# Hướng dẫn gán nhãn bộ dữ liệu cho bài toán sinh câu mô tả ảnh Tiếng Việt


1. Bùi Cao Doanh
2. ThS. Võ Duy Nguyên
3. TS. Nguyễn Tấn Trần Minh Khang

# Nội dung

1. Giới thiệu bài toán mô tả ảnh
2. Các bộ dữ liệu hiện có trên Tiếng Việt
3. Giới thiệu dữ liệu thô
4. Hướng dẫn gán nhãn

# Nội dung

1. Giới thiệu bài toán mô tả ảnh
2. Các bộ dữ liệu hiện có trên Tiếng Việt
3. Giới thiệu dữ liệu thô
4. Hướng dẫn gán nhãn

# 1. GIỚI THIỆU BÀI TOÁN MÔ TẢ ẢNH

# Giới thiệu bài toán mô tả ảnh

Đầu vào

Mô hình

Dòng người cầm hoa đào đang di chuyển trên một con phố có treo nhiều lồng đèn màu vàng ở hai bên.

Đầu ra

# 2. CÁC BỘ DỮ LIỆU TIẾNG VIỆT HIỆN CÓ

# Các bộ dữ liệu Tiếng Việt hiện có

| Bộ dữ liệu | Mô tả dữ liệu |
|---|---|
| UIT-ViIC [1] | - Số điểm ảnh: 3,850.<br>- Tỷ lệ câu mô tả trên 01 bức ảnh: 5<br>- Miền dữ liệu: ảnh thể thao, lấy từ bộ dữ liệu MS-COCO caption. |
| VieCap4H [2] | - Số điểm ảnh: 10,068<br>- Tỷ lệ câu mô tả trên 01 bức ảnh: xấp xỉ 1.15.<br>- Miền dữ liệu: ảnh y tế. |


- [1]. Lam, Quan Hoang, et al. "UIT-ViIC: A Dataset for the First Evaluation on Vietnamese Image Captioning." International Conference on Computational Collective Intelligence. Springer, Cham, 2020.
- [2]. T. M. Le, L. H. Dang, T.- S. Nguyen, T. M. H. Nguyen, and X. -S. Vu, "VLSP2021 - VieCap4H Challenge: Automatic Image Caption Generation for Healthcare Domain in Vietnamese," VNU Journal of Science: Computer Science and Communication Engineering, vol. 38, no. 1, 2022.

# 3. GIỚI THIỆU BỘ DỮ LIỆU THÔ

# Giới thiệu bộ dữ liệu thô

- Bộ dữ liệu thô bao gồm 13,100 ảnh thô được thu thập chủ yếu từ Google Hình ảnh và một phần ở Instagram và Facebook.
- Các ảnh thu thập có miền dữ liệu mở, bối cảnh Việt Nam.
- Bộ dữ liệu hiện tại được chia thành 10 thư mục con, mỗi thư mục con có 1,310 ảnh.

# 4. HƯỚNG DẪN GÁN NHÃN

# Luật gán nhãn

1. Mỗi câu mô tả có tối thiểu 10 từ.
2. Cố gắng mô tả thật chi tiết các nội dung quan trọng của bức ảnh về mặt thị giác.
3. Câu mô tả là 01 câu đơn (không có dấu phẩy hoặc dấu chấm), sử dụng thì hiện tại tiếp diễn để mô tả.
4. Có thể sử dụng có từ Tiếng Anh thông dụng: TV, laptop, …
5. KHÔNG mô tả những sự kiện sắp xảy ra.
6. KHÔNG đề cập tới các đối tượng văn bản.
7. KHÔNG thể hiện cảm xúc, câu mô tả cần phải khách quan.

# Luật gán nhãn

- Trước căn nhà được sơn màu xanh tím có một chiếc xe tải đang đậu.
- Căn nhà nằm ở mép ngoài cùng bên tay phải sử dụng cửa kéo.
- Khoảng sân phía trước căn nhà nằm ở mép ngoài cùng bên tay phải của bức ảnh người ta để rất nhiều chậu cây.
- Có một chiếc xe tải đang đứng đậu ở bên hông căn nhà ở giữa
- Một dãy gồm ba căn nhà có nhiều hơn một lầu.

# Luật gán nhãn

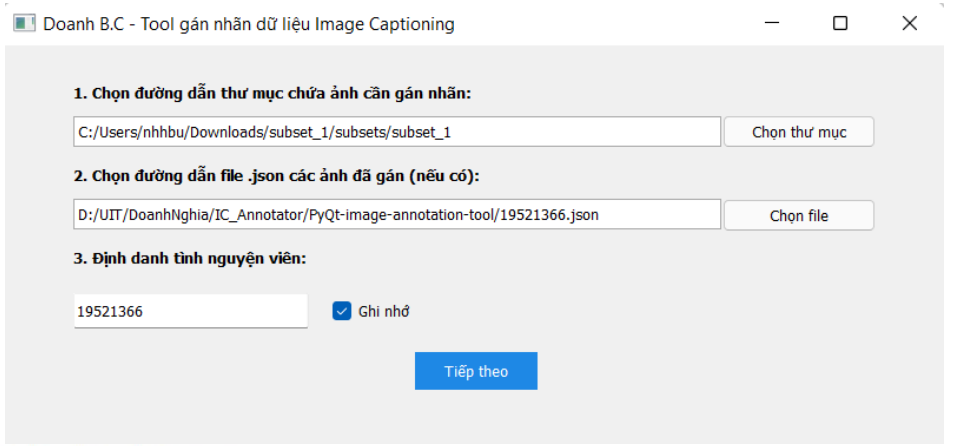
- Một người phụ nữ đang đứng xem những chiếc lồng đèn đang được trưng bày trong một cửa hàng bán lồng đèn.
- Ở một cửa tiệm bán các loại lồng đèn có một người phụ nữ đang đứng xem.
- Người phụ nữ đang đến xem ở một cửa hàng bán các loại lồng đèn giấy.
- Cửa hàng chỗ người phụ nữ đang đứng xem trưng bày rất nhiều loại lồng đèn giấy trang trí với đủ loại màu sắc.
- Một người phụ nữ với mái tóc dài màu vàng óng cùng đôi vai trần đang đứng trước một sạp trưng bày nhiều lồng đèn.

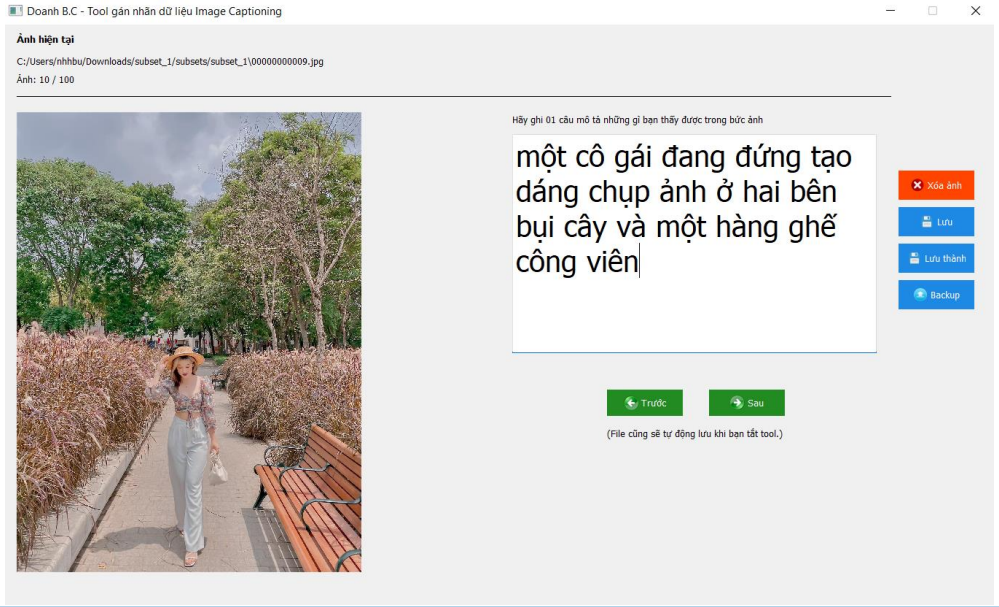
# Phân công

- Bộ dữ liệu có 13,100 ảnh, được chia thành các tập con nhỏ, mỗi tập con bao gồm 1,310 ảnh.
- 01 tình nguyện viên sẽ phụ trách 01 tập con. Sẽ có 05 tình nguyện viên gán chung 01 tập con, để đảm bảo 01 ảnh sẽ có 05 câu mô tả được gán từ 05 tình nguyện viên khác nhau.
- Các tình nguyện viên sẽ không biết các tập dữ liệu được phân công của nhau để đảm bảo tính khách quan.

# Tool gán nhãn

- Được viết bằng Python.
- Thư viện sử dụng để dựng giao diện: PyQt5.
- Thư viện khác: Numpy, Json, PyDrive, Datetime, Os.
- Link github: https://github.com/caodoanh2001/UIT-OpenViIC-labeller

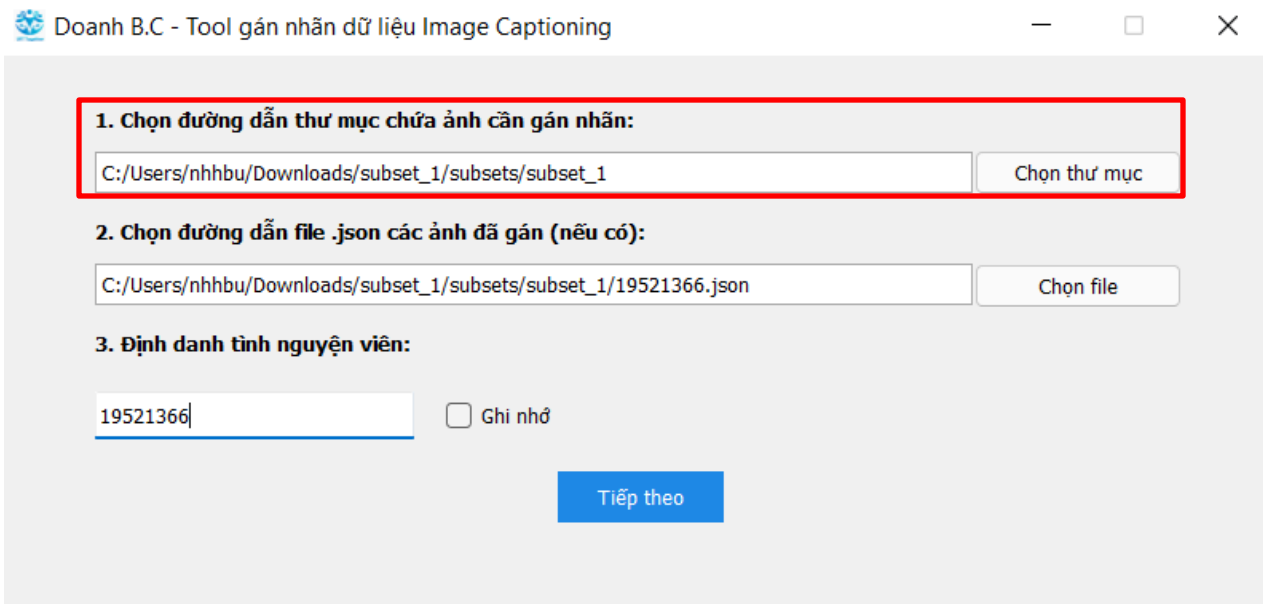


# Hướng dẫn gán nhãn

- Đầu tiên, khi mở tool gán nhãn:

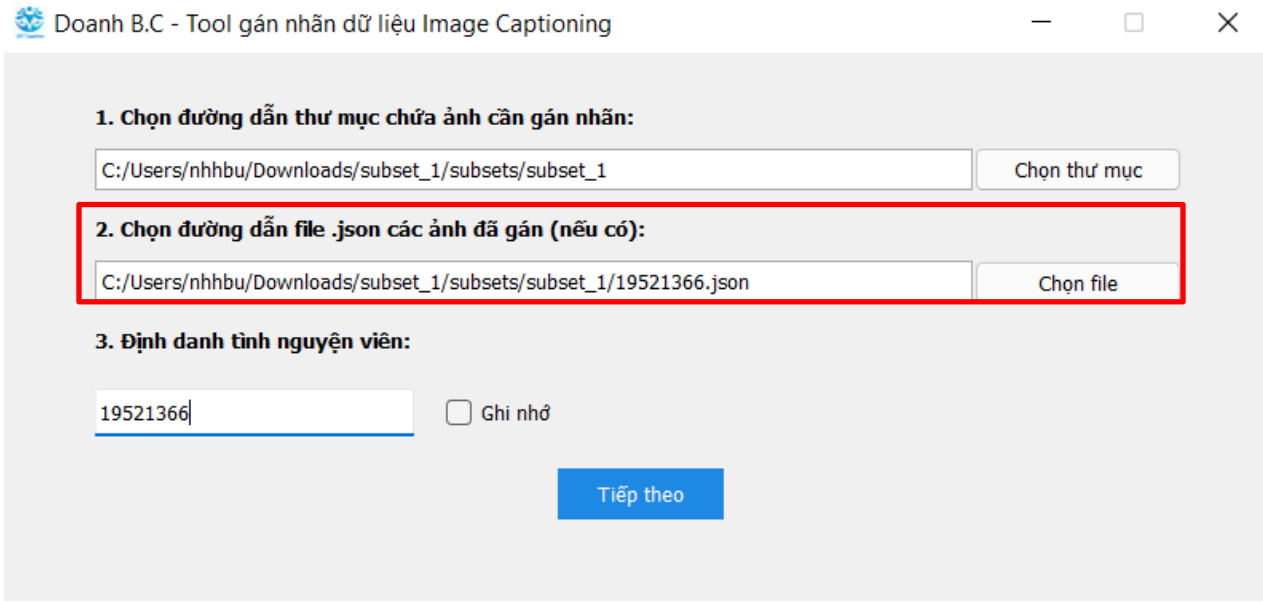
1. Chọn đường dẫn thư mục ảnh được phân công.
2. Chọn đường dẫn file annotation (trong trường hợp muốn load file lên để tiếp tục gán, gán lần đầu thì bỏ trống).
3. Gõ MSSV. File annotation sẽ được đặt tên trùng với MSSV của tình nguyện viên.
4. Chọn tiếp tục để sang màn hình gán nhãn.



# Hướng dẫn gán nhãn

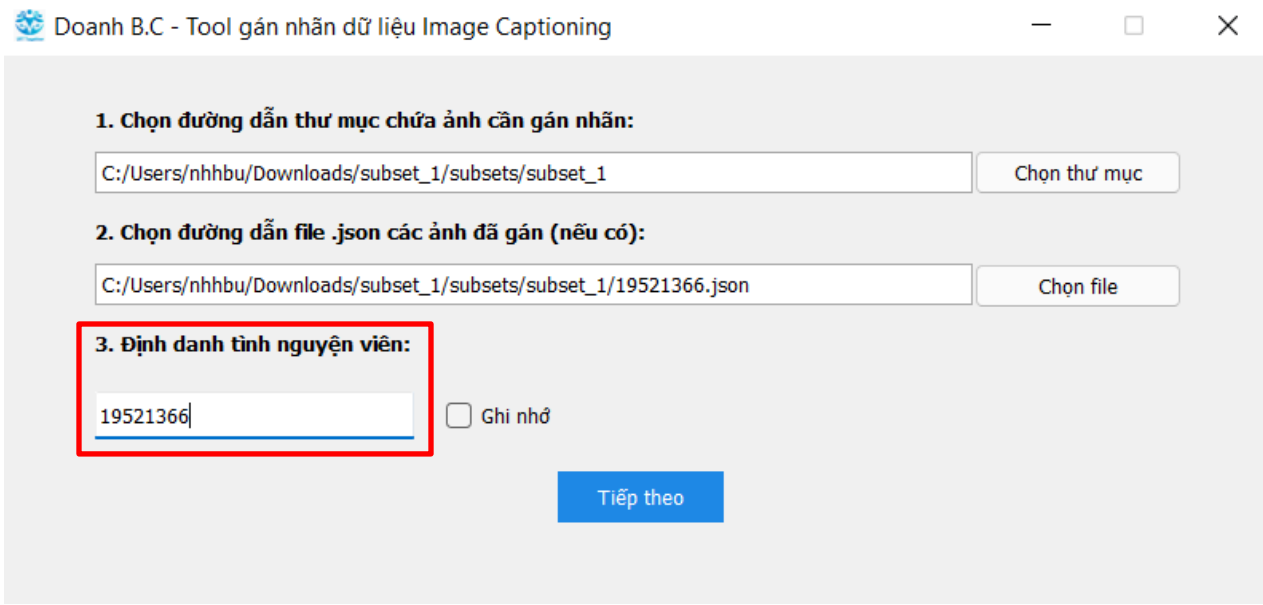
- Đầu tiên, khi mở tool gán nhãn:

1. Chọn đường dẫn thư mục ảnh được phân công.
2. Chọn đường dẫn file annotation (trong trường hợp muốn load file lên để tiếp tục gán, gán lần đầu thì bỏ trống).
3. Gõ MSSV. File annotation sẽ được đặt tên trùng với MSSV của tình nguyện viên.
4. Chọn tiếp tục để sang màn hình gán nhãn.



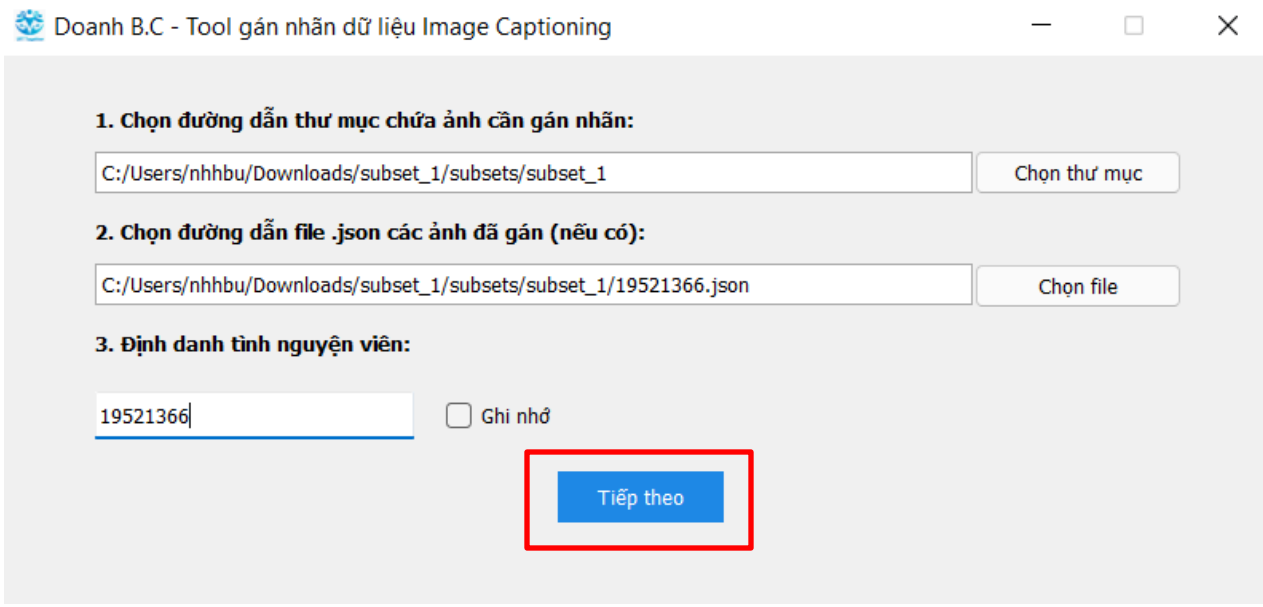
# Hướng dẫn gán nhãn

- Đầu tiên, khi mở tool gán nhãn:
  1. Chọn đường dẫn thư mục ảnh được phân công.
  2. Chọn đường dẫn file annotation (trong trường hợp muốn load file lên để tiếp tục gán, gán lần đầu thì bỏ trống).
  3. Gõ MSSV. File annotation sẽ được đặt tên trùng với MSSV của tình nguyện viên.
  4. Chọn tiếp tục để sang màn hình gán nhãn.



# Hướng dẫn gán nhãn

- Đầu tiên, khi mở tool gán nhãn:
  1. Chọn đường dẫn thư mục ảnh được phân công.
  2. Chọn đường dẫn file annotation (trong trường hợp muốn load file lên để tiếp tục gán, gán lần đầu thì bỏ trống).
  3. Gõ MSSV. File annotation sẽ được đặt tên trùng với MSSV của tình nguyện viên.
  4. Chọn tiếp tục để sang màn hình gán nhãn.



# Hướng dẫn gán nhãn

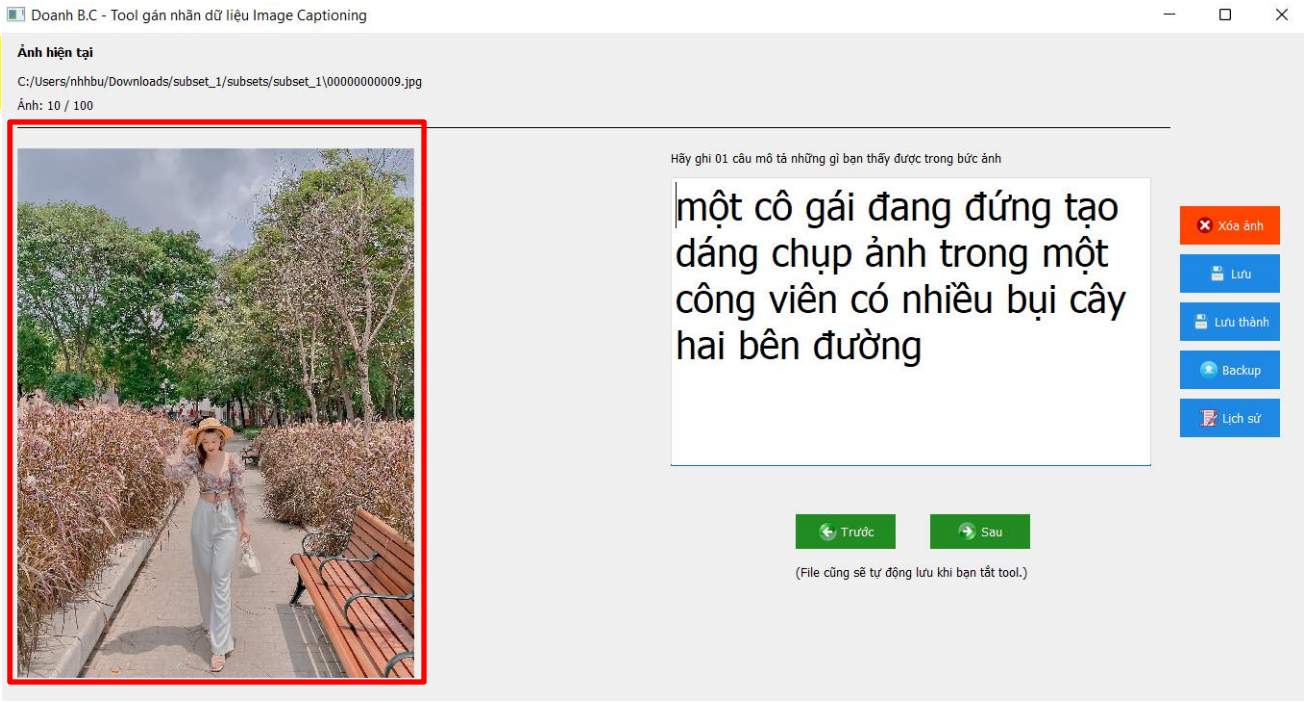
– Ở màn hình gán nhãn:

1. Ảnh cần gán sẽ hiện bên phía tay trái.

2. Các tình nguyện viên gõ 01 câu mô tả cho bức ảnh ở text box phía tay phải.

3. Chọn các nút Trước, Sau để tiến và lùi ảnh.

4. Chọn Lưu hoặc Lưu thành để lưu những câu đã gán ra file .JSON. Tên file mặc định là MSSV.



# Hướng dẫn gán nhãn

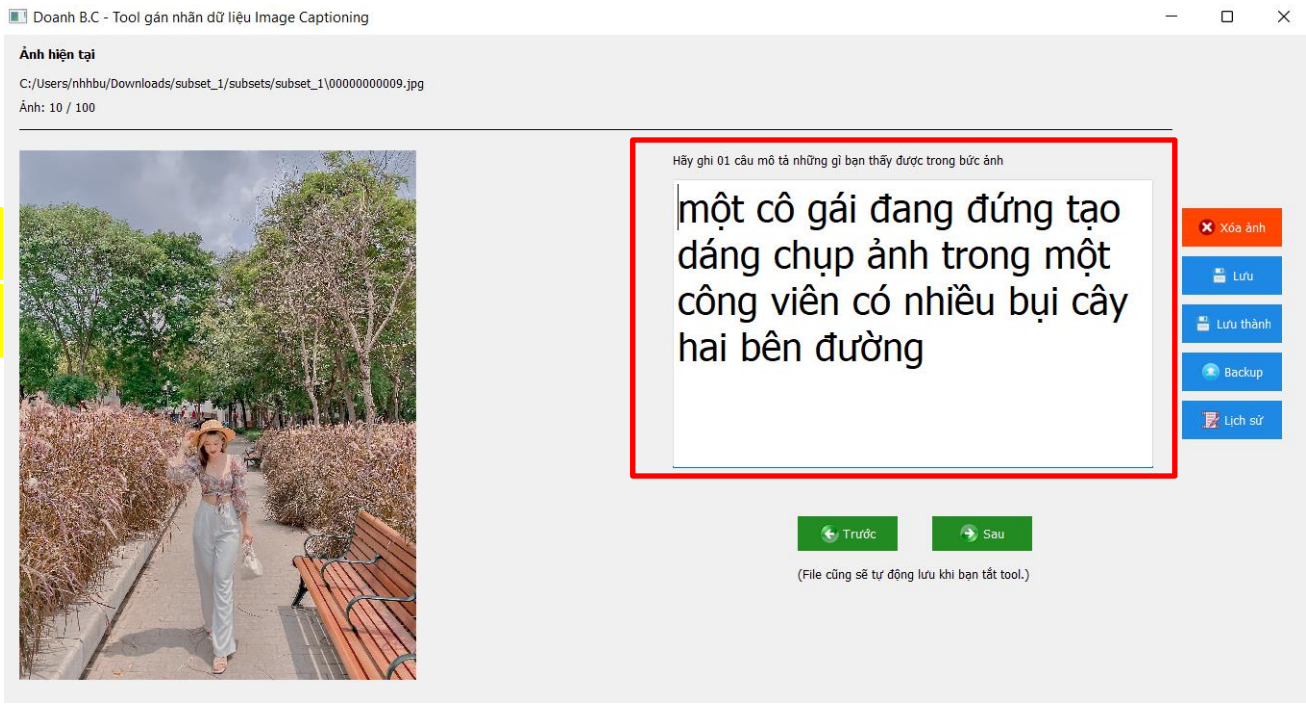
– Ở màn hình gán nhãn:

1. Ảnh cần gán sẽ hiện bên phía tay trái.

2. Các tình nguyện viên gõ 01 câu mô tả cho bức ảnh ở text box phía tay phải.

3. Chọn các nút Trước, Sau để tiến và lùi ảnh.

4. Chọn Lưu hoặc Lưu thành để lưu những câu đã gán ra file .JSON. Tên file mặc định là MSSV.



# Hướng dẫn gán nhãn

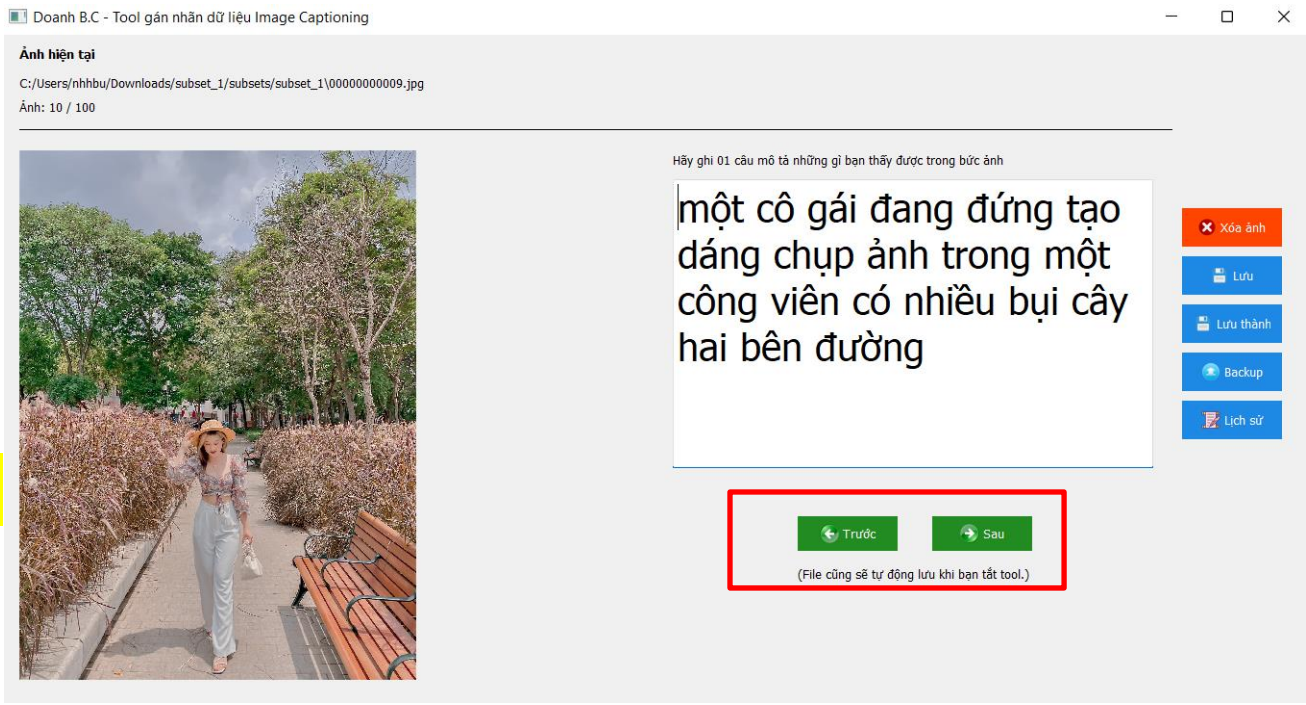
– Ở màn hình gán nhãn:

1. Ảnh cần gán sẽ hiện bên phía tay trái.

2. Các tình nguyện viên gõ 01 câu mô tả cho bức ảnh ở text box phía tay phải.

3. Chọn các nút Trước, Sau để tiến và lùi ảnh.

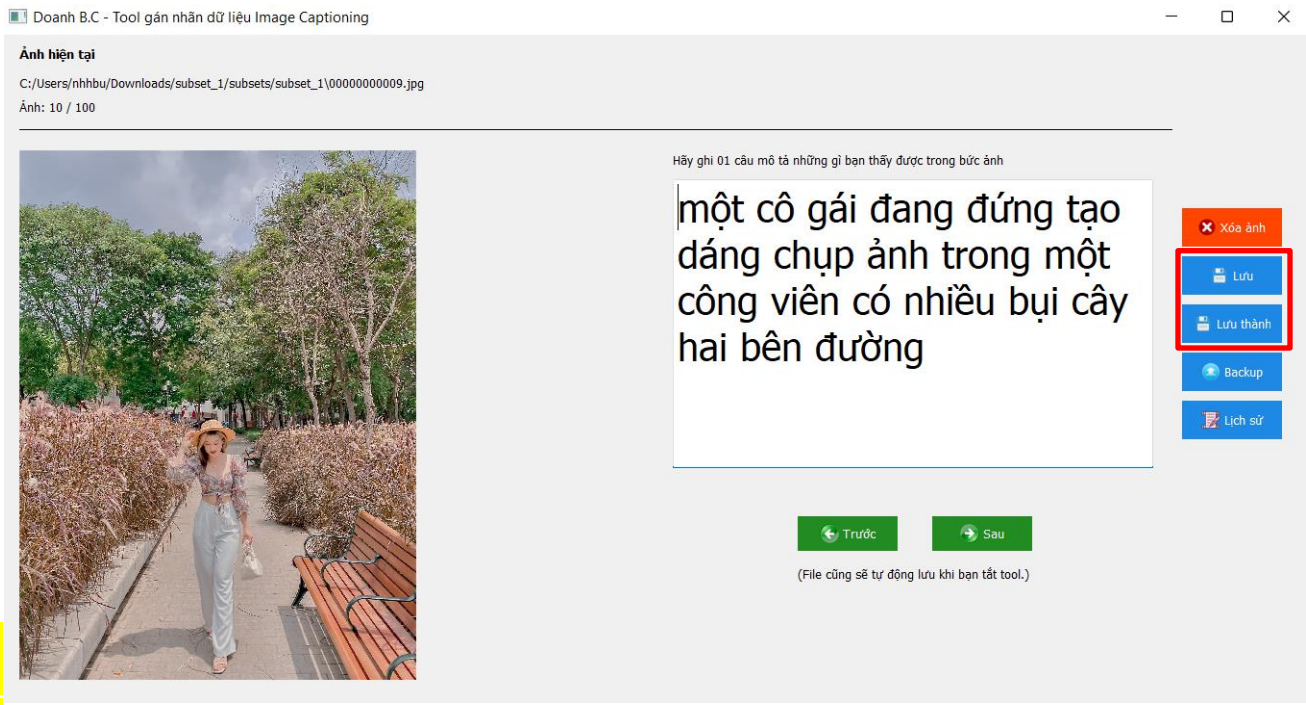
4. Chọn Lưu hoặc Lưu thành để lưu những câu đã gán ra file .JSON. Tên file mặc định là MSSV.



# Hướng dẫn gán nhãn

– Ở màn hình gán nhãn:

1. Ảnh cần gán sẽ hiện bên phía tay trái.
2. Các tình nguyện viên gõ 01 câu mô tả cho bức ảnh ở text box phía tay phải.
3. Chọn các nút Trước, Sau để tiến và lùi ảnh.
4. Chọn Lưu hoặc Lưu thành để lưu những câu đã gán ra file .JSON. Tên file mặc định là MSSV.



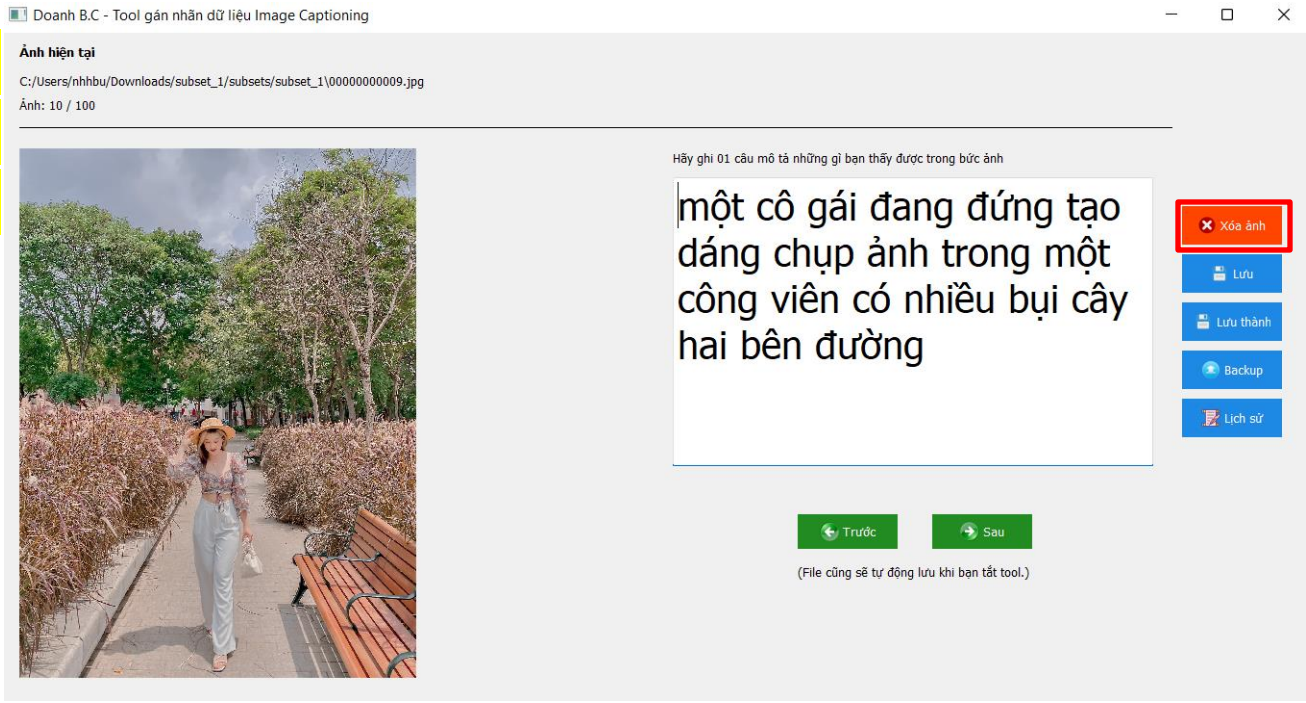
# Hướng dẫn gán nhãn

– Ở màn hình gán nhãn:

5. Khi có những ảnh không thể hiện được tức là ảnh bị lỗi, chọn “Xóa ảnh” để ghi nhận ảnh sẽ không được sử dụng.

6. Tại bất kỳ thời điểm nào, có thể chọn “Backup” để lưu file .JSON hiện tại lên Google Drive. Tên file được upload lên Drive: <MSSV>_<NGÀY-THÁNG-NĂM>

7. Ngoài ra, click nút “Lịch sử” sẽ xem và tải được các file đã lưu theo ngày.



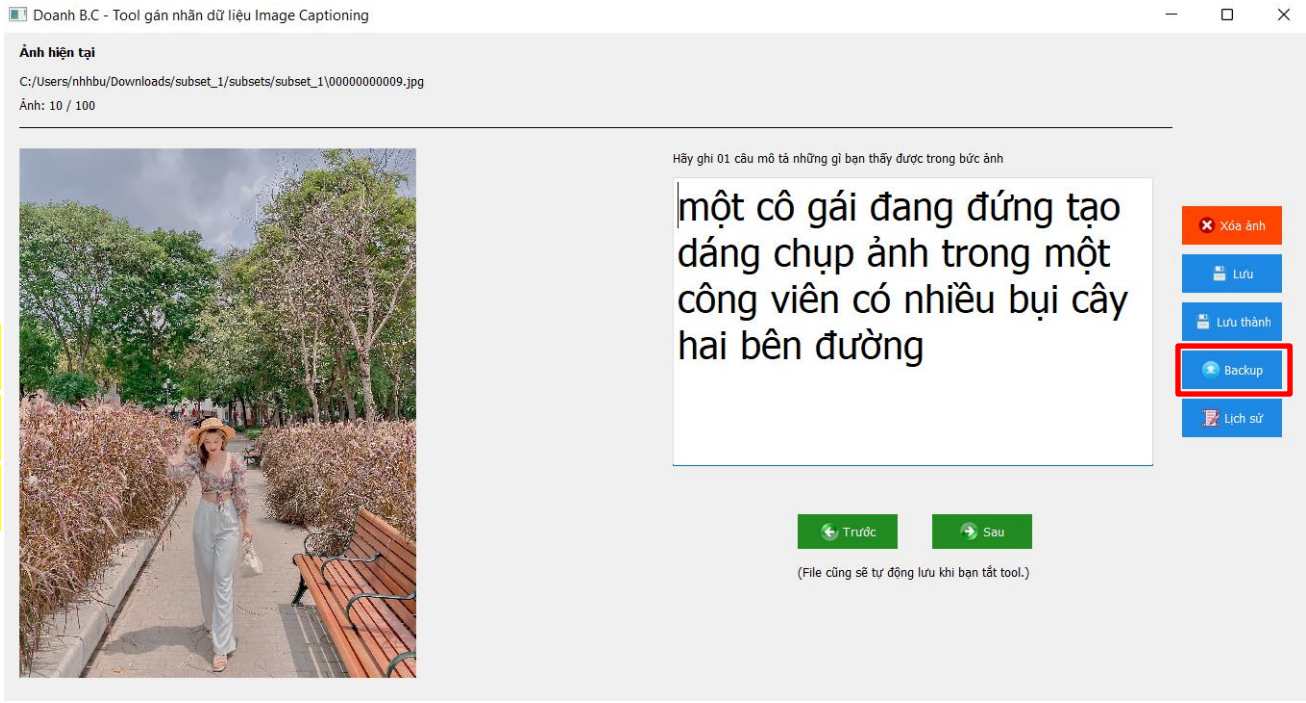
# Hướng dẫn gán nhãn

– Ở màn hình gán nhãn:

5. Khi có những ảnh không thể hiện được tức là ảnh bị lỗi, chọn “Xóa ảnh” để ghi nhận ảnh sẽ không được sử dụng.

6. Tại bất kỳ thời điểm nào, có thể chọn “Backup” để lưu file .JSON hiện tại lên Google Drive. Tên file được upload lên Drive: <MSSV>_<NGÀY-THÁNG-NĂM>

7. Ngoài ra, click nút “Lịch sử” sẽ xem và tải được các file đã lưu theo ngày.



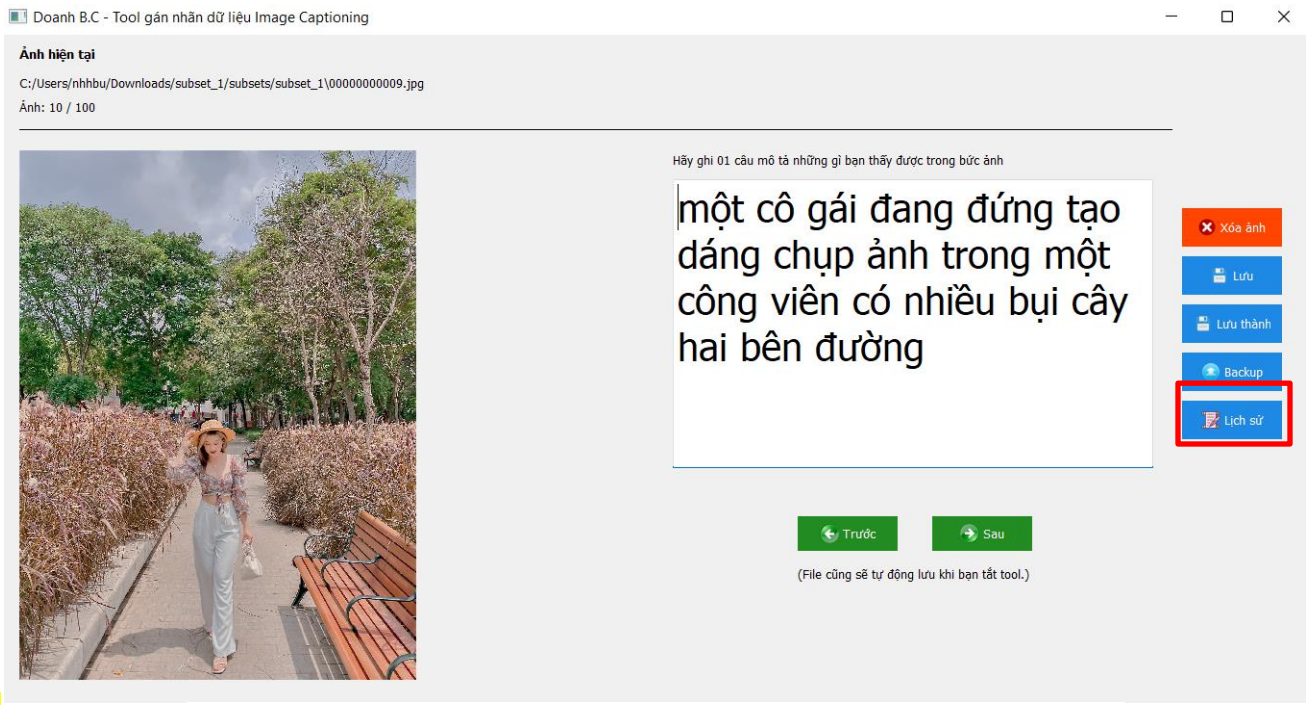
# Hướng dẫn gán nhãn

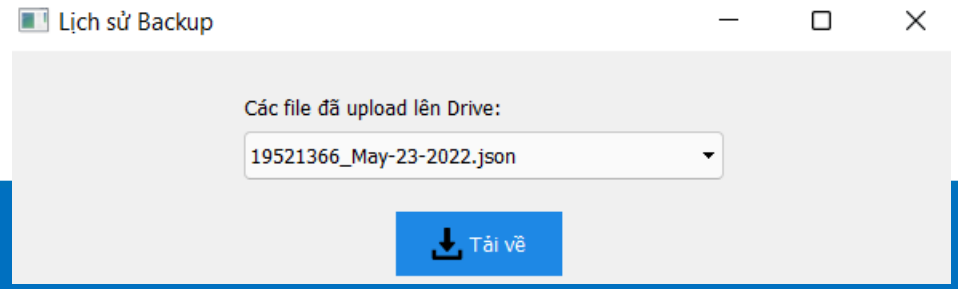
- Ở màn hình gán nhãn:

5. Khi có những ảnh không thể hiện được tức là ảnh bị lỗi, chọn “Xóa ảnh” để ghi nhận ảnh sẽ không được sử dụng.

6. Tại bất kỳ thời điểm nào, có thể chọn “Backup” để lưu file .JSON hiện tại lên Google Drive. Tên file được upload lên Drive: <MSSV>_<NGÀY-THÁNG-NĂM>

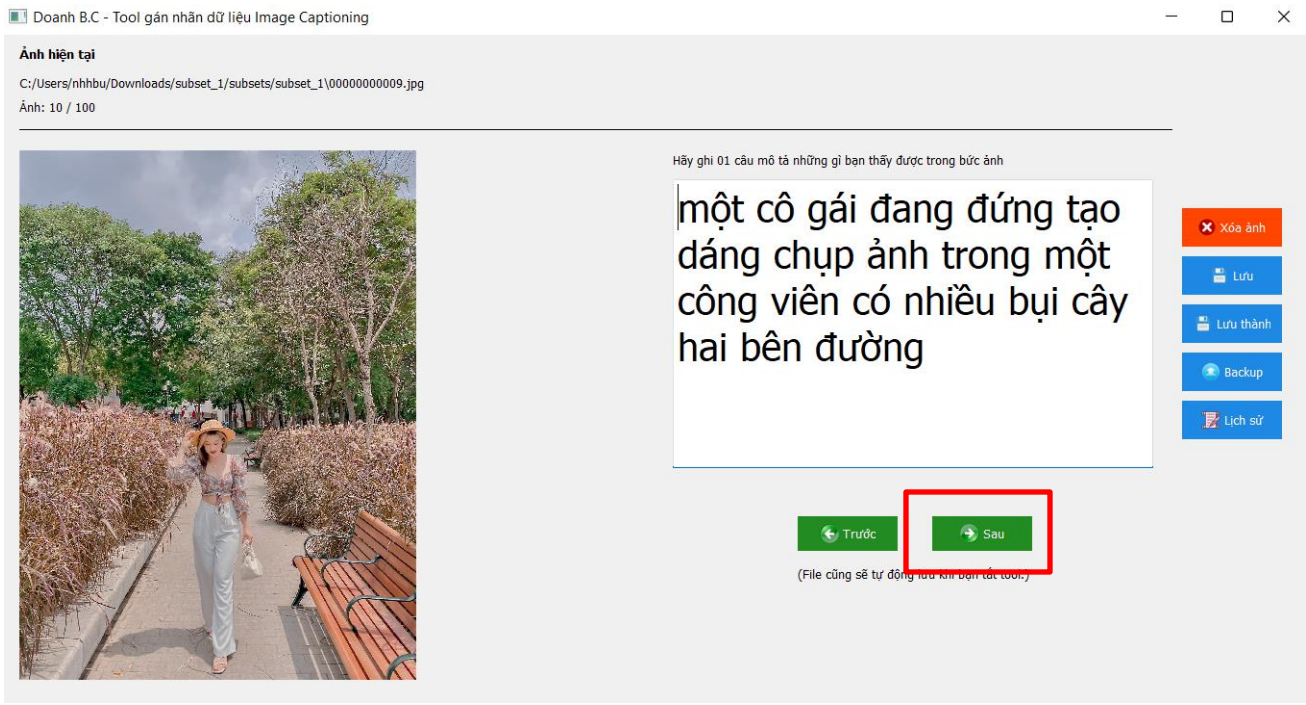
7. Ngoài ra, click nút “Lịch sử” sẽ xem và tải được các file đã lưu theo ngày.



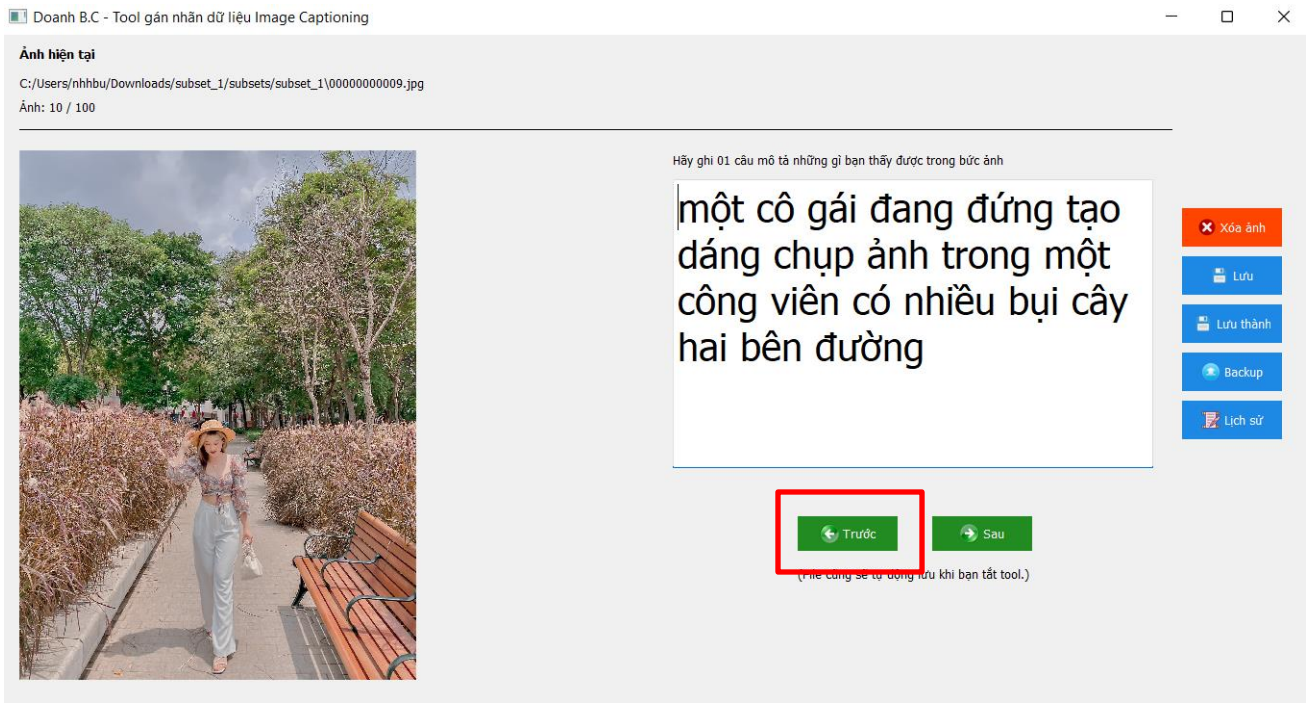
# Hướng dẫn gán nhãn

- Các phím nóng:
  - + Ctrl →: sang ảnh kế tiếp.
  - + Ctrl ←: lùi lại ảnh phía trước.
  - + Ctrl + S: lưu file nhãn .JSON.
  - + Ctrl + D: xóa ảnh.
  - + Ctrl + Tab: tắt con trỏ chuột ở text box.



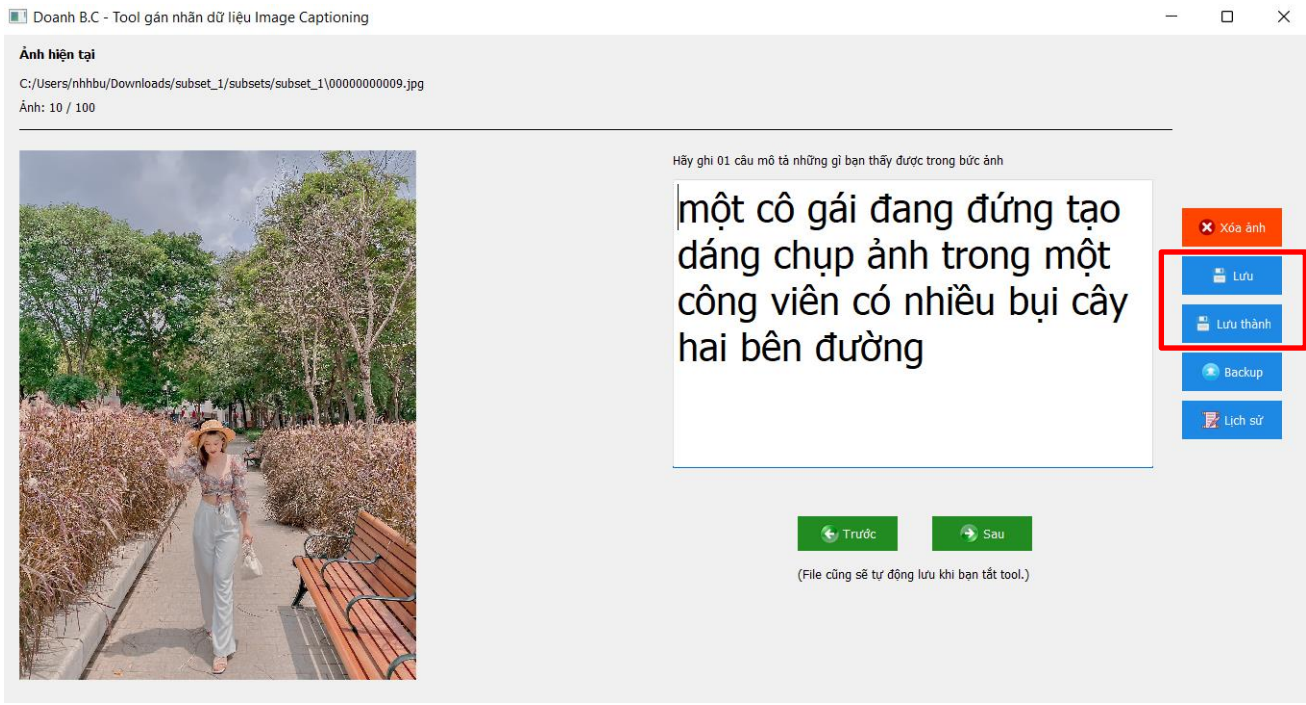
# Hướng dẫn gán nhãn

- Các phím nóng:
  - + Ctrl →: sang ảnh kế tiếp.
  - + Ctrl ←: lùi lại ảnh phía trước.
  - + Ctrl + S: lưu file nhãn .JSON.
  - + Ctrl + D: xóa ảnh.
  - + Ctrl + Tab: tắt con trỏ chuột ở text box.



# Hướng dẫn gán nhãn

- Các phím nóng:
  - + Ctrl →: sang ảnh kế tiếp.
  - + Ctrl ←: lùi lại ảnh phía trước.
  - + Ctrl + S: lưu file nhãn .JSON.
  - + Ctrl + D: xóa ảnh.
  - + Ctrl + Tab: tắt con trỏ chuột ở text box.



# Hướng dẫn gán nhãn

– Các phím nóng:

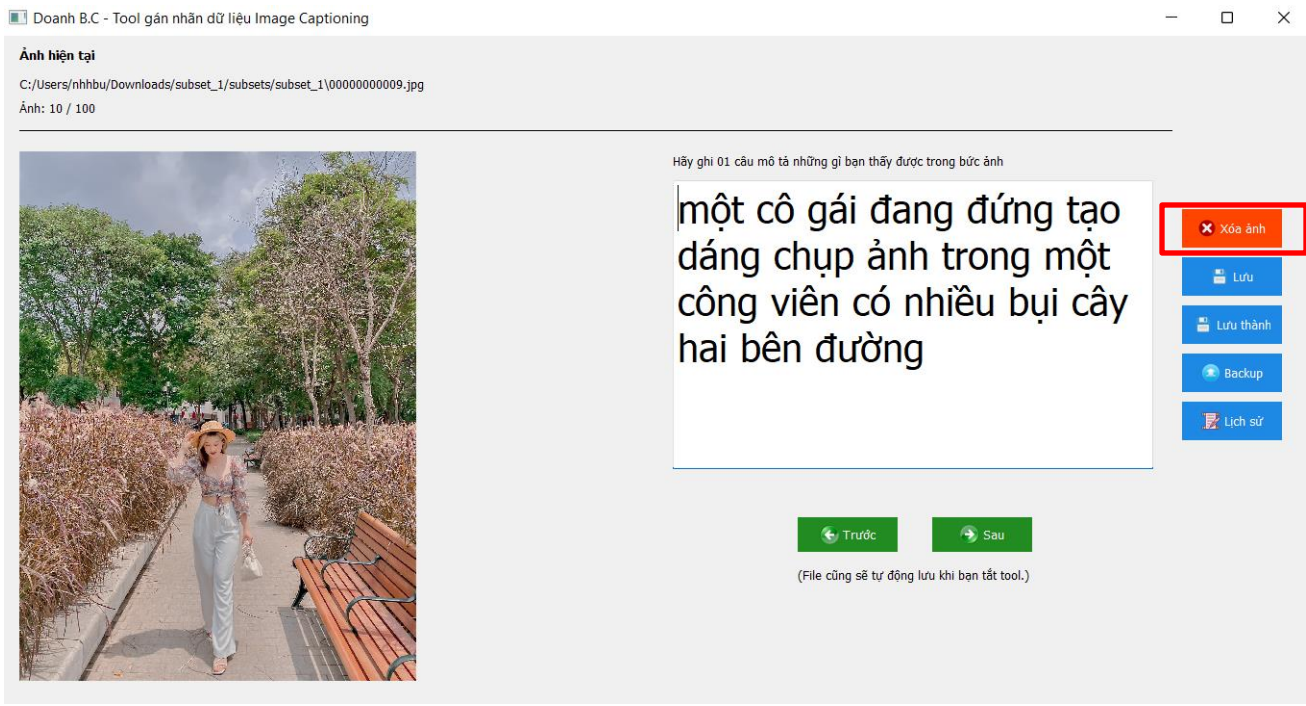
+ Ctrl →: sang ảnh kế tiếp.

+ Ctrl ←: lùi lại ảnh phía trước.

+ Ctrl + S: lưu file nhãn .JSON.

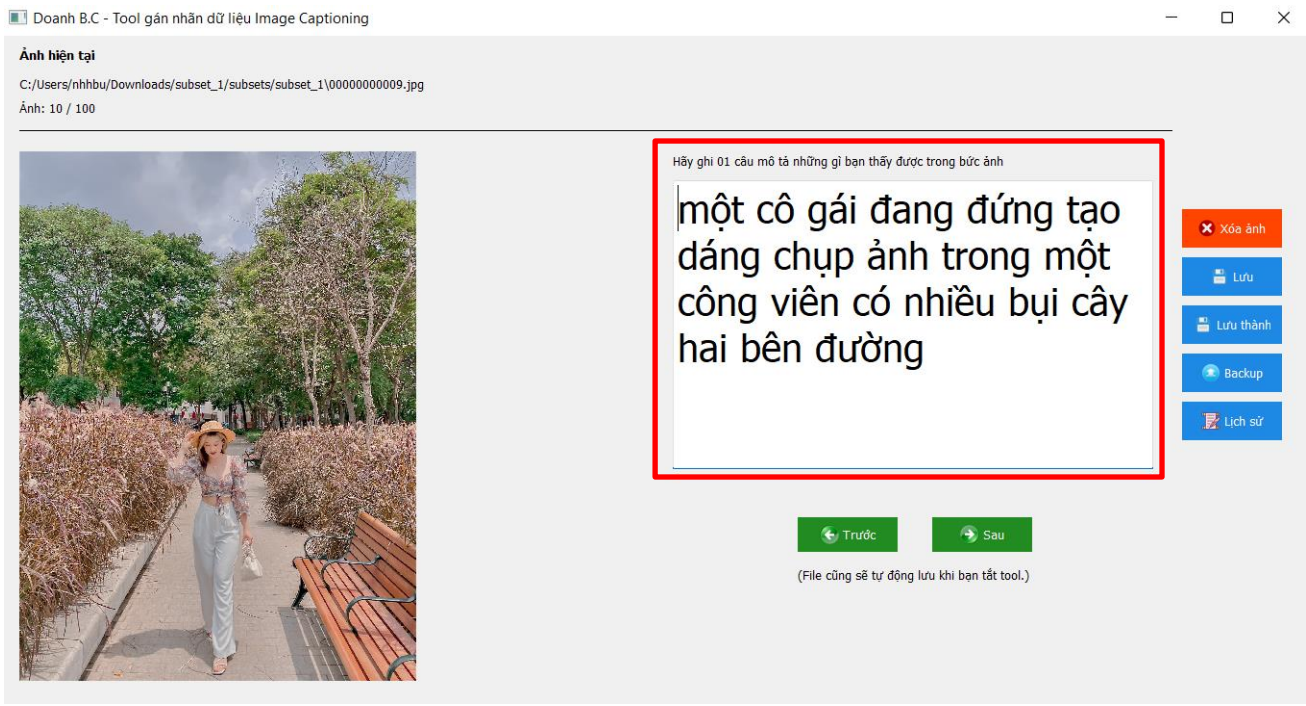
+ Ctrl + D: xóa ảnh.

+ Ctrl + Tab: tắt con trỏ chuột ở text box.



# Hướng dẫn gán nhãn

- Các phím nóng:
  - + Ctrl →: sang ảnh kế tiếp.
  - + Ctrl ←: lùi lại ảnh phía trước.
  - + Ctrl + S: lưu file nhãn .JSON.
  - + Ctrl + D: xóa ảnh.
  - + Ctrl + Tab: tắt con trỏ chuột ở text box.



**Cảm ơn quí vị đã lắng nghe**

**Nhóm tác giả**
Bùi Cao Doanh
ThS. Võ Duy Nguyên
TS. Nguyễn Tấn Trần Minh Khang

# Một số vấn đề thường gặp

- Q: Gán giữa chừng xong có thể lưu lại gán tiếp không?
  + A: Có. Bất kỳ lúc nào có thể sử dụng phím tắt Ctrl+S để lưu file ra file .JSON, sau đó load file này lên để tiếp tục gán.
- Q: Phương án backup?
  + A: Bất kỳ thời điểm nào, các tình nguyện viên có thể click vào nút “Backup”, file .JSON đang gán sẽ được upload lên Google Drive của tài khoản uit.open.viic@gmail.com, tên file là: <MSSV>_<NGÀY-THÁNG-NĂM>.json. Lưu ý rằng cần có kết nối Internet để thực hiện backup.
- Q: Ảnh như thế nào thì nên xóa?
  + Xóa những ảnh không hiển thị được trên tool. Còn lại nên cố gắng thực hiện viết câu mô tả, nếu ảnh đó không có gì để mô tả cũng có thể xóa.