Performance Ceiling Analysis: Why 0.18-0.19 Spearman Correlation?

ICF Estimation Project · 2025-12-06

This document analyzes the observed performance ceiling of 0.18-0.19 Spearman correlation, attributing it to information-theoretic bounds, Kolmogorov complexity, architectural limitations, and loss-metric mismatch.

0.1. Notation

- **ρ**: Spearman rank correlation coefficient
- **H(ICF)**: Shannon entropy of ICF distribution
- **I(Characters; ICF)**: Mutual information between character patterns and ICF
- **K(x)**: Kolmogorov complexity of x (length of shortest program producing x)
- **K(M)**: Kolmogorov complexity of model M
- **K(D)**: Kolmogorov complexity of dictionary D (word → ICF mapping)
- **σ^2**: Variance
- **E[·]**: Expected value

0.2. Observed Performance

Observed Results:

- **Best result:** loss_ablation_balanced_hybrid (0.1891)
- **Iter4 distillation:** 0.1875
- **Residual balanced:** 0.1864
- **Consistent ceiling:** 0.18-0.19 across multiple experiments

This suggests a **fundamental limit**, not just an optimization issue.

0.3. Theoretical Limits

0.3.1. Information-Theoretic Bound

Maximum Spearman correlation is bounded by:

$$\rho_{\max} \leq \sqrt{\frac{I(X;Y)}{H(Y)}}$$

Where:

- **X** = Character features
- **Y** = ICF values
- **I(X; Y)** = Mutual information between characters and ICF
- **H(Y)** = Shannon entropy of ICF distribution

0.3.2. Formal Derivation

$$\rho^2_{\max} \leq \frac{I(X;Y)}{H(Y)} \leq \frac{H(X) - H(X|Y)}{H(Y)}$$

This follows from the data processing inequality and the relationship between correlation and mutual information.

0.3.3. Key Insight

If `I(Characters; ICF)` is low relative to `H(ICF)`, the maximum achievable correlation is inherently limited.

Hypothesis: Character-level features capture approximately 18-19% of ICF variance because:

- Character patterns → semantic frequency is an **indirect mapping**
- Many words with similar character patterns have different ICF values
- ICF depends on corpus/domain characteristics (not just characters)

0.3.4. Kolmogorov Complexity Bound

K(ICF | Characters) measures the minimum information needed beyond character patterns to predict ICF.

If `K(ICF | Characters)` is large:
- Character patterns are insufficient
- Additional information (semantics, context) is required
- Performance ceiling is reached

0.4. Architectural Limitations

0.4.1. Current Design Captures

- ✅ Character-level morphological patterns
- ✅ N-gram features (3, 5, 7 character windows)
- ✅ Word-level patterns (via pooling)

0.4.2. Missing Information

- ❌ Semantic understanding (word meaning)
- ❌ Document context (domain, type, style)
- ❌ Co-occurrence patterns (word relationships)
- ❌ Temporal/domain trends
- ❌ Corpus-specific characteristics

0.4.3. Why This Matters

ICF fundamentally depends on:
1. **Word meaning** (semantic frequency)
2. **Document context** (domain-specific usage)
3. **Corpus characteristics** (training data distribution)

Character patterns alone cannot capture this information.

0.5. Loss-Metric Mismatch

0.5.1. Problem

- **Training objective:** MSE/Huber loss (minimizes absolute error)
- **Evaluation metric:** Spearman correlation (measures ranking quality)

Problem:
- Model optimizes for absolute accuracy
- But we care about relative ordering (ranking)
- Mismatch causes suboptimal optimization

Solution: Direct Spearman optimization (already implemented via `rank-relax`)

0.6. Data Quality & Noise

- ICF computed from specific corpus (bias)
- Measurement noise in frequency counts
- Domain mismatch between train/test
- Limited training data

0.7. Why 0.18-0.19 Specifically?

0.7.1. Hypothesis

Character features capture 18-19% of ICF variance because:

1. **Information content:** `I(Characters; ICF) / H(ICF) ≈ 0.18-0.19`
   - Characters provide limited information about frequency
   - Semantic/contextual information is missing
2. **Mapping complexity:** Character patterns → ICF is:
   - **Many-to-one:** Different words → same ICF
   - **One-to-many:** Similar patterns → different ICF
   - **Ambiguous:** Requires additional information
3. **Architectural limits:** Current CNN design:
   - No semantic understanding
   - No document context
   - Limited receptive field (word length only)

0.8. Breaking the Ceiling

0.8.1. Option 1: Add Semantic Features

- Word embeddings (capture meaning)
- Pre-trained language model features
- Semantic similarity to known words

0.8.2. Option 2: Add Context

- Document type/domain
- Co-occurrence patterns
- Temporal trends

0.8.3. Option 3: Larger Architecture

- Attention mechanisms (long-range dependencies)
- Transformer-based (semantic understanding)
- Multi-scale features

0.8.4. Option 4: Direct Spearman Optimization

- Train directly on Spearman loss (already implemented)
- Better alignment with evaluation metric

0.8.5. Option 5: Multi-Task Learning

- Predict multiple related tasks
- Share semantic representations
- Improve ICF prediction

0.9. Mathematical Formulation

0.9.1. Current

$$f : C^n \to [0, 1]$$

Where `C` = character vocabulary, `n` = word length. Limited because `I(C^n; ICF)` is bounded.

0.9.2. Better

$$f : (C^n, S, D) \to [0, 1]$$

Where `S` = semantic features, `D` = document context. This increases `I(Features; ICF)`.

0.10. Conclusion

The 0.18-0.19 ceiling is likely due to:

1. **Information-theoretic limit:** Characters provide limited ICF information
2. **Architectural limitations:** Missing semantic/contextual features
3. **Task difficulty:** Character patterns → frequency is indirect
4. **Loss-metric mismatch:** Suboptimal optimization (partially addressed)

To break the ceiling, we need to:

- Add semantic features (word embeddings, LM features)
- Add document context
- Use larger architectures with attention
- Continue direct Spearman optimization
- Explore multi-task learning