Test Time Scaling 방법론을 활용한 산업분야데이터에 대한 LLM의 추론 방법

임세훈, 주철우, 조현중*

고려대학교, *고려대학교

tpgns621@korea.ac.kr, jupd2000@korea.ac.kr, *raycho@korea.ac.kr

LLM's Inference Method for Industrial Data Using Test Time Scaling Methodology

Lim Se Hun, Joo Chul Woo, Cho Hyeon Joong*

Korea Univ, *Korea Univ.

요 약

본 논문은 스마트 팩토리 등 산업 현장에서 요구되는 산업 이상 탐지(Industrial Anomaly Detection, IAD)를 수행하기 위해, 대규모 멀티모달 모델(MLLM)에 Test-Time Scaling(TTS) 기법을 적용하는 방법론을 제안한다. 최근 MLLM은 뛰어난 범용성을 보이고 있으나, 미세한 결함을 다루는 산업 데이터의 특성상 zero-shot 추론만으로는 한계를 보이고 있다. 이에 본 연구에서는 모델의 구조적 변경이나 추가적인 학습 없이 추론 단계에서의 연산 자원 배분만으로 성능을 개선하고자 하는 Self-Consistency 알고리즘을 도입하였다. Qwen3-VL 모델을 기반으로 산업 데이터셋에 대해 다수결 투표(Majority Voting) 방식의 추론을 수행하였으며, 이를 통해 응답 개수와 추론 정확도 간의 상관관계를 분석하였다. 실험 결과를 통해 제한된 컴퓨팅 자원 하에서의 단순 스케일링이 갖는 한계점을 규명하고, 향후 산업용 에이전트가 나아가야 할 효율적인 추론 전략과 알고리즘 고도화 방향(Weighted Voting 등)을 제시한다.

Ⅰ. 서론

스마트 팩토리와 무인 공장 실현을 위해 자동 비전 검사는 산업 분야에서 핵심이다. 기존의 산업 이상 탐지(Industrial Anomaly Detection, IAD)[1]의 판별 모델들은 특정 데이터에 과적합 되어 있다. 판별 모델들은 생산 라인을 변경할 경우 재학습이 필요하고, 결함에 대해 보고서를 자세하게 제공하지 못한다는 한계가 있었다. 대규모 멀티모달 모델(MLLM)[2]은 이러한 문제점을 해결하고 산업 이상 탐지에서 새로운 대안으로 떠오르고 있다.

그러나 IAD분야의 벤치마크인 MMAD(Multi-Modal Anomaly Detection)[3]의 연구 결과에 따르면, 현재 최첨단 MLLM(Multimodal Large Language Models)조차 산업 현장이 요구하는 정밀도에는 적합하지 않다는 추세이다. GPT-4o와 같은 상용 모델조차 복잡한 결함의 유무를 판단하거나 그 원인을 추론하는 질문에서 뚜렷한 성능 저하를 보였으며, RAG(Retrival-Augmented Generation)나 Expert Agent 방식의 접근도 모델 자체의 근본적인 시각적 추론 한계를 극복하지는 못했다. 이는 단순한 데이터의 양적 증가나 파라미터의 확장만으로는 미세한 결함을 다루는 고도의 산업적 추론 작업을 해결하기 어렵다는 것을 시사한다.

최근 LLM에 대한 연구에서는 사전 학습에서의 스케일링이 가진 자원 효율성 한계를 극복하기 위해서 추론 시점에 추가적인 연산 자원을 할당하는 Test-Time Scaling(TTS)[4]이 새로운 흐름으로 자리 잡고 있다. OpenAI의 o1이나 DeepSeek의 R1과 같은 모델들이 복잡한 수학이나 코딩 문제에서 "System2" 적인 사고 과정을 통해 추론 성능을 개선할 수 있음을 보였다. 이는 모델이 바로 답변을 생성하기보다는 문제를 깊이 생각하는 과정을 통하여 성능을 최대한으로 발휘할 수 있음을 보여준다.

본 논문은 TTS의 개념을 산업 이상 탐지 영역으로 확장하여, MLLM의 결함 탐지 성능을 개선하는 것을 목표로 한다. 기존 MMAD 벤치마크에서 드러난 추론에서의 결함을 보완하기 위하여 TTS에서 Sequential Scaling에 해당하는 Self-Consistency 전략을 적용하였다. 이를 통해 모델의 크기를 키우지 않고, 추론 시점에서의 연산 배분만으로 산업 이상 탐지 영역에서의 성능을 보완할 수 있음을 보이고자 한다.

Ⅱ. 본론

2.1 MLLM(Multimodal Large Language Model)

CLIP(Contrastive Language-Image Pre-training)[5]과 같고, 강력한 zero-shot 분류 성능을 보여주는 VLM(Vision-Language Model)은 사전 학습된 모델이나 특징을 특정 작업에 맞게 조정하여 활용하는 다운스트림 비전 태스크에 다양하게 적용되어 왔다. 이와 같은 VLM의 인코더와 LLM을 결합한 것이 MLLM이다. MLLM은 시각적 콘텐츠와 관련된 텍스트 기반 상호작용을 할 수 있도록 만든다. 최근 연구들은 이러한 다운스트림 태스크에 대한 추론 근거를 제공하기 위하여 MLLM을 활용하고 있다.

2.2 IAD(Instuction Anomaly Detection)

기존 IAD 연구는 결함 샘플 없이 결함을 식별하고 위치를 파악하는 것을 목표로 했다. 기존 IAD 방법은 수많은 정상 샘플에 대해 학습한 다음 이상 감지 기술을 사용하여 테스트 샘플에서 이상을 식별했다. 최근 연구는 IAD의 일반화 기능에 초점을 맞췄다. CLIP과 Vision-Language Model을 활용하여 예시 샘플이 주어지는 few-shot 모델, 예시 샘플 없이 처리하는 zero-shot 모델이 등장했다. 그러나 이러한 판별 모델들은 CLIP 모델에서 미리 정의된 이상 개념을 너무 의존하여 새로운 시나리오로 일반화하는 기능을 제한한다. MLLM은 시각적 구성요소와 함께 복잡한 텍

스트 입력을 이해하고 다양한 응답을 제공할 수 있으므로 이러한 문제를 해결하는데 도움이 된다.

2.3 Test-Time Scaling(TTS)

초기의 LLM은 더 많은 데이터와 파라미터를 사용하여 모델을 훈련시키는 "Training-Time Scaling"을 통하여 언어의 이해, 추론, 지식 적용 능력을 학습했다. 그러나 이러한 training-time scaling 방식은 자원을 많이 사용하고 사람과 관련된 데이터에 접근할 수 있는 능력이 제한적이라는 문제에 부딪히면서 발전 속도가 둔화되기 시작했다. 이러한 한계로 연구자들은 LLM의 지능을 추론 시점에서 발현시키는 방법에 주목하기 시작했다. 인간이 복잡한 문제에 직면했을 때 더 깊고 신중하게 사고하여 더 나은 결과를 도출하는 것에서 영감을 받아서 추론 시점에 추가적인 연산을 할당하여 성능을 높이는 방법들이 등장했다. 이를 "Test-Time Scaling"(TTS)이라고 부르며, 모델의 지능을 추론 시점에서 발현시키는 방법이다.

Ⅲ. 실험

3.1 Qwen3-vl-8B-Instruct

본 연구에서는 TTS의 효과를 극대화하기 위하여 Qwen3-vl-8B-Instruct[6]를 베이스 모델로 채택했다. Qwen3는 모델 아키텍처 내에 복잡한 다단계 추론을 위한 '사고 모드'와 신속한 응답을 위한 '비사고 모드'를 통합한 모델이다. Qwen3가 도입한 '사고 예산' 기능은 사용자가 추론에 투입할 계산 리소스를 직접 제어할 수 있게 하고, 작업의 복잡도에 따라 추론의 정도를 유연하게 조절할 수 있게 한다.

아래 표는 MMAD 데이터셋에 대하여 Qwen2.5-vl-7B-Instruct와 Qwen3-vl-8B-Instruct의 성능을 평가한 결과이다. 이전 모델 Qwen2.5보다 Qwen3의 성능이 더 좋은 것을 확인할 수 있다. 이는 Qwen3가 추론 시점의 연산 배분을 통한 성능 향상을 이뤄냈다는 것을 알 수 있다.

3.2 Self consistency

Self-Consistency는 모델이 생성한 다수의 후보 응답 중 가장 빈도수가 높은 답변을 최종 결과로 채택하여 추론의 성능을 높인다. 이를 Qwen3 모델의 추론 단계에 적용하여, 같은 입력에 대해 3회의 반복하여 3가지 응답을 생성한 뒤 제일 많이 나온 응답을 고르는 다수결 투표를 통해 최종 응답을 결정하는 방식을 사용했다.

실험 결과, 첨부된 표와 같이 전체 데이터셋에 대해 Average 72.34%, F1 Score 73.29%를 기록했다. MVTec-AD 데이터셋에서는 84.27%로 3회 반복이 비교적 높은 성능을 보였으나, GoodsAD와 같이 난이도 높은 데이터셋에서는 66.75%에 머물렀다. 결과적으로 3회 반복 추론에서는 단일 추론 대비 통계적으로 유의미한 성능 향상이 어렵다는 것을 확인했다.

이에 대해서 찾아본 원인은 다음과 같다. Self-Consistency의 효과를 보여주기 위해서 충분한 수의 응답을 통하여 오답의 분산을 줄이고 정답의 일관성을 얻어야 하지만, 3회라는 적은 횟수는 모델의 추론 성능을 개선하기에는 부족했던 것으로 판단 된다. 이후 연구에서는 응답 개수를 대폭 늘리거나, 단순 다수결이 아닌 각 응답의 신뢰도를 반영한 가중 투표 방식을 도입하는 것도 고려하여 성능 개선할 필요가 있다.

	Anomaly Detection	Defect Classification	Defect Localization	Defect Description	Defect Analysis	Object Classification	Object Analysis	Average	Recall	Precision	F1
MVTec-AD	80.672	65.64315	72.50629	72.21764	88.46473	99.13793	90.27067	81.2732	76.18271	93.12064	83.8044
MVTec-LOCO	60.19007	42.36253	48.98167	57.90554	70.12579	84.375	77.5724	63.07328	56.85484	72.96248	63.90935
VisA	69.96054	40.5042	60.81871	59.7479	67.66982	91.22995	81.46828	67.34277	51.58598	84.89011	64.17445
Goods AD	50.61397	57.43707	53.05445	60.54014	77.67123	88.02642	77.24978	66.37044	81.05457	56.6167	66.66667
Average	65.33914	51.48674	58.84028	62.60281	75.98289	90.69233	81.64028	69.51492	66.41952	76.89748	69.63872

표 1. Qwen2.5-vl-7B-Instruct 모델 성능 결과

	Anomaly Detection	Defect Classification	Defect Localization	Defect Description	Defect Analysis	Object Classification	Object Analysis	Average	Recall	Precision	F1
MVTec-AD	84.62769	73.11203	71.75189	78.31822	91.45228	97.84483	93.05048	84.3082	91.19086	91.63934	91.41455
MVTec-LOCO	62.99093	47.65784	51.12016	63.44969	74.10901	86.58088	79.79051	66.52843	65.42339	74.17143	69.5233
VisA	77.29695	51.59664	55.80618	65.46218	77.55804	94.0107	81.75339	71.9263	68.69783	86.08787	76.41597
Goods AD	55.2027	62.62395	48.07437	61.44036	74.17808	89.67795	75.98098	66.73977	50.70509	61.80867	55.70899
Average	70.02957	58.74762	56.68815	67.16761	79.32435	92.02859	82.64384	72.37568	69.00429	78.42683	73.2657

표 2. Qwen3-vl-8B-Instruct 성능 결과

	Anomaly Detection	Defect Classification	Defect Localization	Defect Description	Defect Analysis	Object Classification	Object Analysis	Average	Recall	Precision	F1
MVTec-AD	84.66847	73.27801	71.66806	78.31822	91.36929	97.62931	92.97732	84.27267	91.27243	91.64619	91.45893
MVTec-LOCO	63.07819	47.55601	51.42566	63.7577	74.21384	85.47794	79.85213	66.48021	65.42339	74.25629	69.56056
VisA	77.30824	51.0084	55.6391	65.54622	77.64402	94.11765	81.78902	71.86466	68.61436	86.16352	76.39405
Goods AD	55.29738	62.4714	48.00797	61.5904	74.0411	89.67795	76.18058	66.7524	50.70509	61.9012	55.74655
Average	70.08807	58.57845	56.6852	67.30313	79.31706	91.72571	82.69976	72.34248	69.00382	78.4918	73.29002

표 3. Self consistency를 적용한 Qwen3-vl-8B-Instruct 성능 결과

Ⅳ. 결론

본 연구를 통해 MMAD 데이터셋에 대한 Self-Consistency 적용은 모델의 추론 성능을 개선하기에 부족하며, 단순 다수결 투표만으로는 의미 있는 성능 향상을 끌어내기 어렵다는 한계를 확인하였다. 결과적으로 추론의 전략을 단순하게 연산을 반복하기보다 통계적 신뢰성 확보를 위한 응답 개수의 확대와 신뢰도 기반의 가중 투표 도입 등 추론에 추가적인 연산을 부여하는 방법을 다양하게 적용하면 산업 현장에서 요구하는 정밀도를 달성할 수 있다는 결론을 도출하였다.

ACKNOWLEDGMENT

본 연구는 과학기술정보통신부의 재원으로 한국연구재단 중견연구(창의연구형) 사업의 지원을 받아 수행되었음(과제번호: RS-2025-16066493).

참 고 문 헌

[1] Xi Jiang, Guoyang Xie, Jinbao Wang, Yong Liu, Chengjie Wang, Feng Zheng, and Yaochu Jin. "A survey of visual sensory anomaly detection." arXiv preprint arXiv:2202.07006 (2022).

[2] Y. Jin et al., "Efficient Multimodal Large Language Models: A Survey," Vis. Intell., vol. 3, no. 1, Art. no. 27, Dec. 2025, doi: 10.1007/s44267-025-00099-6.

[3] X. Jiang et al. "MMAD: A Comprehensive Benchmark for Multimodal Large Language Models in Industrial Anomaly Detection." in Proc. 13th Int. Conf. Learn. Represent. (ICLR), Singapore, Apr. 2025, pp. 48346 - 48368.

[4] X. Wang et al., "Self-consistency improves chain of thought reasoning in language models," in Proc. 11th Int. Conf. Learn. Represent. (ICLR), Kigali, Rwanda, May 2023.

[5] A. Radford et al., "Learning Transferable Visual Models From Natural Language Supervision," in Proc. 38th Int. Conf. Mach. Learn. (ICML), vol. 139, July 2021, pp. 8748 - 8763.

[6] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang et al. "Qwen3 Technical Report." arXiv preprint arXiv:2505.09388 (2025).