the pairs of context words and the target word as positive examples, and pairs of random words and the input target word as negative examples.

It is very expensive to compute due to the involvement of all words.

1. What are suitable medical applications of Word2vec?
   1. Similarity search: To find similar medical concepts based on the Euclidian distance of the word2vec embeddings of those medical concepts.
   2. Algebra operation: To perform summation and subtraction of word2vec embeddings in order to better understand a combination of medical concepts.
   3. Features for predictive modelling: To utilize word2vec embeddings as input feature vectors to support downstream classification or regression models.
2. What are the challenges in visualising high-dimensional data?
   1. To preserve the relationship of original data when visualizing in 2D space.
   2. To ensure similar data points mapped to nearby location in 2D space.
   3. To provide computationally efficient methods to produce the visualization.
3. What is true when comparing PCA and t-SNE?
   1. PCA is a dimensionality reduction algorithm that tries to preserve global distance between all pairs of data points.
   2. t-SNE tries to project high-dimensional data into 2D space while preserving local distance.
   3. t-SNE is more appropriate for visualization high-dimensional data
4. What is true about t-SNE method?
   1. Input distribution utilizes a Gaussian kernel
   2. Output distribution is the student t distribution with 1 degree of free.
   3. The objective function is KL divergence between input and output distributions.
5. What is true about the Med2Vec method?
   1. It models two level hierarchy of medical data namely visit level and patient level.
   2. Med2Vec is designed to model sequences of clinical codes.
   3. Mec2Vec is a generalization of word2vec for electronic health record data.
6. Which is true about MiME: Multilevel medical embedding method?
   1. It leverages the dependency from diagnosis to treatments.
   2. It models EHR data with multi-llevel hierarchy (treatment -> diagnosis -> visit - patient).
   3. It introduces auxiliary prediction task to predict diagnosis and treatment within a visit.

## Week 2 – Convolutional Neural Networks

1. What is NOT true about convolutional neural networks (CNN)?
   1. CNN usually has many more parameters than fully connected neural networks.
2. Given input sequence = [1, 2, 3, 4, 5, 6] and a filter [1 2 1], what is the output of the convolution operation with stride 2 with no padding?
   1. [8, 16]
3. What is the size of the output feature map given 6x6 input and two filters of 4x4 with stride 1 and no padding?
   1. 3x3x2
4. What is the number of parameters for a convolutional layer given 6x6 input and two filters of 4x4 with stride 1 and no padding?
   1. 34

Parameters per filter = (height x width x number of channels)  
Total parameters = (parameters per filter x number of filters)

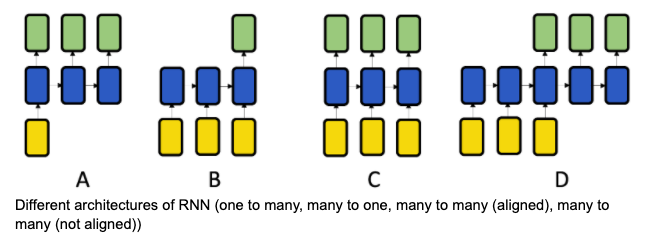
1. What is the output of 1D max pooling over input [1, 2, 3, 2, 3, 5] with filter of size 2 and stride 2?
   1. [2, 3, 5]
2. What is NOT true about AlexNet?
   1. AlexNet uses a larger filter of 11x11, which is computationally more efficient than the subsequent CNN models such as VGG and ResNet.
3. What is NOT true about CNN architectures?
   1. Inception Net uses 1x1 filters to adjust the output size so that the dimensions across different parallel paths are the same.

The purpose of a 1x1 filter is to reduce the dimensions before applying filters.

1. Which of the following is true about CNN models?
   1. CNN can be applied to both images and text data.
2. Why did the CNN model succeed in diabetic retinopathy detection application?
   1. Many standardized retina images exist.
   2. Many labelled images exist.
   3. Multiple labels are collected to generate gold standard labels.
3. What neural network architecture was used to detect skin cancer in the [Esteva 2017] paper?
   1. Inception net v3.

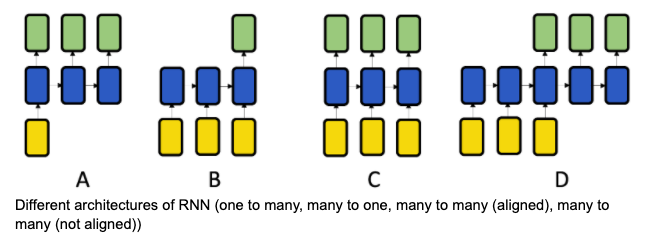
## Week 3 – Recurrent Neural Networks (RNNs)

1. What data are suitable for RNN to model?
   1. Longitudinal medical records
   2. Time series such as electrocardiogram
2. Which of the following architectures is best for x-ray image report generation?



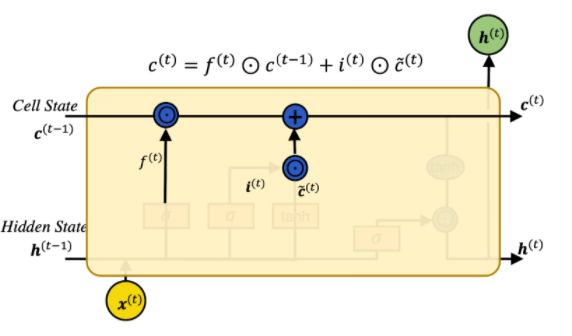
* 1. One to many
  2. Many to one
  3. Many to many of the same sequence length
  4. Many to many of different sequence lengths

1. Which of the following architectures is best for sequential diagnosis prediction is (i.e., predicting the disease diagnosis of the current visit based on a patient’s visit history)?



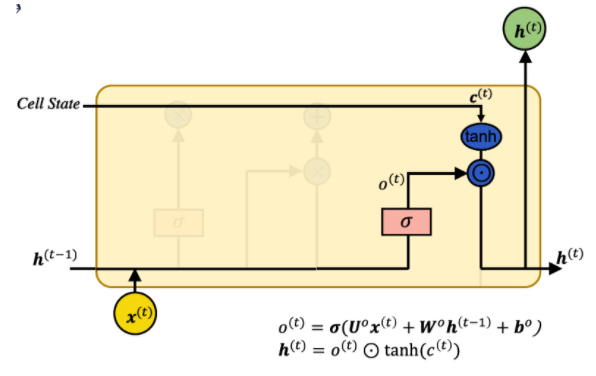
* 1. One to many
  2. Many to one
  3. Many to many of the same sequence length
  4. Many to many of different sequence lengths

1. What is the potential issue of backpropagation through time (BPTT) algorithm?
   1. Computational complexity
   2. Gradient explosion problem
   3. Vanishing gradient problem
   4. Curve of dimensionality
2. In LSTM model, what is the in the equation for updating cell state?



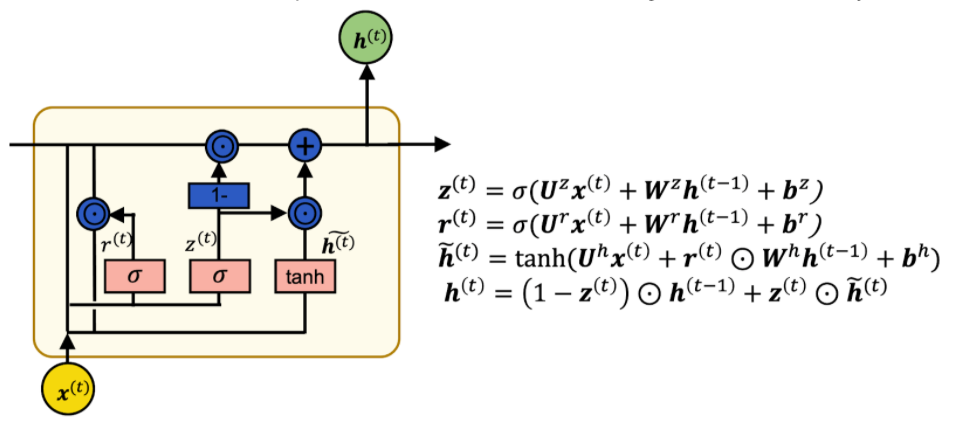
* 1. Input gate
  2. Forget gate
  3. Update gate
  4. Update to the cell state

1. In LSTM model, what is the in the equation?



* 1. Update gate
  2. Output gate
  3. New cell state
  4. New hidden state

1. In the GRU model, can we replace the tanh activation function with sigmoid activation?

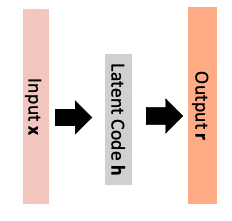


* 1. Yes
  2. No

1. What is true about bidirectional RNNs?
   1. Two RNN models are trained from the opposite directions?
   2. Hidden states of two RNNs are concatenated before predicting the output
   3. Bidirectional RNNs are suitable for modelling languages.
   4. Two RNN models have to be trained separately.
2. What is NOT true about the Seq2seq model?
   1. Seq2seq model requires training two RNNs: one for encoding and another for decoding.
   2. Seq2seq models can be used for machine translation applications.
   3. Encoder RNN produces a context vector c which will be used as part of the input in the decoder RNN.
   4. Seq2seq model ensures input sequence and output sequence to be the same length.
3. What is true about RNN applications to healthcare?
   1. We need training data of the same length to train RNN models.
   2. RNN model can be used to model longitudinal EHR data.
   3. The output of RNN models can be binary or multi-class classification.
   4. RNN can also be used to model clinical notes.

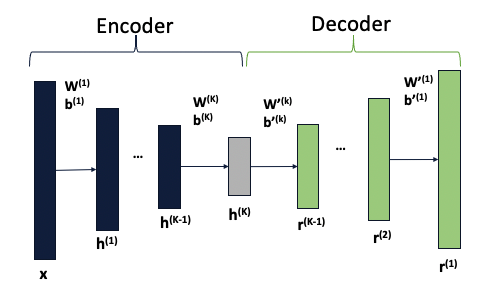
## Week 4 – Autoencoders

1. What is the most analogous method to Autoencoders?
   1. K-means clustering
   2. Principal component analysis
   3. Support vector machine
   4. Hierarchical clustering
2. Which of the following is NOT true about autoencoders?
   1. It is an unsupervised method.
   2. It is a lossless compression technique.
   3. It is a dimensionality reduction method.
   4. It is a feedforward neural network.
3. What is NOT true about sparse autoencoder?



* 1. It introduces sparsity in the latent code **h**.
  2. Sigmoid activation function is used to produce latent code **h**.
  3. The same loss function to the standard autoencoder is used for sparse autoencoder.
  4. Sparsity level on each dimension of **h** needs to be specified.

1. What is true about denoising autoencoder?
   1. It adds random noise to the original input before applying the autoencoder model.
   2. Its loss function is between reconstruction and the original input x without adding random noise.
   3. It is more expensive to train because of random noise added to the original input.
   4. It is more robust against noises thanks to the introduction of corrupted input.
2. What are the model parameters first learned in a stacked autoencoder?



1. What is true about stacked autoencoder?
   1. It applies multiple encoders first, then applies the corresponding decoders in reverse orders.
   2. It is a deep neural network of 2K layers where K is the number of autoencoders.
   3. It is trained in an end-to-end fashion as a single model using backpropagation.
   4. It is trained sequentially as K separate autoencoders.
2. What other model is used before applying autoencoder in the phenotype discovery paper?

*Lasko et al. 2013. “Computational Phenotype Discovery Using Unsupervised Feature Learning over Noisy, Sparse, and Irregular Clinical Data.” PloS One 8 (6):e66341*.

* 1. CNN model
  2. RNN model
  3. Gaussian process
  4. ARIMA model

1. What is the variant of autoencoder model used in the deep patient paper?

*Miotto, Riccardo, Li Li, Brian A. Kidd, and Joel T. Dudley. 2016. “Deep Patient: An Unsupervised Representation to Predict the Future of Patients from the Electronic Health Records.” Scientific Reports 6 (May):26094*.

* 1. Stacked denoising autoencoder
  2. Sparse autoencoder
  3. Contractive autoencoder
  4. Variational autoencoder