

PROJECT.1

석사 연구 1

REGEN: RECURRENT ENSEMBLE

METHODS FOR GENERATIVE MODELS

01

ABOUT PROJECT

저의 주 연구 분야인 생성 모델에서 앙상블에 대한 논문입니다. 석사 졸업 논문이고 현재 2023 IEEE Access에 투고 준비 중입니다.

Abstract

Title: REGEN: Recurrent Ensemble Methods for Generative Models

Author: Ahjeong Park, Youngmi Park, Chulyun Kim

- 진행 기간: 2021년 9월 ~ 2022년 10월
- 본인이 공헌한 점: 전체 연구 및 논문 진행

▶ [결과/성과]

1. 특허 출원

- 제목: 재귀 신경망 모델의 앙상블 방법 및 시스템
- 등록 심사중
- 출원일자: 2022년 12월 12일
- 발명자: 김철연, **박아정**, 박영미

2. IEEE Access 투고 준비 중

3. 자연어 번역 및 생성 모델 동작원리 이해, Seq2Seq, Transformer 구조에 대한 이해

4. 소스코드

- <https://github.com/aaajeong/Ensemble-for-Generative-Models.git>

REGEN: Recurrent Ensemble Methods for Generative Models




Ahjeong Park, Youngmi Park, Chulyun Kim

앙상블은 여러 모델을 활용하여 단일 구성 모델보다 더 나은 예측 성능을 얻습니다. 대부분의 기존 앙상블은 모델이 블랙박스로 간주되어 최종 결과만을 취합합니다. 이러한 고려는 다양한 종류의 기계학습 모델에 앙상블 모델을 적용할 수 있게 합니다. 특히 최종 출력의 Diversity가 제한적인 Discriminative 모델은 모델 간 의견 수렴이 쉽기 때문에 기존의 앙상블을 적용하기에 적합했습니다. 하지만 Generative 계열의 모델은 최종 출력의 길이와 범위에 제한이 없고 Diversity가 높아 모델 간 합의를 보는데 문제가 있습니다. 따라서 이 문제를 고려하기 위해 Generative 계열의 모델에 대한 **새로운 앙상블인 REGEN을 제안**했습니다.

새로운 앙상블은 **Consensus, Survival Ensemble** 입니다. 또한 기존의 앙상블을 새롭게 재해석한 Majority Ensemble도 설계하여 비교 실험을 진행했습니다.

앙상블 구성 모델로 Seq2Seq, Transformer을 활용했고 각 모델의 Decoder는 매 단계마다 합의를 진행한 후 다음 생성에 영향을 미치며 앙상블을 진행합니다. 기계번역 및 문자열 사칙연산에 대해 실험 결과, **REGEN은 단일 구성 모델 뿐만 아니라 기존 앙상블 보다 성능이 우수함을 확인**했습니다.

기여도

실험 설계		95
실험 진행		100
논문 작업		95

현재 상태

2023년 3월 IEEE Access 투고 준비 중입니다.

총 3가지 앙상블 방법 제안

- 1) Baseline: Majority
- 2) Recurrent Ensemble with Survival
- 3) Recurrent Ensemble with Consensus

실험 방법

각 앙상블에 대해 2개의 Case Study를 진행했습니다.

Case Study: Application of Recurrent Ensemble

- 1) Seq2Seq
- 2) Transformer

REGEN: Recurrent Ensemble Methods for Generative Models

Ahjeong Park, Youngmi Park, Chulyun Kim

1-1) Majority Ensemble in Seq2Seq

- Generative(RNN) 구조에 전통적인 앙상블 방법을 재해석한 Baseline 앙상블 방법입니다.
- 전통적인 앙상블과 동일하게 time-step의 중간 출력을 고려하지 않고 각 모델의 최종 output을 결합합니다.

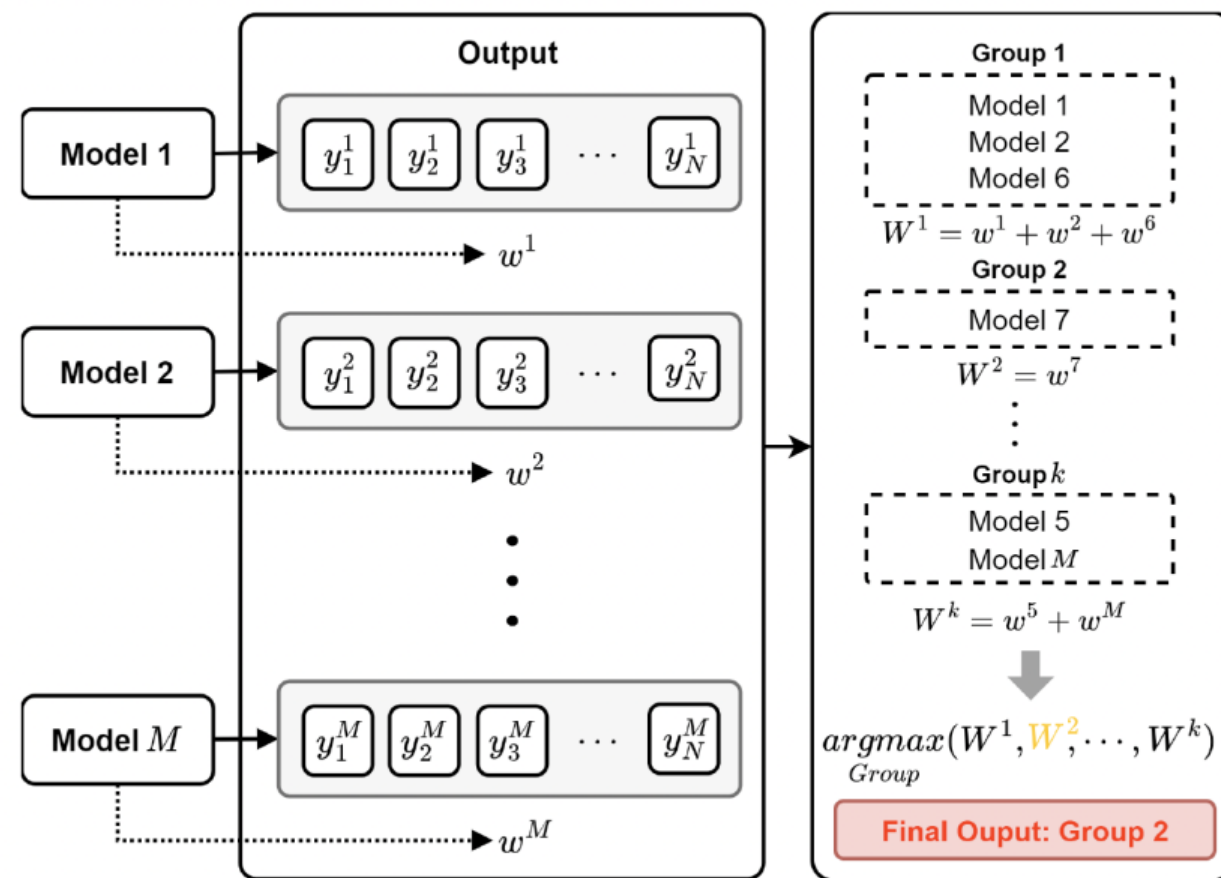


Figure 3.1: Overall Structure of Traditional Majority in Seq2Seq

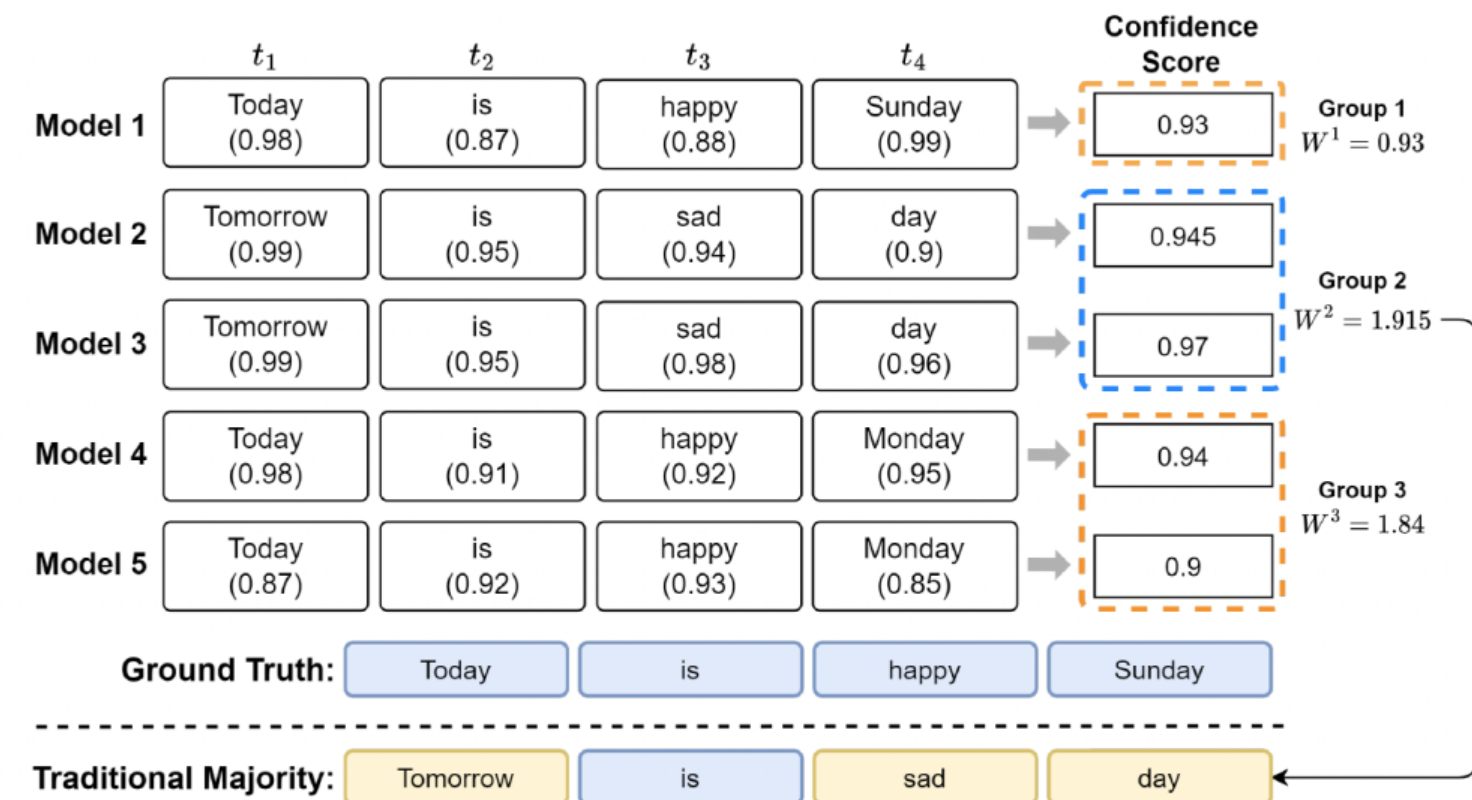


Figure 3.2: A toy example: Traditional Majority

REGEN: Recurrent Ensemble Methods for Generative Models

Ahjeong Park, Youngmi Park, Chulyun Kim

1-2) Survival Ensemble in Seq2Seq

- Recurrent Ensemble의 한 종류로, 게임 방식과 비슷해 Survival로 명칭했습니다.
- 살아남은 Winner(승자) 모델만이 다음 time-step 예측에 참여할 수 있도록 해서 최종 결과를 결정합니다.

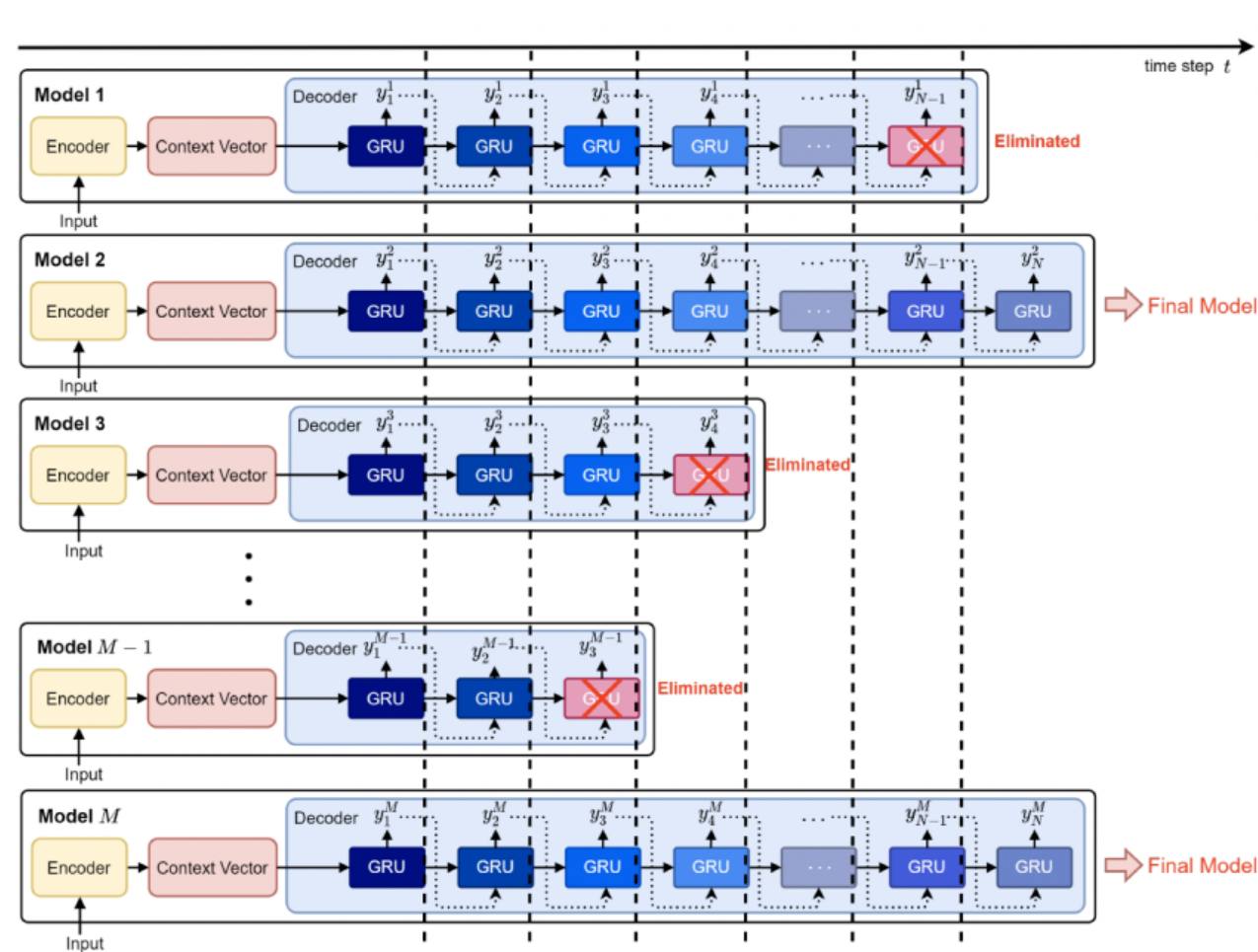


Figure 3.3: Overall Structure of Survival in Seq2Seq

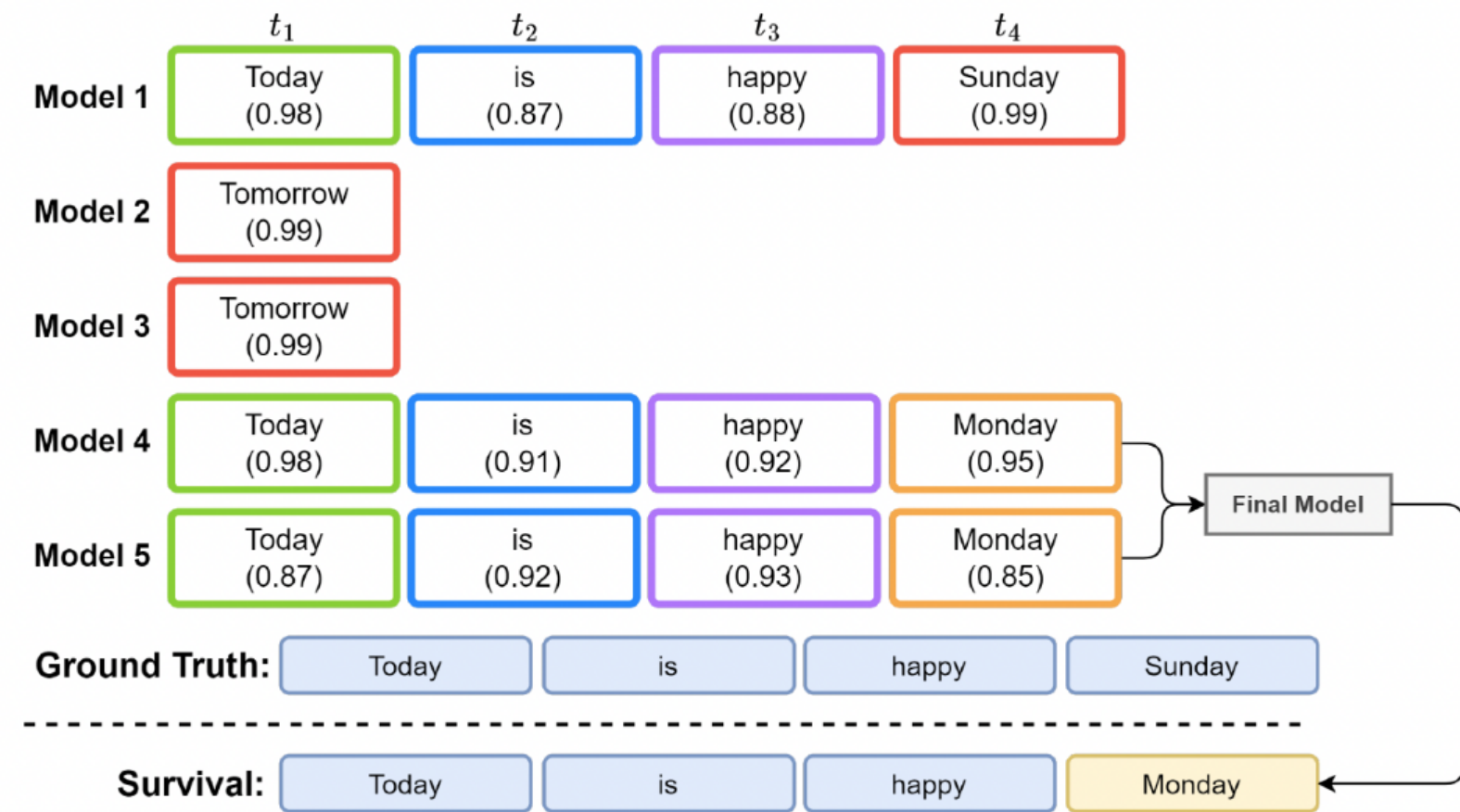


Figure 3.4: A toy example: Recurrent Ensemble using Survival

REGEN: Recurrent Ensemble Methods for Generative Models

Ahjeong Park, Youngmi Park, Chulyun Kim

1-3) Consensus Ensemble in Seq2Seq

- Recurrent Ensemble의 한 종류로, 가장 높은 성능을 달성한 앙상블 방법입니다.
- 모든 모델의 매 Time-step의 중간 출력을 고려합니다.
- 단일 모델의 Confidence를 고려한 Voting 결과를 넘겨주기 위해 Hard Voting이 아닌 Soft Voting 방식을 선택했습니다.

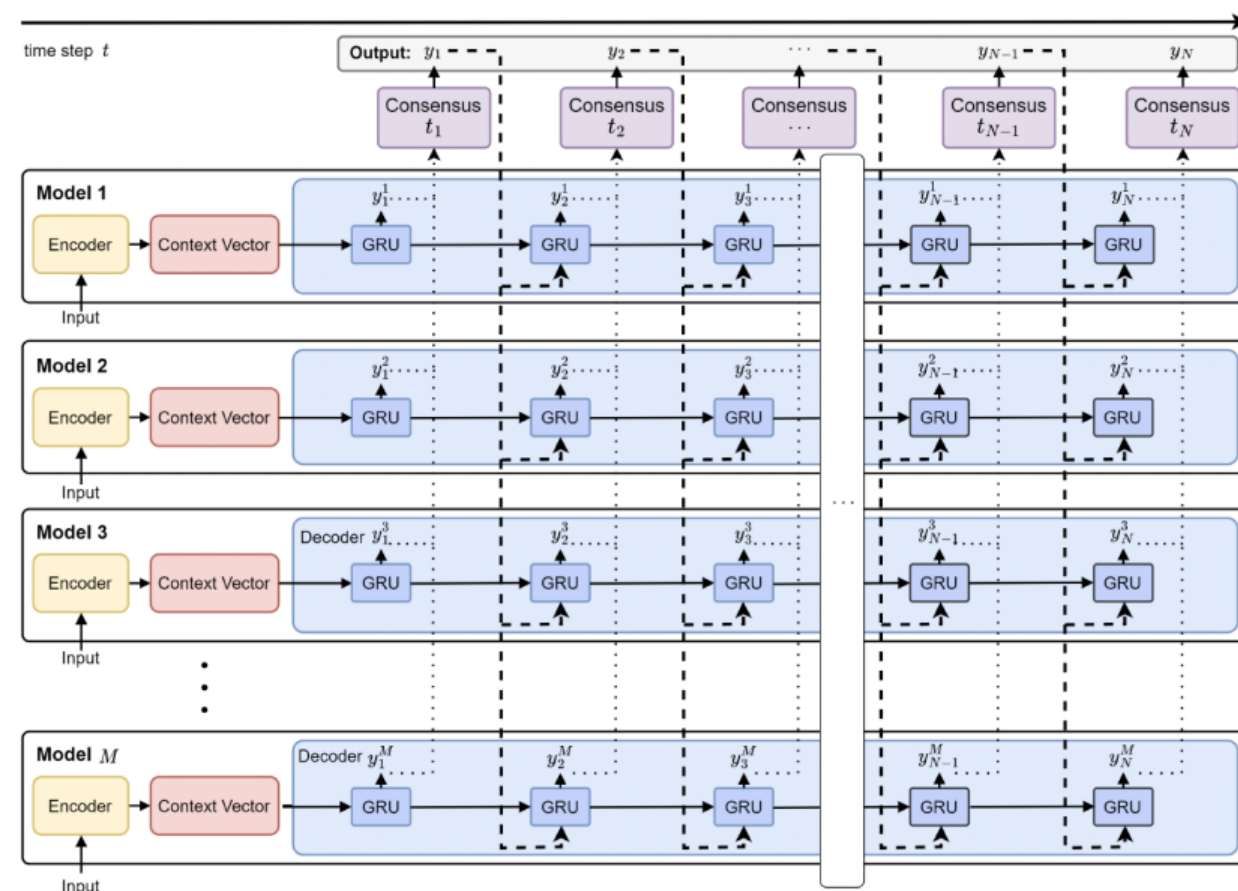


Figure 3.5: Overall Structure of Consensus in Seq2Seq

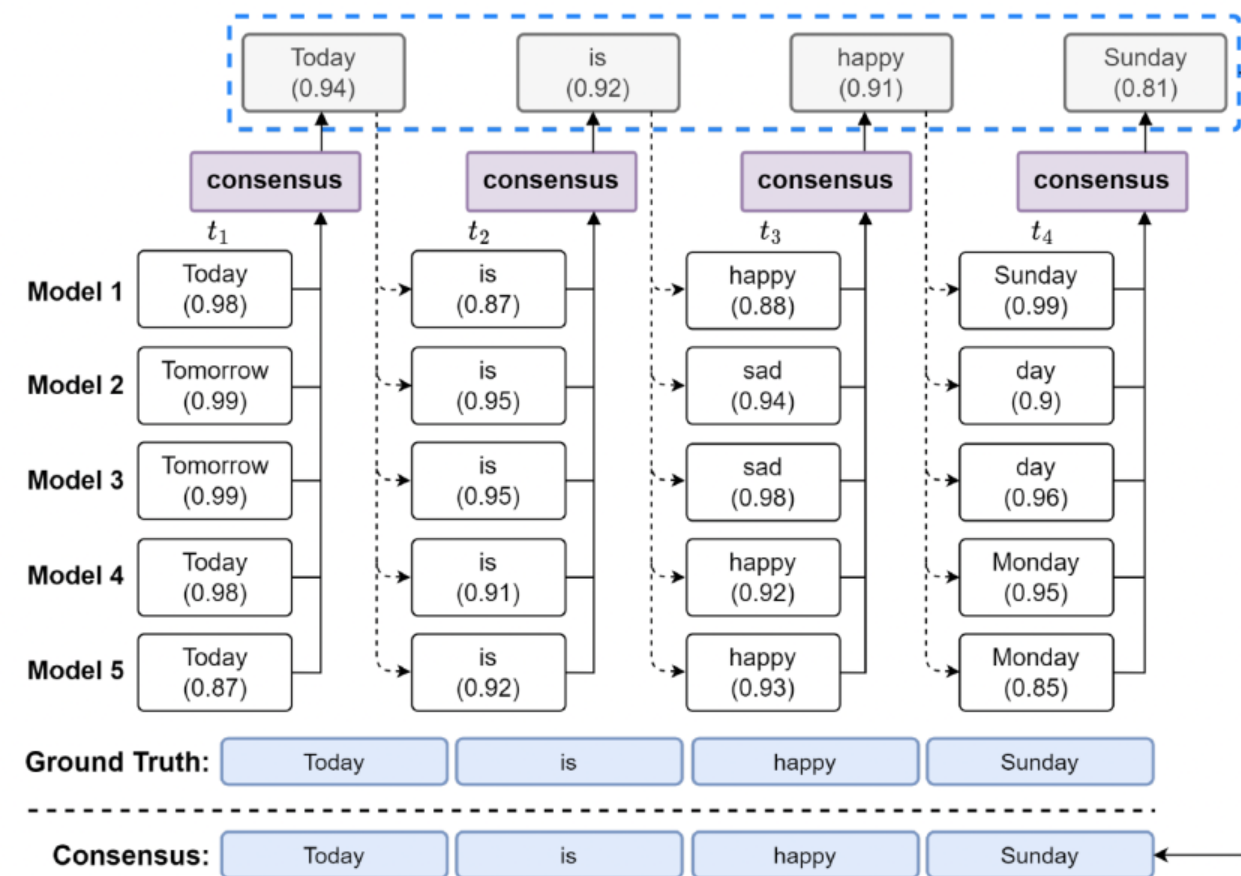


Figure 3.6: A toy example: Recurrent Ensemble using Consensus

REGEN: Recurrent Ensemble Methods for Generative Models

Ahjeong Park, Youngmi Park, Chulyun Kim

2) Majority, Survival and Consensus Ensemble in Transformer

- Transformer 모델에서도 앙상블 알고리즘은 동일하게 적용됩니다.
- 한가지 다른 점은 Transformer는 Seq2Seq과 달리 이전의 output이 하나씩 되먹임 되는 방식이 아니라 마스킹 되는 부분을 고려해주었습니다.

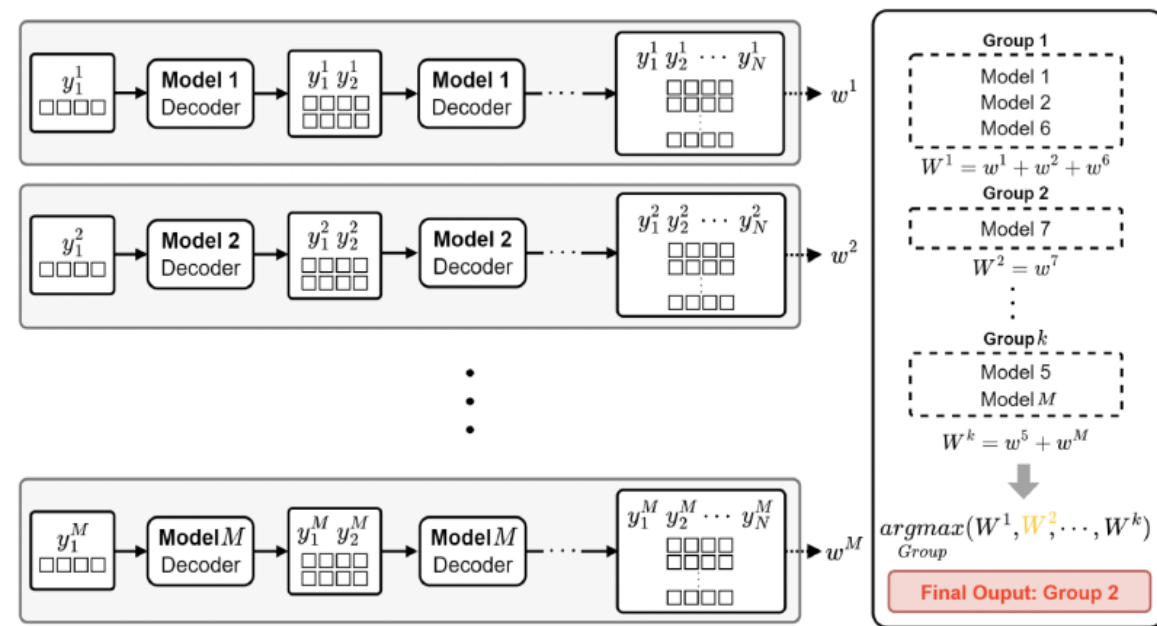


Figure 3.7: Overall Structure of Traditional Majority in Transformer

Majority Ensemble

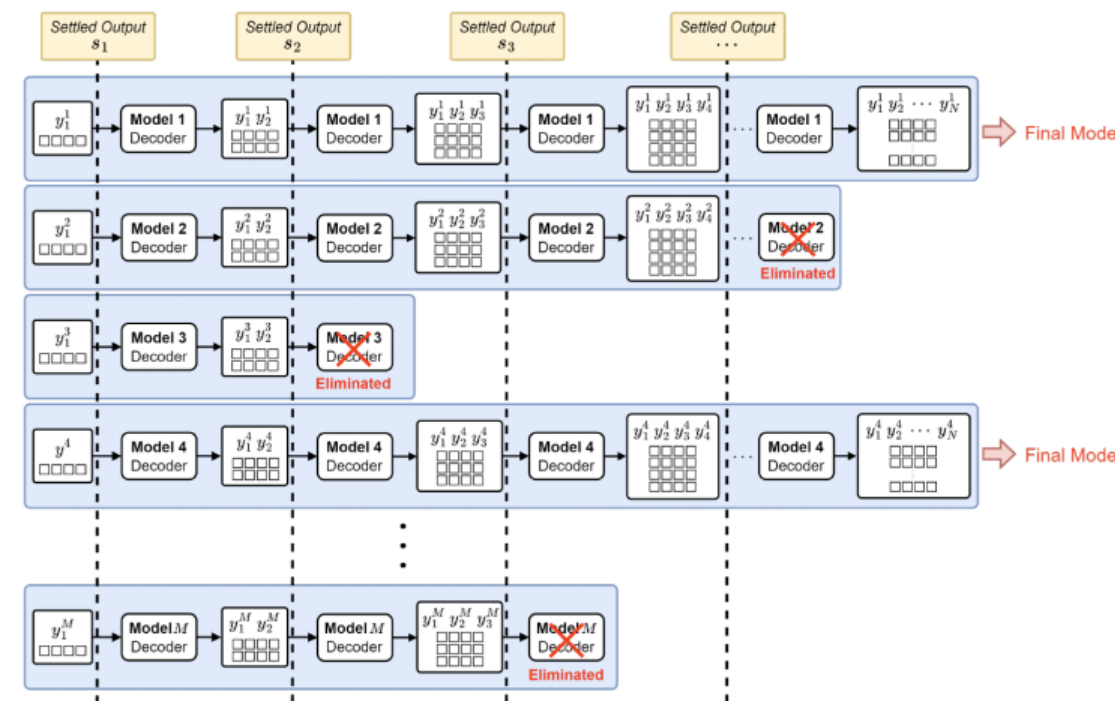


Figure 3.9: Overall Structure of Survival in Transformer

Survival Ensemble

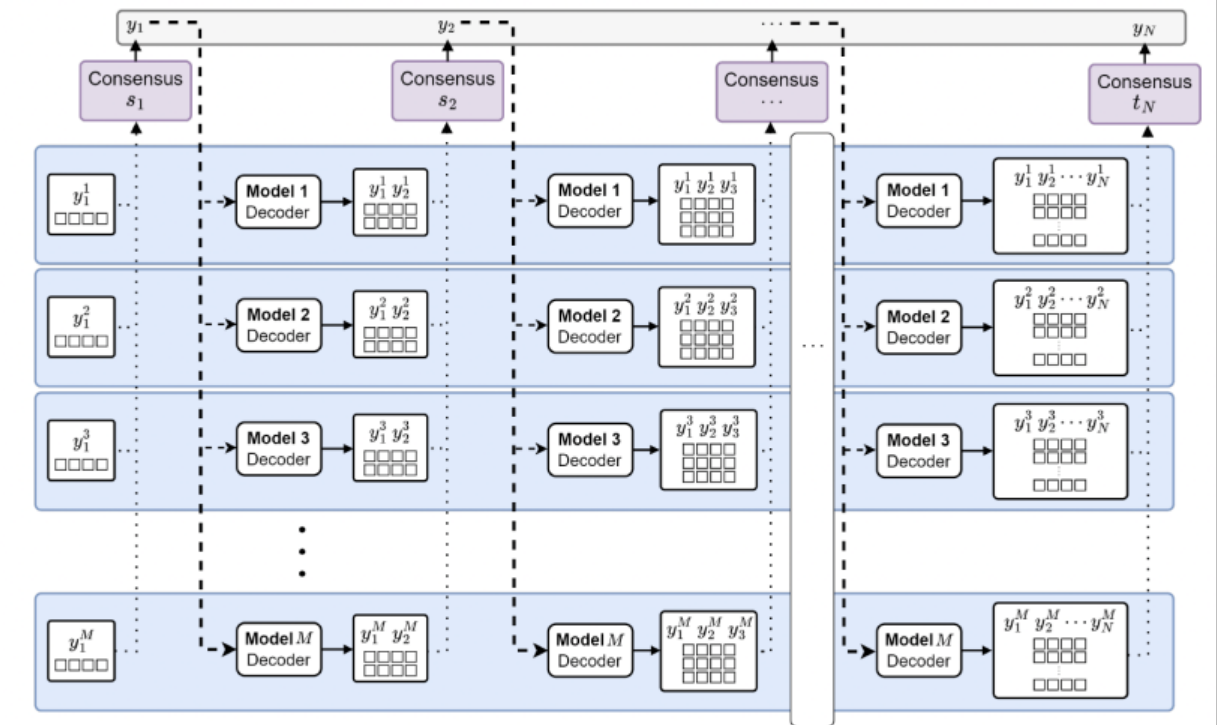


Figure 3.11: Overall Structure of Consensus in Transformer

Consensus Ensemble

REGEN: Recurrent Ensemble Methods for Generative Models

Ahjeong Park, Youngmi Park, Chulyun Kim

Experiments Setup

Seq2Seq

1) Neural Machine Translation (Spain-English)

- 단일 모델 총 15개
- Baselines: Majority, Independent Ensemble
- Metric: TQE(Translation Quality Estimation, BERTScore, BLEU, ROUGE)

2) String Arithmetic

- 단일 모델 총 5개
- Metric: Accuracy

Transformer

1) Neural Machine Translation(German-English)

- 단일 모델 총 10개
- Baselines: Majority, Checkpoint Ensemble
- Metric: TQE(Translation Quality Estimation, BERTScore, BLEU, ROUGE)

REGEN: Recurrent Ensemble Methods for Generative Models

Ahjeong Park, Youngmi Park, Chulyun Kim

Experiments Results

Neural Machine Translation in Seq2Seq

Table 4.2: Comparison of the performance of TQE (%), BLEU (%), and F1 BERT (%) for various ensemble methods and single model (avg) in Spanish-English machine translation

Model	TQE	BLEU	F1 BERT
Independent (top 1) [35]	60.34	19.49	78.40
Independent (top 2) [35]	62.59	20.5	78.91
Independent (top 3) [35]	62.83	20.64	79.06
Traditional Majority	72.43	21.9	80.20
Survival (REGEN)	73.10	21.28	80.31
Consensus (REGEN)	74.33	22.71	80.53
Single (Avg, 10 epoch)	69.09	19.56	79.28

- 본 연구의 3가지 앙상블 모두 단일 모델을 뛰어 넘은 것을 확인할 수 있었습니다.
- Recurrent Ensemble > 기존 앙상블(Independent Ensemble)

String Arithmetic in Seq2Seq

Table 4.7: Comparison of performance of Accuracy for various ensemble methods and single model in String Arithmetic

Model	Accuracy (%)
Traditional Majority	76.38
Survival (REGEN)	77.53
Consensus (REGEN)	86.75
Single (Avg)	64.70
Single 1	65.46
Single 2	68.32
Single 3	67.90
Single 4	58.14
Single 5	63.70

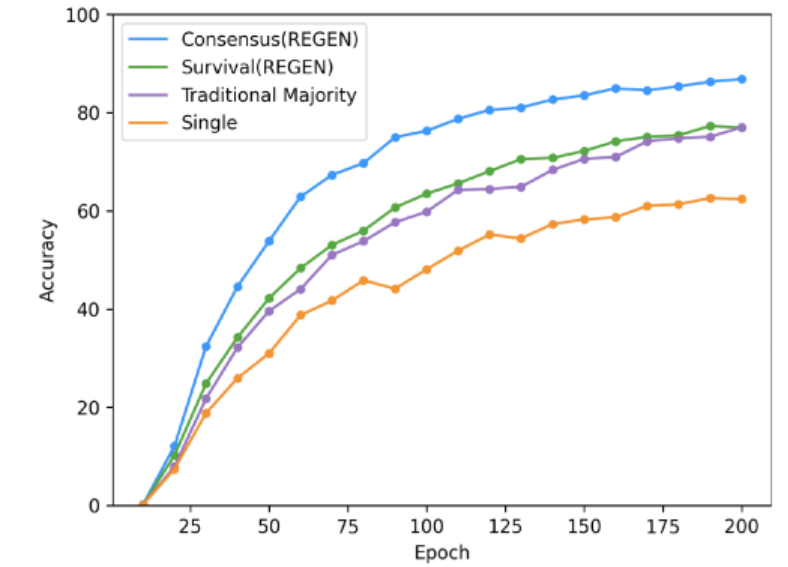


Figure 4.2: Accuracy evaluated during Training

- Recurrent Ensemble > baseline: Majority Ensemble
- Consensus > Survival > Majority
- 기계 번역 실험과 마찬가지로 여전히 Recurrent Ensemble이 기존의 앙상블 방법보다 뛰어남을 확인할 수 있었습니다.

REGEN: Recurrent Ensemble Methods for Generative Models

Ahjeong Park, Youngmi Park, Chulyun Kim

Experiments Results

Neural Machine Translation in Transformer

Table 4.9: Comparison of the performance of TQE (%), BLEU (%), and F1 BERT (%) scores for the various ensemble methods and single model in German-English machine translation

Model	TQE	BLEU	F1 BERT
Checkpoint (190, 195, best) [9]	81.23	23.62	92.91
Checkpoint (190, 195) [9]	81.5	23.76	93.01
Traditional Majority	83.10	24.02	93.56
Survival (REGEN)	84.48	25.04	94.21
Consensus (REGEN)	84.55	25.15	94.23
Single (Avg, 200 Epoch)	84.19	24.93	94.09

Table 4.8: TQE, BLEU, F1 BERT 성능 비교

- Recurrent Ensemble > Baseline(Majority, 기존 앙상블)
- Recurrent Ensemble > Single Models

Table 4.10: Comparison of the performance of n-gram BLEU (%) for various ensemble methods and single model (avg) in

Model	BLEU 1	BLEU 2	BLEU 3	BLEU 4
Checkpoint (190, 195, best) [9]	41.68	32.27	21.08	13.07
Checkpoint (190, 195) [9]	41.76	32.43	21.21	13.2
Traditional Majority	42.3	32.6	21.43	13.38
Survival (REGEN)	42.63	33.96	22.55	14.14
Consensus (REGEN)	42.8	34.09	22.65	14.21
Single (Avg, 200 Epoch)	42.7	33.78	22.41	14.05

Table 4.9: N-gram BLEU 성능 비교

- Recurrent Ensemble > Majority > Checkpoint Ensemble(기존 앙상블)
- Recurrent Ensemble > Single Model