Comparing Diverse Methodologies for Dialogue Act Recognition on Conversational Speech

Souvik Maji[1], Arjun Bhattad[1]
[1]Indian Institute of Technology Jodhpur
{b22cs089, b22ai051}@iitj.ac.in

https://github.com/majisouvik26/Dialogue_Act_Recognition

Abstract

Dialogue Act Recognition (DAR) is crucial for understanding communicative intent in conversations. This paper presents a comparative study of four distinct methodologies for DAR, evaluated on the widely-used Switchboard Dialog Act Corpus. We investigate approaches spanning different architectural philosophies: (i) a context-aware self-attention model (CASA) focusing on explicit modeling of conversational turns; (ii) traditional machine learning models (XGBoost, MLP) using engineered acoustic and textual features; (iii) an efficient Fourier-Mixing Transformer (FNet) that replaces attention with spectral transforms; and (iv) a generative large language model (GPT-2) fine-tuned for the classification task. For each method, we explore the contribution of textual and acoustic features through ablation studies. Our experiments provide insights into the strengths and weaknesses of these diverse approaches, highlighting the effectiveness of context modeling and multimodal fusion, while also considering computational efficiency. The results offer guidance for selecting appropriate DAR models based on specific application requirements.

1. Introduction

Understanding conversations requires interpreting not just *what* is said, but also *why* it is said—the underlying communicative intent. Dialogue Act Recognition (DAR) addresses this by classifying utterances into semantic categories like 'Statement', 'Question', 'Backchannel', or 'Agreement'. Accurate DAR is fundamental for building robust conversational AI systems, including dialogue state tracking, summarization (as explored with datasets like DialogSum [5]), response generation, and human-robot interaction.

Early DAR systems often relied on acoustic prosodic features combined with statistical models like Hidden Markov Models [9, 20, 24]. With the advent of deep learning, recurrent neural networks (RNNs) and later Transformers [25] demonstrated superior performance by learning richer representations from text transcripts. However, different architectural choices embed distinct inductive biases, impacting performance, efficiency, and reliance on context or specific modalities.

The contribution of this paper is a systematic comparison of four representative DAR methodologies applied to the benchmark Switchboard Dialog Act Corpus [11, 24]. We aim to understand the trade-offs associated with different modeling philosophies:

1. **Explicit Context Modeling:** Re-implementing and enhancing Context-Aware Self-Attention (CASA) [**?**] to explicitly model dependencies between adjacent turns.
2. **Feature Engineering:** Using classical machine learning models (Gradient Boosting and MLP) with carefully extracted acoustic and TF-IDF text features, serving as strong baselines.
3. **Attention-Free Transformation:** Employing FNet [16], a Transformer variant replacing self-attention with efficient Fourier transforms for global mixing.
4. **Generative Pre-training:** Fine-tuning a pre-trained generative language model (GPT-2 Medium [21]) using task-specific prompts and exploring parameter-efficient techniques like LoRA [12].

For each approach, we evaluate the impact of using text-only, audio-only, and combined text-audio features via early concatenation. This comparative analysis provides valuable insights into choosing effective and efficient DAR models for conversational speech understanding.

2. Related Work

Dialogue Act recognition has a rich history. Early work focused on statistical methods using lexical and acoustic-prosodic features, often coupled with sequential models like

HMMs to capture dialogue flow [13, 24]. Hand-crafted features capturing turn-taking patterns, cue phrases, pitch contours, and energy levels were common [9, 23].

The rise of deep learning brought significant advances. Recurrent Neural Networks (LSTMs, GRUs [6]) became popular for modeling sequential dependencies within utterances [14, 15]. Attention mechanisms [25] further improved performance by allowing models to focus on relevant parts of the input, both within and across utterances. Raheja and Tetreault [**?**] specifically proposed Context-Aware Self-Attention (CASA) to explicitly model the influence of neighboring turns, demonstrating strong improvements. Our work builds upon their CASA model with architectural refinements.

More recently, large pre-trained language models (PLMs) like BERT [10] and GPT [21] have achieved state-of-the-art results on many NLP tasks, including classification. Fine-tuning PLMs for DAR has shown promise [**?**], leveraging their powerful text representations learned from vast corpora. Concurrently, alternative Transformer architectures like FNet [16] emerged, offering computational efficiency by replacing the quadratic self-attention mechanism with linear-time Fourier transforms, proving effective on various sequence tasks.

Traditional machine learning methods, such as Support Vector Machines [8] and Gradient Boosted Trees (like XGBoost [4]), remain relevant baselines, especially when combined with strong feature engineering [1]. They often require less data and compute resources than large neural models.

Our work situates itself within this landscape by directly comparing models representing these different paradigms—context-aware attention, classical ML with feature engineering, attention-free Transformers, and fine-tuned PLMs—on the same benchmark dataset and feature sets, specifically examining the role of context and multimodality (text and audio).

3. Problem Formulation

The goal of Dialogue Act Recognition is to assign a predefined semantic label y_i from a set of C possible dialogue acts (e.g., Statement, Question, Backchannel) to each utterance U_i within a conversation. An utterance U_i consists of a sequence of words $U_i = \langle w_{i,1}, \ldots, w_{i,T_i} \rangle$.

The input to a DAR model typically includes the text transcript of the utterance U_i. Additionally, acoustic features derived from the corresponding raw audio signal $s_i(t)$ can be incorporated. As detailed in Sec. 4.5, we extract an acoustic feature vector $\mathbf{x}_i^{\text{audio}} \in \mathbb{R}^{D_{\text{audio}}}$ summarizing prosodic and spectral characteristics.

$$\mathbf{x}_i^{\text{audio}} = \mathcal{F}_{\text{audio}}(s_i(t)), \tag{1}$$

where $\mathcal{F}_{\text{audio}}$ represents the acoustic feature extraction pipeline.

Textual information is processed differently depending on the model. It might involve creating TF-IDF vectors $\mathbf{x}_i^{\text{text}} \in \mathbb{R}^{D_{\text{text}}}$ (Method 2), generating contextualized embeddings using RNNs or Transformers (Methods 1, 3), or directly using the raw text sequence as input to a pre-trained language model (Method 4).

$$\mathbf{z}_i = \mathcal{F}_{\text{text}}(U_i, \mathcal{C}_i), \tag{2}$$

where $\mathcal{F}_{\text{text}}$ is the text processing function, which may also consider the conversational context $\mathcal{C}_i = \{U_{i-k}, \ldots, U_{i+k}\}$, a window of $2k+1$ turns around the target utterance U_i.

The DAR model itself is a function f that maps the available input features to a probability distribution over the C dialogue act classes:

$$p(y_i | U_i, \mathcal{C}_i, \mathbf{x}_i^{\text{audio}}; \theta) = f(U_i, \mathcal{C}_i, \mathbf{x}_i^{\text{audio}}; \theta), \tag{3}$$

where θ represents the model parameters learned during training. The predicted label $\hat{y}_i$ is typically the one with the highest probability:

$$\hat{y}_i = \underset{c \in \{1,\ldots,C\}}{\arg\max}\ p(y_i = c | U_i, \mathcal{C}_i, \mathbf{x}_i^{\text{audio}}; \theta). \tag{4}$$

Training involves minimizing a suitable loss function over a labeled dataset $\{(U_i, \mathcal{C}_i, \mathbf{x}_i^{\text{audio}}, y_i)\}_{i=1}^N$. For discriminative models (Methods 1, 2, 3), this is typically the cross-entropy loss (potentially smoothed):

$$\mathcal{L}_{\text{CE}} = -\frac{1}{N} \sum_{i=1}^{N} \sum_{c=1}^{C} \mathbb{I}\{y_i = c\} \log p(y_i = c | \cdot; \theta), \tag{5}$$

where $\mathbb{I}\{\cdot\}$ is the indicator function. For the generative approach (Method 4), the loss is typically the negative log-likelihood of generating the target DA label sequence y_i given the input prompt, as shown in Sec. 4.4.

4. Methodologies

We experiment with four philosophically distinct approaches to dialogue-act (DA) recognition on the Switchboard Dialog Act Corpus [11, 24], organised from most to least task-specific inductive bias. To make comparisons fair, every method is trained and evaluated on the same standard data splits, uses the same text and/or acoustic feature pipelines described in Sec. 4.5 where applicable, and predicts the canonical 42-way Switchboard tag set [13]. Throughout, let $U_i = \langle w_{i,1}, \ldots, w_{i,T_i} \rangle$ denote the i-th utterance in a conversation and let $\mathcal{C}_i = \{U_{i-k}, \ldots, U_{i+k}\}$ be its $2k{+}1$-turn symmetric context window (we use $k{=}2$ unless stated otherwise).

4.1. Context-Aware Self-Attention (Method 1)

Motivation. Raheja & Tetreault [**?**] showed that a speaker's DA depends strongly on the pragmatic function of neighbouring turns. Merely concatenating context tokens forces a model to learn context selection implicitly; their *Context-Aware Self-Attention* (CASA) makes this selection *explicit* and *differentiable*. We re-implement CASA with three architectural refinements: (i) a bidirectional GRU encoder [6] that supplies token-level features, (ii) multihead rather than single-head attention [25], and (iii) a learnable gate that interpolates utterance-internal and inter-utterance representations.

Token encoder. Each token $w_{i,t}$ is mapped to a d_e-dimensional embedding $\mathbf{e}_{i,t} = \mathrm{Embed}(w_{i,t})$ and passed through a BiGRU:

$$\overrightarrow{\mathbf{h}}_{i,t},\ \overleftarrow{\mathbf{h}}_{i,t} = \mathrm{GRU}\big(\mathbf{e}_{i,t}, \overrightarrow{\mathbf{h}}_{i,t-1}, \overleftarrow{\mathbf{h}}_{i,t+1}\big). \tag{6}$$

We concatenate directions to obtain $\mathbf{h}_{i,t} \in \mathbb{R}^{d_h}$ and summarise U_i by mean-pooling: $\mathbf{u}_i = \frac{1}{T_i}\sum_t \mathbf{h}_{i,t}$.

Context-aware multihead attention. For every current utterance U_i we form query, key, and value matrices

$$\mathbf{Q}_i = \mathbf{W}_Q\,\mathbf{u}_i, \quad \mathbf{K}_j = \mathbf{W}_K\,\mathbf{u}_j, \quad \mathbf{V}_j = \mathbf{W}_V\,\mathbf{u}_j, \tag{7}$$

for all context utterances $U_j \in \mathcal{C}_i$ (including U_i itself, $j = i$). With multi-head index $h \in \{1, \dots, H\}$, the attention weights are

$$\alpha_{ij}^{(h)} = \mathrm{softmax}_{j\in\mathcal{C}_i}\Big(\frac{\mathbf{Q}_i^{(h)}\,\mathbf{K}_j^{(h)\top}}{\sqrt{d_k}}\Big). \tag{8}$$

The context summary is therefore

$$\mathbf{c}_i = \underset{h=1}{\overset{H}{\mathrm{Concat}}}\Big(\sum_{j\in\mathcal{C}_i} \alpha_{ij}^{(h)}\mathbf{V}_j^{(h)}\Big). \tag{9}$$

Gated fusion. To balance evidence that is *internal* to U_i against evidence gleaned from $\mathcal{C}_i$, we compute a sigmoid gate

$$g_i = \sigma\big(\mathbf{w}_g^\top[\mathbf{u}_i; \mathbf{c}_i] + b_g\big), \tag{10}$$

where $[\cdot;\cdot]$ denotes concatenation, and fuse representations as

$$\mathbf{r}_i = g_i \odot \mathbf{u}_i + (1 - g_i) \odot \mathbf{c}_i. \tag{11}$$

Finally, DA logits are produced by $\mathbf{o}_i = \mathbf{W}_o\mathbf{r}_i + \mathbf{b}_o$ and optimised with a smoothed cross-entropy loss. Dropout, layer normalisation, and L_2 decay follow the original paper's hyper-parameters.

Why it matters. CASA explicitly models pragmatic dependencies (e.g., question→answer, statement→backchannel), and the gating term adapts to varying discourse styles. This explicit context modeling is expected to improve performance compared to context-agnostic or implicit context models.

4.2. Gradient-Boosted / MLP Baseline (Method 2)

Feature engineering. This approach relies on traditional, engineered features. Each utterance U_i is represented by a single concatenated vector $\mathbf{x}_i = [\mathbf{x}_i^{\text{text}}; \mathbf{x}_i^{\text{audio}}]$. Text features $\mathbf{x}_i^{\text{text}} \in \mathbb{R}^V$ are TF–IDF weights over a unigram + bigram vocabulary V derived from the training set. Acoustic features $\mathbf{x}_i^{\text{audio}} \in \mathbb{R}^{D_{\text{audio}}}$ are the classical descriptors detailed in Sec. 4.5 (MFCCs, pitch, formants, energy statistics), typically aggregated over the utterance (e.g., mean, std dev).

XGBoost. We use XGBoost [4], an efficient implementation of gradient boosted decision trees. Given a training set $\{(\mathbf{x}_i, y_i)\}$, XGBoost builds an ensemble of M decision trees $f_m \in \mathcal{F}$ additively: $\hat{y}_i = \sum_{m=1}^{M} f_m(\mathbf{x}_i)$. It minimizes the regularized objective:

$$\mathcal{L} = \sum_i \ell\big(y_i, \hat{y}_i\big) + \sum_{m=1}^{M} \Omega(f_m), \tag{12}$$

where ℓ is the multi-class logarithmic loss, and $\Omega(f) = \gamma T + \frac{1}{2}\lambda\|w\|^2$ penalizes tree complexity (number of leaves T) and leaf weights w. We use standard hyperparameters (e.g., 300 trees, max depth 6, learning rate 0.1).

MLP. For comparison within the feature-based paradigm, we train a simple Multi-Layer Perceptron (MLP) on the same input vector $\mathbf{x}_i$. It consists of two hidden layers with ReLU activation ρ and dropout:

$$\mathbf{h}_1 = \rho(\mathbf{W}_1\mathbf{x}_i + b_1),\ \mathbf{h}_2 = \rho(\mathbf{W}_2\mathbf{h}_1 + b_2),\ \mathbf{o}_i = \mathbf{W}_3\mathbf{h}_2 + b_3.$$

The MLP provides a neural network baseline using the identical engineered features as XGBoost.

4.3. Fourier-Mixing Transformer (Method 3)

Motivation. FNet [16] proposed replacing the computationally expensive self-attention mechanism in Transformers with unparameterized Fourier Transforms for token mixing. This significantly reduces computational complexity while maintaining strong performance on language tasks.

Spectral mixing layer. Each token embedding matrix $\mathbf{X}^{(l)} \in \mathbb{R}^{T_i \times d}$ at layer l (where T_i is utterance length and d is hidden dimension) is mixed globally using 2D Discrete Fourier Transform (DFT):

$$\tilde{\mathbf{X}}^{(l)} = \Re\big(\mathcal{F}_d(\mathcal{F}_{T_i}(\mathbf{X}^{(l)}))\big),$$

where $\mathcal{F}_{T_i}$ applies DFT along the sequence dimension and $\mathcal{F}_d$ along the feature dimension. The real part $\Re(\cdot)$ is retained. This is followed by the standard Transformer feed-forward network (FFN) block:

$$\mathbf{Y}^{(l)} = \text{LayerNorm}(\mathbf{X}^{(l)} + \tilde{\mathbf{X}}^{(l)}) \tag{13}$$

$$\mathbf{X}^{(l+1)} = \text{LayerNorm}(\mathbf{Y}^{(l)} + \text{FFN}(\mathbf{Y}^{(l)})) \tag{14}$$

Because the 2D FFT costs $O(T_i d \log(T_i d))$, this is much faster than self-attention's $O(T_i^2 d)$.

Classifier. We use a standard pre-trained FNet model as the text encoder. After L Fourier-mixing blocks, we obtain token representations. We apply mean pooling over the final layer's token representations to get an utterance summary $\mathbf{u}_i^{\text{FNet}}$. For multimodal fusion, we concatenate this with the acoustic summary vector $\mathbf{x}_i^{\text{audio}}$:

$$\mathbf{z}_i = [\mathbf{u}_i^{\text{FNet}}; \mathbf{x}_i^{\text{audio}}], \quad \mathbf{o}_i = \text{Linear}(\mathbf{z}_i).$$

The model is trained using cross-entropy loss. Despite lacking explicit attention, spectral mixing captures global token interactions; we expect combining this with acoustic features to improve robustness.

4.4. Generative LLM Fine-Tuning (Method 4)

Motivation. Large Language Models (LLMs) pre-trained on vast text corpora possess strong language understanding capabilities. Fine-tuning them for specific downstream tasks like classification has proven highly effective. We explore this approach by framing DAR as a conditional generation task.

Prompt design. Following approaches like Kawamura , each training instance (U_i, y_i) is formatted into a textual prompt. For example, the input might be structured as:

```
Prompt: "Classify the intent of the
following utterance:"
Utterance: U_i
Dialogue Act:
```

The model is then trained to generate the target label y_i (e.g., `Statement`, `Question`) following this prompt. Here, U_i is the text of the utterance. We fine-tune a pre-trained GPT-2 Medium model [21] using this format.

Training and Loss. The model is trained using the standard causal language modeling objective, minimizing the negative log-likelihood (NLL) of the target DA label tokens given the prompt and preceding label tokens (teacher forcing):

$$\mathcal{L}_{\text{NLL}} = -\sum_{t=1}^{|y_i|} \log p_\theta(y_{i,t} \mid \text{Prompt}, U_i, y_{i,<t}). \tag{15}$$

Inference. To classify a new utterance U_{new}, we construct the prompt ending with `Dialogue Act:` . The model then generates the subsequent tokens. We typically use greedy decoding and take the first complete generated label as the prediction. Logits beyond the required label tokens are ignored.

Parameter-efficient tuning. To manage computational resources, we also experiment with Low-Rank Adaptation (LoRA) [12]. This involves freezing the pre-trained LLM weights and injecting small, trainable rank-decomposition matrices into specific layers (e.g., attention layers). We use standard LoRA hyperparameters (r=8, α=16) and freeze adapters early if validation loss plateaus. This significantly reduces the number of trainable parameters and VRAM usage.

4.5. Feature Extraction and Modal Ablation

Acoustic Features ($\mathbf{x}_i^{\text{audio}}$). Standard acoustic features are extracted from the raw audio signal $s_i(t)$ using libraries like Librosa or OpenSMILE. We use a frame size of 25 ms with a 10 ms hop, applying a Hamming window. Key features include:

- Mel-Frequency Cepstral Coefficients (MFCCs): Typically the first 13 coefficients, plus their delta and delta-delta, summarizing spectral shape. Calculated as:

$$c_n(m) = \sum_b \log S_b(m) \cos\Big[\frac{\pi}{B} n(b - 0.5)\Big] \tag{16}$$

  where $S_b(m)$ is Mel-filterbank energy.
- Pitch (f_0): Fundamental frequency, estimated using methods like autocorrelation (ρ_x):

$$f_0 = f_s / \arg\max_\tau \rho_x(\tau) \tag{17}$$

  where f_s is the sampling frequency.
- Energy/Loudness: Root Mean Square (RMS) energy or perceived loudness.
- Formants: Resonant frequencies of the vocal tract, estimated via Linear Predictive Coding (LPC) root analysis.
- Zero-Crossing Rate (ZCR): Rate at which the signal changes sign.

For Methods 2 and 3, these frame-level features are aggregated into utterance-level statistics (e.g., mean, standard deviation, min, max) to form the fixed-size vector $\mathbf{x}_i^{\text{audio}}$. Method 1 (CASA) and Method 4 (GPT-2) in our setup primarily use text, but acoustic features could potentially be integrated via fusion layers (not explored in the current CASA/GPT-2 implementations here).

Textual Features. Text processing varies by method:

- Method 1 (CASA): Word embeddings + BiGRU encoder.
- Method 2 (XGB/MLP): TF-IDF vectors (unigrams+bigrams).
- Method 3 (FNet): Subword token embeddings processed by FNet layers.
- Method 4 (GPT-2): Subword token embeddings processed by GPT-2 layers, using the raw prompt format.

Ablation Settings. To understand the contribution of each modality, we evaluate three variants for applicable methods (Methods 1, 2, 3; Method 4 is text-only by design in this setup): [label=()]

Text-only: Using only features derived from the transcript ($\mathcal{F}_{\text{text}}$).

Audio-only: Using only aggregated acoustic features ($\mathbf{x}_i^{\text{audio}}$). This applies mainly to Method 2. For Methods 1/3, this would require a separate audio encoder branch.

Early-concat: Concatenating text and acoustic feature vectors before the final classification layer(s). For Method 1/3, this means $[\mathbf{r}_i; \mathbf{x}_i^{\text{audio}}]$ or $[\mathbf{u}_i^{\text{FNet}}; \mathbf{x}_i^{\text{audio}}]$. For Method 2, this is the default setup $[\mathbf{x}_i^{\text{text}}; \mathbf{x}_i^{\text{audio}}]$. This ablation allows us to quantify the performance gain from multimodal fusion.

5. Experiments and Results

5.1. Dataset and Setup

We conducted experiments on the Switchboard Dialog Act Corpus (SwDA) [11, 24], a widely used benchmark for DAR. It contains telephone conversations annotated with the 42 dialogue act tags defined by the DAMSL standard [13]. We used the standard train/validation/test splits commonly employed in prior work [**?**, 24]. While DialogSum [5] provides dialogue data, SwDA's detailed utterance-level DA annotations make it more suitable for this comparative study.

All models were implemented using standard libraries: PyTorch [19] for neural models (CASA, FNet, MLP, GPT-2), Hugging Face Transformers [26] for FNet and GPT-2 backbone/fine-tuning, XGBoost library [4] for the gradient boosting model, and scikit-learn [22] for TF-IDF and MLP baseline. Acoustic features were extracted using Librosa [18]. Hyperparameters for each model were tuned based on performance on the validation set (e.g., learning rates, dropout rates, layer sizes, number of trees, LoRA parameters $r = 8, \alpha = 16$). For context-aware models (CASA), we used a context window of $k = 2$ (previous 2 turns, current turn, next 2 turns).

5.2. Evaluation Metrics

We report standard classification metrics:

- **Accuracy**: Overall percentage of correctly classified utterances.
- **Macro-F1 Score**: The unweighted average of the F1-score for each of the 42 DA classes. This metric is crucial for SwDA due to its significant class imbalance, giving equal importance to frequent and infrequent dialogue acts.

5.3. Main Results

Table 1 presents the primary results comparing the four methodologies and their modal variants on the SwDA test set.

5.4. Generative Model Results (Method 4)

Method 4 fine-tuned Flan-T5-base for text generation of DA labels. Table 2 shows the ROUGE F1-scores on the test set, comparing the base model (Original), full fine-tuning (Instruct), and parameter-efficient fine-tuning using LoRA (PEFT).

Comparison of Approaches. The results show that the neural approaches leveraging contextual information (CASA, FNet, GPT-2) significantly outperform the traditional feature-based methods (XGBoost, MLP). Among the neural models, fine-tuned GPT-2 Medium achieves the highest text-only performance in both accuracy and Macro-F1, demonstrating the power of large pre-trained models. CASA, with its explicit context modeling, also performs very strongly, slightly behind the fully fine-tuned GPT-2 but benefiting from multimodal fusion. FNet offers a competitive alternative, achieving results close to CASA but likely with better training/inference speed due to the absence of quadratic attention (though not benchmarked here). The LoRA variant of GPT-2 provides performance close to full fine-tuning with significantly fewer trainable parameters.

Impact of Modality. As observed in Table 1 and aligning with the statement in Sec. 4.5, incorporating acoustic features via early concatenation consistently improves performance for methods designed to handle them (CASA, XGBoost, MLP, FNet). The gains are noticeable in both accuracy and Macro-F1 score (around 1-2 points Macro-F1 improvement). This confirms that acoustic cues related to prosody (pitch, energy) and voice quality provide complementary information to the lexical content for DAR. While the text modality dominates overall performance, audio adds valuable robustness. An audio-only baseline using Method 2 features yielded significantly lower results (Acc. 55

Role of Context. The strong performance of CASA, which explicitly models inter-utterance dependencies using

Table 1. Main Dialogue Act Recognition results on the Switchboard test set. Accuracy (%) and Macro-F1 Score (%) are reported. Best results in bold.

Method	Text-Only		Early-Concat (Text+Audio)	
	Acc.	Macro-F1	Acc.	Macro-F1
Method 1: CASA [?]				
(BiGRU + Attn + Gate)	81.5	54.2	82.1	55.5
Method 2: Feature-Based				
XGBoost [4]	62.2	38.8	62.8	40.1
MLP	61.8	26.1	62.2	29.5
Method 3: FNet [16]				
(Fourier Mixing)	77.6	51.9	78.3	53.0

Table 2. ROUGE F1-scores for Flan-T5-base generating DA labels on the DialogueSum test set. Higher is better.

Model Variant	ROUGE-1	ROUGE-2	ROUGE-L
Original (Base)	0.233	0.076	0.201
Instruct FT (Full)	**0.422**	**0.180**	**0.338**
PEFT (LoRA)	0.408	0.163	0.325

attention over a context window, underscores the importance of conversational context for accurate DAR. While FNet and GPT-2 implicitly capture some context through their pre-training and architecture, CASA's explicit mechanism proves highly effective. The feature-based methods (Method 2) inherently lack this sequential context modeling beyond N-grams, contributing to their lower performance.

6. Future Research Directions

This comparative study opens several avenues for future work:

- **Sophisticated Multimodal Fusion:** Explore more advanced fusion techniques beyond early concatenation, such as cross-modal attention mechanisms or dedicated fusion layers, potentially allowing models like GPT-2 to effectively leverage acoustic features.
- **Longer-Range Context:** Investigate methods to incorporate context beyond the fixed $k = 2$ window used here, possibly using hierarchical models or memory networks to capture discourse structure over longer conversational spans.
- **Self-Supervised Pre-training:** Leverage large unlabeled speech and text corpora for self-supervised pre-training of joint audio-text encoders specifically tailored for dialogue tasks, potentially improving robustness and data efficiency.
- **Efficiency vs. Performance Trade-offs:** Conduct a more detailed analysis of the computational costs (training time, inference latency, parameter count) associated with each method (e.g., FNet vs. CASA, full LLM FT vs. LoRA) relative to their performance.
- **Cross-Dataset Generalization:** Evaluate the best-performing models and fusion strategies on other dialogue datasets (e.g., AMI Corpus [3], DialogSum [5]) to assess the generalizability of the findings.
- **Error Analysis:** Perform a detailed error analysis to understand which dialogue acts are challenging for different architectures and modalities, guiding further model improvements.

7. Conclusion

We presented a systematic comparison of four diverse methodologies for Dialogue Act Recognition on the Switchboard corpus: context-aware attention (CASA), traditional feature-based ML (XGBoost/MLP), attention-free Fourier mixing (FNet), and generative LLM fine-tuning (GPT-2). Our experiments highlight the effectiveness of modern neural architectures, particularly those leveraging contextual information (CASA) and large pre-trained models (GPT-2), which significantly outperformed feature-based baselines. We confirmed the value of incorporating acoustic features alongside text via early fusion, leading to consistent performance gains across applicable methods. FNet presented an efficient and competitive alternative to attention-based models, while LoRA offered a parameter-efficient way to adapt LLMs. This comparative study provides insights into the strengths and trade-offs of different DAR approaches, informing the selection of models for building more capable conversational AI systems. Future work will focus on advanced fusion, longer context modeling, and cross-dataset generalization.

References

[1] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001. 2

[2] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei, "Language models are few-shot learners," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 33, 2020, pp. 1877–1901.

[3] J. Carletta, S. Ashby, S. Bourban, M. Flynn, M. Guillemot, T. Hain, J. Kadlec, V. Karaiskos, W. Kraaij, M. Kronenthal, G. Lathoud, M. Lincoln, A. Lisowska, I. McCowan, W. Post, D. Reidsma, and P. Wellner, "The AMI meeting corpus: A pre-announcement," in *Proc. International Workshop on Machine Learning for Multimodal Interaction (MLMI)*, 2005, pp. 28–39. 6

[4] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016, pp. 785–794. 2, 3, 5, 6

[5] Y. Chen, K. Liu, J. Chen, L. Chen, and R. Yan, "DialogSum: A real-life scenario dialogue summarization dataset," in *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, 2021, pp. 1940–1951. 1, 5, 6

[6] K. Cho, B. van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using RNN encoder-decoder for statistical machine translation," in *Proc. Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2014, pp. 1724–1734. 2, 3

[7] H. W. Chung, L. Hou, S. Longpre, B. Zoph, Y. Tay, W. Fedus, Y. Li, X. Wang, M. Dehghani, S. Brahma, A. Webson, S. S. Gu, Z. Dai, M. Suzgun, X. Chen, A. Chowdhery, A. Narang, S. Mishra, V. Yu, V. Zhao, Y. Huang, A. Dai, H. Yu, S. Petrov, E. H. Chi, J. Dean, X. Zhai, A. Ghorbani, M. P. Marcus, L. Yu, J. Wei, and X. Liang, "Scaling instruction-finetuned language models," *arXiv preprint arXiv:2210.11416*, 2022.

[8] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, no. 3, pp. 273–297, 1995. 2

[9] S. B. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 28, no. 4, pp. 357–366, 1980. 1, 2

[10] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, 2019, pp. 4171–4186. 2

[11] J. J. Godfrey, E. C. Holliman, and J. McDaniel, "SWITCHBOARD: Telephone speech corpus for research and development," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 1992, pp. 517–520. 1, 2, 5

[12] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "LoRA: Low-rank adaptation of large language models," in *International Conference on Learning Representations (ICLR)*, 2022. 1, 4

[13] D. Jurafsky, E. Shriberg, and D. Biasca, "Switchboard SWBD-DAMSL shallow-discourse-function annotation coders manual," *University of Colorado, Boulder, Institute of Cognitive Science Technical Report*, 97-02, 1997. 2, 5

[14] N. Kalchbrenner and P. Blunsom, "Recurrent continuous translation models," in *Proc. Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2013, pp. 1700–1709. 2

[15] J. Lee and F. Dernoncourt, "Sequential short-text classification with recurrent and convolutional neural networks," in *Proc. Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, 2016, pp. 515–520. 2

[16] J. Lee-Thorp, J. Ainslie, I. Eckstein, and S. Ontanon, "FNet: Mixing tokens with Fourier transforms," *arXiv preprint arXiv:2105.03824*, 2021. 1, 2, 3, 6

[17] C.-Y. Lin, "ROUGE: A package for automatic evaluation of summaries," in *Proc. ACL Workshop on Text Summarization Branches Out*, 2004, pp. 74–81.

[18] B. McFee, C. Raffel, D. Liang, D. P. Ellis, M. McVicar, E. Battenberg, and O. Nieto, "librosa: Audio and music signal analysis in Python," in *Proc. 14th Python in Science Conference (SciPy)*, 2015, pp. 18–24. 5

[19] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilam, S. Chintala, T. K. Chawla, M. Schildhaut, D. Li, L. L. Stern, G. Bradski, and J. Bertozzi, "PyTorch: A deep learning framework for computer vision," in *Advances in Neural Information Processing Systems (NIPS)*, 2019, pp. 6397–6407. 5

[20] L. Rabiner and B.-H. Juang, *Fundamentals of Speech Recognition*, Prentice-Hall, 1993. 1

[21] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, "Language models are unsupervised multitask learners," *OpenAI Blog*, vol. 1, no. 8, 2019. 1, 2, 4

[22] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011. 5

[23] E. Shriberg, R. Bates, A. Stolcke, P. Taylor, D. Jurafsky, K. Ries, N. Coccaro, R. Martin, M. Meteer, and C. Van Ess-

Dykema, “Can prosody aid the automatic classification of dialog acts in conversational speech?” *Language and Speech*, vol. 41, no. 3-4, pp. 443–492, 1998. 2

[24] A. Stolcke, K. Ries, N. Coccaro, E. Shriberg, R. Bates, D. Jurafsky, P. Taylor, R. Martin, C. Van Ess-Dykema, and M. Meteer, “Dialogue act modeling for automatic tagging and recognition of conversational speech,” *Computational Linguistics*, vol. 26, no. 3, pp. 339–373, 2000. 1, 2, 5

[25] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, “Attention is all you need,” in *Advances in Neural Information Processing Systems (NIPS)*, 2017, pp. 5998–6008. 1, 2, 3

[26] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Lou, M. Funtowicz, J. Davison, and G. Brew, “Transformers: State-of-the-art natural language processing,” in *Proc. Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020, pp. 38–45. 5