SAS 2016

스코틀랜드, 에딘버러

2016년 9월 7일 - 2016년 9월 10일

서울대학교 허기홍

들어가며

프로그램 정적 분석 분야 전문 학회인 SAS 에 참석차 에딘버러에 다녀왔다. 기쁘게도 그동안 즐겁게 연구한 결과가 채택되어 학회에서 발표할 수 있는 기회도 갖게 되었다. SAS 학회의 특성에 걸맞은 골수 분석 전문가들과 부대낀 뜻깊은 일주일이었다. 이번에는 TAPAS 워크샵도 학회 하루전에 함께 열린 까닭에, 거기서도 좋은 발표를 들을 수 있었다. 학회가 열린 에딘버러는 2014년 PLDI 에 참가했을때 이후 2년만이었다. 그 때도, 그리고 올해도 채택된 논문을 들고 가게 된 행운이 가득한 도시이다. 공항에서부터 한적한 시골길을 지나 고즈넉한 구 도심으로 향하는 길은 여전히 아름다웠다.

논문 발표

Learning a Variable-Clustering Strategy for Octagon From Labeled Data Generated by a Static Analysis

허기홍, 오학주, 양홍석

이 연구에서는 기계학습을 이용하여 팔각 관계 분석 (octagon relational analysis) 를 꼭 필요한 부분에만 적용하는 방법을 제시하였다. 관계 분석은 비관계 분석에 비해 정확하지만 비용이 너무 커서 곧바로 적용하기가 어렵다. 따라서 필요한 부분만 잘 선별해서 적용하는 것이 중요한데, 관계 분석이 꼭 필요한 부분만 선별해내는 방법은 2년전 PLDI 에 제출한 바 있다. 그 때는 같은 목표를 달성하기 위해서 정적분석을 사용하였다. 관계 분석의 효과를 쉽고 빠르게 가늠하는 예비분석을 통해서 였다. 예비분석을 이용하면 관계 분석을 효율적으로 사용할 수 있지만, 예비분석 자체의 비용도 여전히 큰 것이 문제였다. 그래서 이 연구에서는 예비분석을 학습과정에서 선생님으로 사용하고 기계학습 알고리즘을 이용해 관계 분석이 필요한 부분만 선별하는 기술을 배우는 방법을 제시한 것이다. 이를 통하면 기존 방식보다 수십배 빠르게 예측할 수 있어 효율적이다.

내 발표는 다행히 첫날이어서 많은 사람들의 집중력이 생생한 때였고, SAS 참석자들은 대부분 관계 분석에 대해 잘 알고 있어서 내용이 잘 전달된 듯하였다. 이 논문의 핵심은 학습 데이터를 정적 분석기가 자동으로 생성해준다는 것인데, 그래도 여전히 특질(feature)을 사람이 만들어 주어야한다는 단점이 있다.

아니나 다를까 발표 후에는 그에 관해 궁금해 했다. 현재 진행하고 있는 특질 자동 생성 연구에 대해 간략히 대답해주었다. 우리가 가고 있는 방향에 더 확신이 들었고 자신감이 생겼다.

쉬는 시간에도 분석기를 만들고 있는 여러 사람들과 이야기를 나누었는데, 기계학습으로 정적분석의 성능을 높이는 것에 흥미를 보였다. 특히 호주에서 포인터 분석을 연구하는 중국인 Tian 이라는 친구와는 자주 이야기를 나누었고, 우연히도 한국으로 올 때 중간 경유지인 암스테르담까지 같은 비행기를 타게 되어 비행기안에서도 열띤 논의를 했다. 다들 기계학습이 좋은 도구라는 것은 들어서 알고 있는데, 분석 문제에다 어떻게 적용해야할지 막연해 하는 듯 하였다. 그도 그럴것이, 엄밀한 의미 정의를 바탕으로 하는 정적 분석과 통계적인 추론을 바탕으로 하는 기계학습은 얼핏 전혀 어울리지 않는 한 쌍처럼 보인다. 우리도 처음에는 막연했지만, 이제 갓 걸음마를 떼면서 쉬운 문제부터 하나씩 풀어가고 있는 단계인것 같다. 비슷한 고민을 하는 사람들에게 우리의 논문이 좋은 참고자료가 되었으면 하는 바람이다.

초청 강연

Probabilistic Programming: Formal Techniques and Applications

Aditya Nori

과거에는 프로그램 분석 연구를 하다가 요즘은 기계학습 분야에서 훌륭한 연구를 많이 하고 있는 Aditya Nori 의 발표였다. 확률적 프로그래밍 (probabilistic programming) 의 개념에 관해 설명해주었고 다양한 적용사례들을 이야기해 주었다. 확률적 프로그램은 온갖 의존성이 복잡하게 얽혀있는 확률 모델을 전통적인 방법으로 표현하는 대신 프로그램으로 표현하는 방법이다. 이를 이용해서 다양한 확률적 추론을 쉽게 할 수 있는데, 현실에 적용한 대표적인 사례로는 Microsoft 의 게임기 XBox 에서 적절한 온라인 대전상대를 연결해주는 TrueSkill 시스템이 있다.

확률적 프로그래밍보다도 사실 더 흥미로웠던 것은 발표 후반부에 나온 기계학습 관련 이야기였다. 발표자는 궁극적으로 인공신경망을 사람이 더 잘 이해하도록 도와주는 연구를 하고 싶어했다. 인공신경망도 결국 입력을 받아서 계산하고 출력을 내는 프로그램으로 볼 수 있기 때문에, 전통적인 PL 문제가 그대로 존재한다. 이미 Aditya Nori 는 최근에는 인공신경망을 디버깅하거나, 오류가 있는 입력을 찾아내는 연구 결과를 발표하였다. 정말 흥미롭다. 기계학습에 부정적인 많은 사람들이 말하는 것 중 하나가, 그 내부가 마치 블랙박스 같아서 결과를 설명하기 어렵다는 것인데 이러한 노력이 지속되면 곧 그런 우려가

희미해질 것이다. 그 길을 가는데 또 어떤 기술이 필요할까? 프로그램 정적 분석, 검증을 인공신경망 세계로 옮겨보면 그것은 어떤 의미가 있으며 어떻게 하면 되는 것일까?

Learning from Programs: Probabilistic Models, Program Analysis and Synthesis

Martin Vechev

요즘 기계학습을 이용한 프로그래밍 언어 연구로 각광받고 있는 Martin Vechev 의 초청강연이다. 나와 관심사도 비슷하고, 지난 PLDI 에서 그의 초청강연을 한 번 들은 터라 쉽게 내용이 다가왔다. 과거 수년간 POPL, PLDI, ICML 등에 낸 논문에 관한 전반적인 이야기와 경험, 계획등을 이야기해 주었다. 가장 유명한 연구로는 난잡하게 쓰여 있는 자바스크립트 코드를 깔끔하게 바꾸어주는 JSNice 가 있고, 그 이외에도 코드 자동완성, 동시성 프로그램 분석, 그리고 효율적인 팔각 관계 분석 라이브러리 등 다양한 분야를 연구하는 사람이다.

우리는 프로그램 분석을 위해서 기계학습을 이용한다면, 이 연구진에서는 기계학습[1]을 위해 프로그램 분석을 이용하는 점이 인상적이다. 발표에서 강조한 내용 중, 프로그램을 있는 그대로 학습한 결과와, 포인터 분석 결과를 함께 학습한 결과를 보여주었다. 그들의 실험에 따르면, 정적 분석 정보는 대략 데이터 10배에 해당하는 효과를 가진다고 했다. 즉, 같은 결과를 내는 인공지능을 만드는데 정적 분석 정보를 이용하면 10배 더 적은 데이터로 쉽게 학습할 수 있다는 것이다.

지난번에 이어서 정말 재미있는 발표였다. 알찬 내용, 자신감있는 표정과 어조도 여전했다. 전날 Aditya Nori 가 강조한 것과 마찬가지로 기계학습 결과가 왜 그렇게 나왔는지 사람이 잘 이해하는 것에 대해 진지하게 고민하고 있는것도 엿볼수 있었다. 다음 연구도 기대가 된다.

인상깊은 논문 발표

Making k-Object-Sensitive Pointer Analysis More Precise with Still k-Limiting

Tian Tan, Yue Li and Jingling Xue

학회 기간동안 많은 이야기를 나누었던 Tian Tan 의 논문 발표였다. 전 학회를 통틀어 가장 짜임새있는 발표였다고 생각한다. 초반에 문제와 핵심 아이디어 전달을 명료하게 했기 때문에 청중으로서 듣는데 집중력을 잃지 않았다. 이 때문에 청자가 후반의 세부 내용을 다 이해하지 못하더라도 큰 틀에서 무슨 일을 한 것인지 파악한채 논문을 펴게 만들 수 있었다.

일반적으로 프로그램 분석에서 함수 호출의 문맥을 구분할 때 널리 쓰이는 방법이 최신 k 개 호출 이력을 기억하는 것이다. 오른쪽 예제 코드를 분석한다고 쳐보자. 함수 h의 내부가 실행되는 상황은 실제로는 두 경우이다.

```
int main(){
1: f();
2: f();
}

void f(){
3: g();
}

void g(){
4: h();
}
```

[1] 물론 여기서 학습대상은 프로그램이다. 프로그램의 여러 성질을 학습하는 것이 목표.

1) 1에서 f 를 호출한 후, 3에서 g를 호출한 후, 4에서 h를 호출한 경우. (1-3-4)
2) 2에서 f 를 호출한 후, 3에서 g를 호출한 후, 4에서 h를 호출한 경우. (2-3-4)

하지만 k=2 로 정하고 분석을 하면, 최신 함수 호출 이력을 최대 2개까지만 기억하기 때문에 이 두상황을 구분하지 못한다. 즉, 두 경우 모두 3-4 라는 동일한 문맥으로 요약되어서 분석의 정확도를 잃게 된다. 이것이 일반적으로 널리 쓰이는 문맥 구분 방식이다.[2]

이 논문의 문제 의식은 간단하지만 재미있는 관찰에서 출발한다. 이 예제에서 h 로 흘러들어가는 모든 경우는 늘 3번 지점을 지나가는데 문맥 구분시 3번을 고려할 필요가 있냐는 것이다. 대신에 h로 흘러들어가는 문맥은 1-4, 2-4 로만 기록해서 여전히 k=2 인 경우에도 더 정확하게 분석할수 있도록 하자는 것이다. 이렇게 하면 기존 분석 방법보다 비용이 조금 더 들긴하지만 정확도가 대폭 상승하는 효과가 있다고 한다.

나도 문맥을 구분하는 분석에 관해서 연구한 적이 있었기 때문에 흥미를 갖고 들었다. 발표자와 이야기해보니 예비 분석을 통해 필요한 문맥을 예측하는 우리의 이전 연구도 잘 알고 있었다. 밀접한 관심사를 가진 사람과 즐겁게 논의했던 좋은 시간이었다.

Generalized Homogeneous Polynomials for Efficient Template-Based Nonlinear Invariant Synthesis

Kensuke Kojima, Minoru Kinoshita and Kohei Suenaga

이 논문을 발표한 Kohei Suenaga 는 몇 달전 PLDI 에서 처음 만났다. 이번 SAS 에서 내가 발표한 논문 내용을 PLDI 에서 포스터로 발표하고 있을 때 였다. 그때 내 연구에 관심을 많이 보이고 자기도 마침 SAS 에서 수치 관계 분석에 관해 발표하니 그때 더 이야기 하자고 했었는데 다시 반갑게 만났다.

이 연구에서 특정한 성질을 가진 순환문의 불변식 (invariant) 을 대상으로 하고 있다. 일반적으로 프로그램 분석에서 순환문의 불변식을 잘 찾는 것은 중요하지만 매우 어려운 문제이다. 하지만 특별한 경우에는 그에 알맞은 방식으로 다소 쉽게 찾을 수 있는 경우가 있다. 예를 들어 이 논문에서 대상으로 삼는 프로그램처럼, 불변식이 특정 변수들의 다항식 형태라는 것을 우리가 미리 알고 있는 경우에는 그 다항식을 구성하는 각 항의 계수만 잘 찾으면 문제가 풀린다. 즉, 불변식이 가질수 있는 기본 형태(template)를 미리 만들어 놓고, 세부 내용만 알맞게 채우는 방식이다.

하지만 다항식이 고차원이고 등장하는 변수가 많아지면 수많은 조합을 다 고려해야하므로 문제가 복잡해지는데 이 논문에서는 간단한 아이디어를 이용하여 해결하였다. 핵심은 형틀(template) 에서 “개념적”으로 말이 안되는 항은 미리 제거하여 문제 공간을 줄이는 것이다. 이 사람들이 대상으로 하는 기계 제어 프로그램 등은 임의로 구성되는 것이 아니라 물리 법칙 등 잘 정의된 개념을 구현한 것인 경우가 많다. 예를 들어 순환문의 불변식이 변수 s, v, a, t 로 구성되어 있는 n차 식이라고 하자. 그리고 각 변수는 거리 (s), 속도 (v), 가속도 (a), 시간 (t) 를 나타낸다. 기본적으로 이 변수로 구성된 n 차식의 모든 항을 고려할 것이다. 하지만 물리학의 법칙상, 속도 (v) 에 관한 불변식에는 a*t, s/t 등과 같은 항만 사용하지 a/t, s*t 같이 개념적으로 말이 안되는 항은 등장하지 않을 것이다. 따라서 말이 되는 일부만 고려하여 불변식을 잘 찾을 수 있다는 이야기이고, 여기서는 그것을 타입시스템을 이용하여 추려내었다.

[2] 논문에서는 이 예제와 다르게 객체중심 프로그램 분석을 가정하고 있기 때문에, 호출지점 대신 객체로 문맥을 구분하지만 문제와 핵심 아이디어는 일맥상통한다.

PLDI 에서 포스터 발표를 했을때, Kohei 가 자기가 한 일도 말하자면 선별적 관계분석이라고 했는데 발표를 듣고나니 그 의미를 알았다. 재미있는 아이디어이다. 우리가 마주친 문제 중에서도 프로그래머의 상식을 믿고 그에 기대어 과감하게 시도해볼만한 것은 없을까 생각해 본다.

Programming a Dataflow Analysis in Flix

Magnus Madsen, Ming-Ho Yee, Ondřej Lhoták

올해 PLDI 에 발표된 Flix 시스템을 TAPAS 워크샵에서 소개하였다. Flix 는 고차원 언어로 간편하게 기술한 프로그램의 분석 명세에서 분석기를 자동으로 만들어주는 시스템이다. Flix 는 기본적으로 Datalog 라는 언어의 표현력을 확장한 것이다. Datalog 프로그램은 여러 규칙 (rule) 으로 구성되고, 목표는 그 규칙을 만족하면서 가장 작은 모델을 찾는 것이다. 규칙을 기술하기가 쉽고 모델을 효율적으로 찾을수 있기 때문에, 프로그램에서 뽑은 방정식의 최소 고정점을 찾을 때 (즉, 정적 분석을 할 때) 만들때 많이 사용한다. 하지만 Datalog 는 제약이 있어서 비교적 간단한 분석을 기술하는데 밖에 사용하지 못하는데, 이 연구는 그 제약을 어느정도 극복하여 좀 더 일반적인 분석을 할수 있게 만든 것이 핵심이다. Datalog 는 모든 규칙을 명시적으로 하나씩 나열해야 했기 때문에 무한한 규칙을 기술할 수 없다. 따라서 pointer 분석 등 유한한 대상을 다루는 분석에 주로 적용을 했다. 반면 Flix 는 무한한 래티스 (lattice) 와 함수를 기술할 수 있기 때문에 상수 분석등도 만들어 낼 수 있다.

PLDI 에서도 재미있게 발표를 들은 터라 발표후에 찾아가서 간단히 이야기를 나누었다. 상수 도메인처럼 옆으로 무한한 래티스는 쉽게 다룰 수 있어 보이는데 구간 도메인처럼 높이가 무한한 도메인도 다룰 수 있는지 궁금했다. 그런 경우는 아직 불완전하다고 했다. 또한 Flix 로 만든 프로그램은 늘 끝난다는 것을 보장할 수가 없기 때문에 추가적인 모듈을 붙여서 검사한다고 했는데 그 부분이 얼마나 강력한지도 궁금한데 그때 미처 물어보지 못해 아쉽다.

학회 즐기기

에딘버러는 이번이 두번째인데 다시금 느끼지만 참 고풍스럽고 아름다운 도시이다. 아침 일찍 공항에 도착해서 새로 생긴 트램을 타고 도심까지 향하는데, 밤서리를 차게 맞은 자연 경관이 일품이었다. 화창한 날씨와 푸른 들이 어우러진 풍경에 익숙해질때 즈음 전통 건물로 가득찬 도심으로 빨려들어가는 느낌도 신선했다. 영국 날씨가 고약하다고들 한다. 지난번 방문 때는 물론이고 첫날까지도 운좋게도 느끼지 못했지만 이번에 제대로 걸렸다. 비는 그리 심하지 않았지만 바람이 너무 많이 불어서 우산이 무용지물이 되었다. 길거리에 아무도 우산을 안들고 다니는 이유를 몸소 느끼고 왔다.

사람들을 만나서 교류하는 것은 즐거운 일이다. 이번 SAS 에는 한국인이라고는 나, 그리고 같이 간 학주형 뿐이었다. (한국에서 프로그램 분석분야 연구가 더 활발해 졌으면 한다.) 그래도 제법 학회를 다녀본 터라 안면있는 사람들이 보여서 인사를 나누었다. 또한 이번에 처음 알게된 사람들도 연구 이야기, 대학원 생활 이야기 등을 하며 즐겁게 쉬는 시간을 보냈다. 만찬은 Dynamic Earth 라는 굉장히 특이한 공간에서 열렸는데 과학/환경 박물관 인듯했다. 그리 크지 않은 건물로 보였지만 지하에 엄청난 공간이 있었고 공룡을 비롯한 각종 동물 모형, 지구환경 등이 전시되어 있었다. 만찬 때는 우연히 분석의 "sensitivity" 조절을 주제로 이번에 논문을 발표한 학생 세 명이 뭉쳤다. structure-sensitivity 를 이야기한 George Balatsouras, object-sensitivity 를 이야기한 Tian Tan, 그리고 나. 셋 다 실제 프로그램을 대상으로 하는 분석기를 만들고 있는 터라 관심사가 비슷해서 재미가 있었다. 연구 이외에도 한국인들은 남북 관계를 어떻게 생각하느냐 (예. 북한이 미사일 쏘면 어떻게 반응하는가?), 한국에서가장 인기 있는 스포츠는 무엇인가 등등 한국에 관해 여러가지를 궁금해했다. 둘 다 내가 안본 한국 영화를 서너편 본적 있다는 공통점도 있었다.

마무리

논문 발표도 하고 여러 사람도 만나 이야기도 나누고 오가며 여행도 즐기고. 학회를 참가하면 늘 많은 것을 얻고 돌아온다. 잠시 일상을 떠나 있으면서, 또 다른 훌륭한 사람들이 연구하는 것을 보면서 앞으로 계획도 다시 바로 잡아본다. 늘 크고 바른 길로 인도해주시고 좋은 기회를 주시는 이광근 교수님, 같이 연구하면서 좋은 결과에 함께 기뻐해 주셨던 양홍석, 오학주 교수님, 연구실 세미나 발표 때마다 함께 논의하면서 이번 논문에 좋은 아이디어를 주신 허충길 교수님과 동료들, 출장 준비 잘 도와주신 최은희 선생님께 감사드리면서 글을 마무리 한다.