**Part A: Literature Exploration and Comparison**

**Objective:** Explore the application of CNN/RNN/Transformer networks in specific domains, identify three significant papers, and conduct a comparative analysis.

* **Application Area Chosen:**🡪

Speech Recognition

* **Identify Three Papers:** Select three research/journal papers from reputed sources like IEEE/Springer or ACM that focus on the application of CNN/RNN/Transformer networks in your chosen domain. Upload all three PDFs as individual files on Canvas.
* **Create a Comparison Table:** Compare the three papers and present your findings in a table with the following titles:
* Group Number, member names, and BITS ID, Domain, PAPER 1, PAPER 2, PAPER 3 (with subheadings: Title, Authors, Year, Network used, Depth of the network, Network application (e.g., feature engineering, classification, regression), Loss function, Evaluation/Performance metric, Dataset used, URL if public dataset) [\*Reference Comparison Table is given below]
* **Conclude:** End the comparison with a proper conclusion highlighting your observations. Justify the choice of one paper over the others for implementation in Part B.

**Submission:** Upload the table as one PDF.

|  |  |  |  |
| --- | --- | --- | --- |
| **Domain: Speech Recognition** | **PAPER 1** | **PAPER 2** | **PAPER 3** |
| **Title of the paper** | [Deep Speech 2: End-to-End Speech Recognition in](https://arxiv.org/abs/1512.02595)  [English and Mandarin](https://arxiv.org/abs/1512.02595) | [SPEECH RECOGNITION WITH DEEP RECURRENT NEURAL NETWORKS](https://arxiv.org/abs/1303.5778) | [EESEN: END-TO-END SPEECH RECOGNITION USING DEEP RNN MODELS AND](https://arxiv.org/abs/1507.08240)  [WFST-BASED DECODING](https://arxiv.org/abs/1507.08240) |
| **Authors** | Baidu Research – Silicon Valley AI Lab∗  Dario Amodei, Rishita Anubhai, Eric Battenberg, Carl Case, Jared Casper, Bryan Catanzaro, | Alex Graves, Abdel-rahman Mohamed and Geoffrey Hinton | Yajie Miao, Mohammad Gowayyed, Florian Metze |
| **Year of publication** | 2015 | 2013 | 2015 |
| **Network used** | "Deep Speech 2" (DS2) network. Deep Speech 2 is an end-to-end deep learning model designed for speech recognition. | recurrent neural networks (RNNs), especially LSTM | automatic speech recognition (ASR) using end-to-end deep recurrent neural network (RNN) models and weighted finite-state transducer (WFST)-based decoding. |
| **Depth of the network** | It's described as having multiple layers, including bidirectional recurrent layers, convolutional layers, and other components like Batch Normalization. | The article discusses various network architectures with different depths. Specifically, it investigates the performance of deep recurrent neural networks (RNNs) with different numbers of hidden layers as below: Single-layer RNN (CTC-1l-622h), Deep RNNs (CTC-3l-500h-tanh), Bidirectional RNN | Not provided |
| **How is the network helping the overall task? eg: feature engg or classification or regression or all** | Deep Speech 2 (DS2) network plays a crucial role in advancing speech recognition using End-to-End Approach, Handling Diverse Speech Variability, Model Capacity and Depth, Convolutional & Bidirectional Recurrent Layers, Batch Normalization & Language Model Integration. | Combination of deep architectures with the Long Short-Term Memory (LSTM) RNN model to leverage the benefits of both deep networks and the context-preserving capabilities of LSTMs. | automatic speech recognition (ASR) using end-to-end deep recurrent neural network (RNN) models and weighted finite-state transducer (WFST)-based decoding. |
| **Loss function used** | Connectionist Temporal Classification (CTC) loss function. The CTC loss function considers all possible alignments between the input sequence and the target sequence and calculates the likelihood of each alignment. The goal of training is to maximize the likelihood of the correct alignment, which corresponds to the correct transcription. | No specific loss function isn't provided. | No specific loss function isn't provided. However, in automatic speech recognition (ASR) systems, the most common loss function used during training is the Connectionist Temporal Classification (CTC) loss. |
| **Evaluation / Performance metric used** | Word Error Rate (WER) and Character Error Rate (CER). These metrics are commonly used in speech recognition tasks to quantify the accuracy of the system's transcriptions compared to the ground truth transcriptions.  These metrics provide a quantitative assessment of the system's accuracy in converting speech audio into text. In the DS2 article, the system's performance is evaluated using these metrics on various test sets with different speech conditions, such as clean speech, noisy speech, accented speech, and more. | The article mentions that the performance of the models is evaluated in terms of the phoneme recognition error rate. This is a common evaluation metric in speech recognition tasks, and it measures the accuracy of the models in correctly predicting the sequence of phonemes (linguistic units) that make up an utterance.  Results showed that the proposed deep LSTM models outperformed single-layer models, achieving state-of-the-art results in phoneme recognition on the TIMIT dataset. | Word Error Rate (WER) as the primary evaluation metric. The WER is a standard metric in ASR and represents the percentage of words in the recognized transcription that differ from the reference transcription. |
| **Name of Dataset used. If a public dataset, provide the URL.** | DS2 article mentions the use of following various datasets for training and evaluation of their speech recognition system:  Internal English Dataset,  VoxForge Dataset (clean speech recorded by speakers with various accents) ,  CHiME Challenge Dataset (recordings of utterances in various noisy environments) ,  LibriSpeech Corpus (constructed from audio books) ,  [LJ Speech Dataset](https://keithito.com/LJ-Speech-Dataset/) (consisting of 13,100 short audio clips of a single speaker reading passages from 7 non-fiction books. A transcription is provided for each clip. Clips vary in length from 1 to 10 seconds and have a total length of approximately 24 hours.) | [TIMIT corpus for phoneme recognition](https://catalog.ldc.upenn.edu/LDC93S1)  It’s a standard dataset used for evaluation of automatic speech recognition systems & consists of recordings of 630 speakers of 8 dialects of American English each reading 10 phonetically-rich sentences. | [Wall Street Journal (WSJ) corpus](https://paperswithcode.com/dataset/sms-wsj).  The corpus consists of various subsets of audio recordings, transcripts, and associated linguistic resources, and it is commonly used to benchmark ASR algorithms and models. |
| **Choice of Paper for Implementation** | **PAPER 1 (Deep Speech 2: End-to-End Speech Recognition in English and Mandarin)**:🡪  This paper explains following distinguished model architecture in great details (along with mathematical proof) making implementation clear & easy & very high accuracy due to its training over different datasets:  Preliminaries:🡪variants of DS2 architecture are explored by varying no. of convolutional layers from 1 to 3 & number of recurrent or GRU layers from 1 to 7.  Batch Normalization for Deep RNNs:🡪 It substantially improves final generalization error while greatly accelerating training for very deep networks of simple RNNs on large data sets  SortaGrad:🡪it uses length of utterance as a heuristic for difficulty, since long utterances have higher cost than short utterances.)  Frequency Convolutions:🡪Since spectral ordering of features is removed by fully-connected and recurrent layers, frequency convolutions work better as first layers of network.  Striding:🡪bigrams allow for larger strides without any sacrifice in word error rate. This allows to reduce number of time-steps of unrolled RNN benefiting both computation and memory usage.)  Row Convolution and Unidirectional Models:🡪Recurrent layers have learned good feature representations, so row convolution layer simply gathers appropriate information to feed to classifier.)  Language Model (RNN Models are trained over millions of unique utterances, which enables network to learn a powerful implicit language model. | | |