Weakly Supervised Dense Video Captioning

Supplementary Materials

1. Region-sequence Generation Algorithm

Algorithm 1 describes the region-sequence generation method, which is based on the CELF (Cost-Effective Lazy Forward selection) algorithm [1]. In this algorithm, m is the number of regions in a sequence, UC and CB are the abbreviation for uniform cost and cost benefit respectively.

2. Response Maps

Figure 1 shows some examples of response maps (heatmaps) generated by the Lexical-FCN model. We first associate the response maps to the words in the sentences based on the computed probabilities, and then visualize the best match.

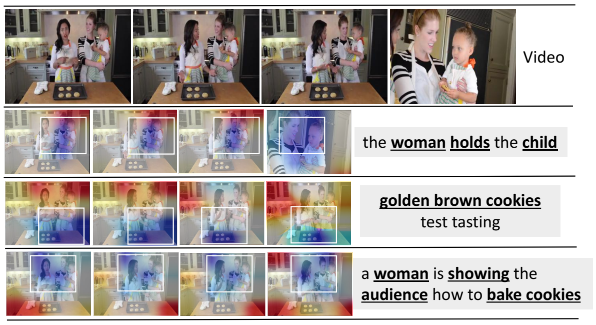

Video
the woman holds the child
golden brown cookies test tasting
a woman is showing the audience how to bake cookies

Figure 1. Visualization of learned response maps from the last CNN layer (left), and the corresponding natural sentences (right). The blue areas in the response maps are of high attention, and the region-sequences are highlighted in white bounding-boxes.

3. Sentence Re-ranking Module

Figure 2 shows the diagram of our sentence re-ranking module, which re-rank multiple predicted sentences from dense video captioning. This module is similar to [2], which learns the cosine similarity between video features and sentence features with a neural network evaluator.

4. More Result Examples

More result examples of our DenseVideoCap system are provided in Figures 3, 4, 5, 6.

Algorithm 1 Region-sequence generation by submodular maximization with the CELF algorithm [1].

```
function LazyForward($\mathcal{S}_\mathbf{v}$, $x_\mathbf{v}$, $R$, $m$, $type$)
    $\mathcal{A} \leftarrow \emptyset$;    ▷ Start with the empty sequence
    for each $r \in \mathcal{S}_\mathbf{v}$ do $\mathcal{L}(\mathbf{w}; r) \leftarrow \infty$    ▷ Init marginal gains
    end for
    while $|\mathcal{A}| < m$ do
        for each $r \in \mathcal{S}_\mathbf{v} \backslash \mathcal{A}$ do $cur_s \leftarrow$ false;
        end for
        while true do    ▷ Begin loop
            if $type = UC$ then    ▷ Uniform cost
                $r^* \leftarrow \arg\max_{r \in \mathcal{S}_\mathbf{v} \backslash \mathcal{A}} \mathcal{L}(\mathbf{w}; r)$;    ▷ Max gain
            end if
            if $type = CB$ then    ▷ Cost benefit
                $r^* \leftarrow \arg\max_{r \in \mathcal{S}_\mathbf{v} \backslash \mathcal{A}} \frac{\mathcal{L}(\mathbf{w};r)}{R(r)}$;    ▷ Max gain / cost
            end if
            if $cue_{r^*}$ then $\mathcal{A} \leftarrow \mathcal{A} \cup \{r^*\}$; break;
            else    ▷ Update marginal gain
                $\mathcal{L}(\mathbf{w}; r) \leftarrow R(\mathcal{A} \cup \{r\}) - R(\mathcal{A})$;
                $cur_{r^*} \leftarrow$ true;
            end if
        end while
    end while
    return $\mathcal{A}$;    ▷ Return region-sequence
end function

function Main($\mathcal{S}_\mathbf{v}$, $x_\mathbf{v}$, $R$, $m$)
    $\mathcal{A}_{UC} \leftarrow$ LazyForward($\mathcal{S}_\mathbf{v}$, $x_\mathbf{v}$, $R$, $m$, $UC$)
    $\mathcal{A}_{CB} \leftarrow$ LazyForward($\mathcal{S}_\mathbf{v}$, $x_\mathbf{v}$, $R$, $m$, $CB$)
    return $\arg\max\{R(\mathcal{A}_{UC}),\ R(\mathcal{A}_{CB})\}$
end function
```

References

[1] J. Leskovec, A. Krause, and et al. Cost-effective outbreak detection in networks. In *ACM SIGKDD*, 2007. 1

[2] R. Shetty and J. Laaksonen. Frame-and segment-level features and candidate pool evaluation for video caption generation. *arXiv:1608.04959*, 2016. 1

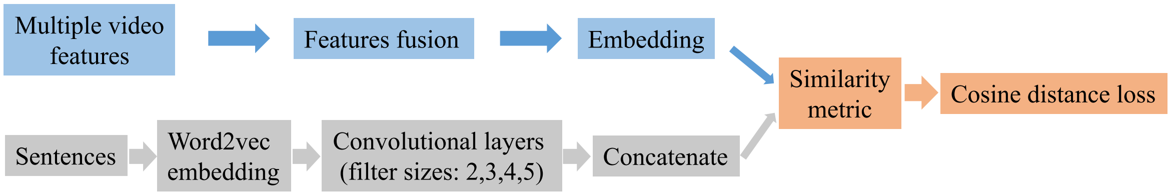
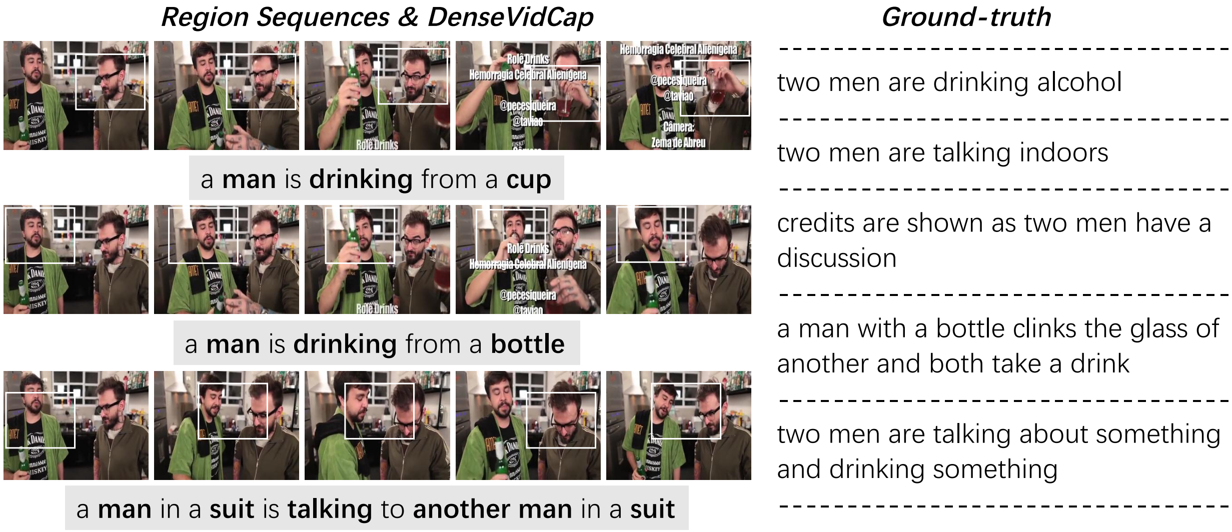


Figure 2. Illustration of the sentence re-ranker module.



Region Sequences & DenseVidCap

a **man** is **drinking** from a **cup**

a **man** is **drinking** from a **bottle**

a **man** in a **suit** is **talking** to **another man** in a **suit**

Ground-truth

two men are drinking alcohol

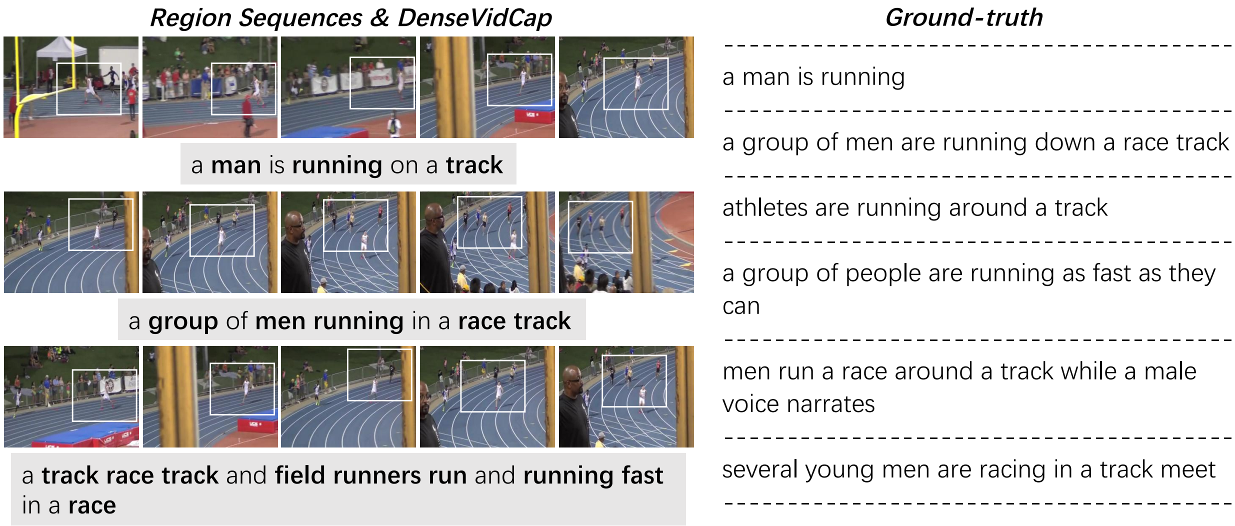
two men are talking indoors

credits are shown as two men have a discussion

a man with a bottle clinks the glass of another and both take a drink

two men are talking about something and drinking something

Figure 3. Left: Examples of dense sentences produced by our ***DenseVidCap*** method and corresponding ***region sequences***; Right: Ground-truth (video6974).



Region Sequences & DenseVidCap

a **man** is **running** on a **track**

a **group** of **men running** in a **race track**

a **track race track** and **field runners run** and **running fast** in a **race**

Ground-truth

a man is running

a group of men are running down a race track

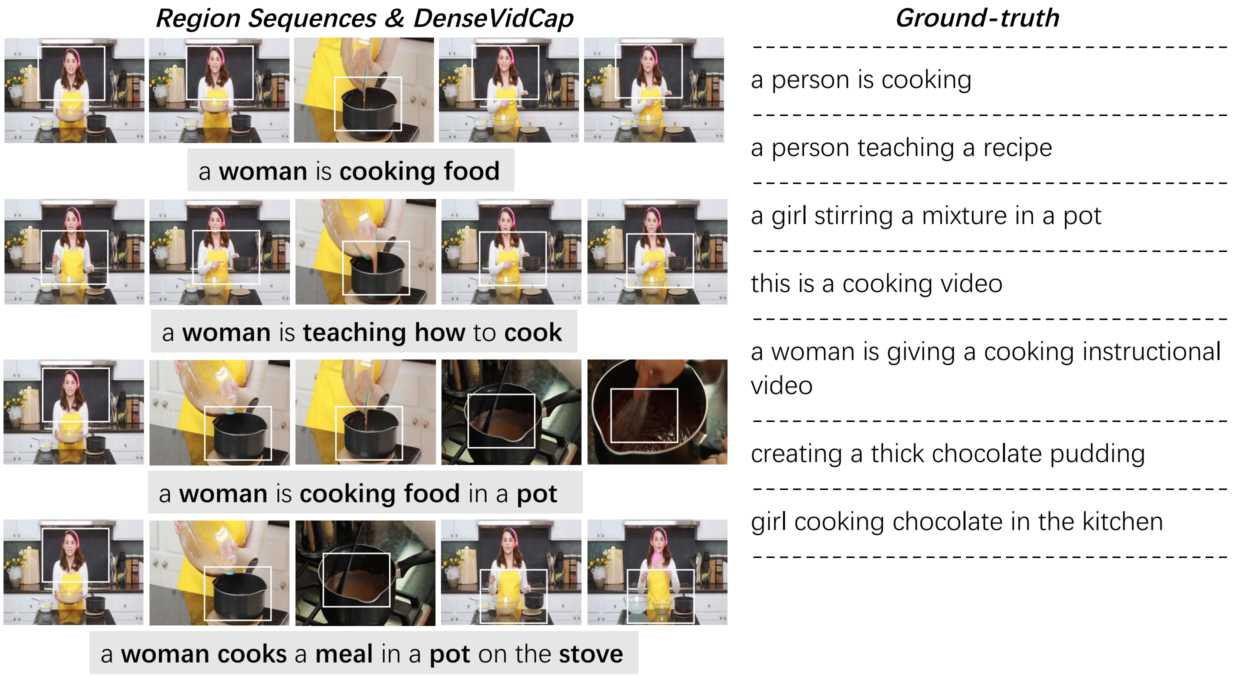
athletes are running around a track

a group of people are running as fast as they can

men run a race around a track while a male voice narrates

several young men are racing in a track meet

Figure 4. Left: Examples of dense sentences produced by our ***DenseVidCap*** method and corresponding ***region sequences***; Right: Ground-truth (video6967).



Region Sequences & DenseVidCap

a **woman** is **cooking food**

a **woman** is **teaching how** to **cook**

a **woman** is **cooking food** in a **pot**

a **woman cooks** a **meal** in a **pot** on the **stove**

Ground-truth

a person is cooking

a person teaching a recipe

a girl stirring a mixture in a pot

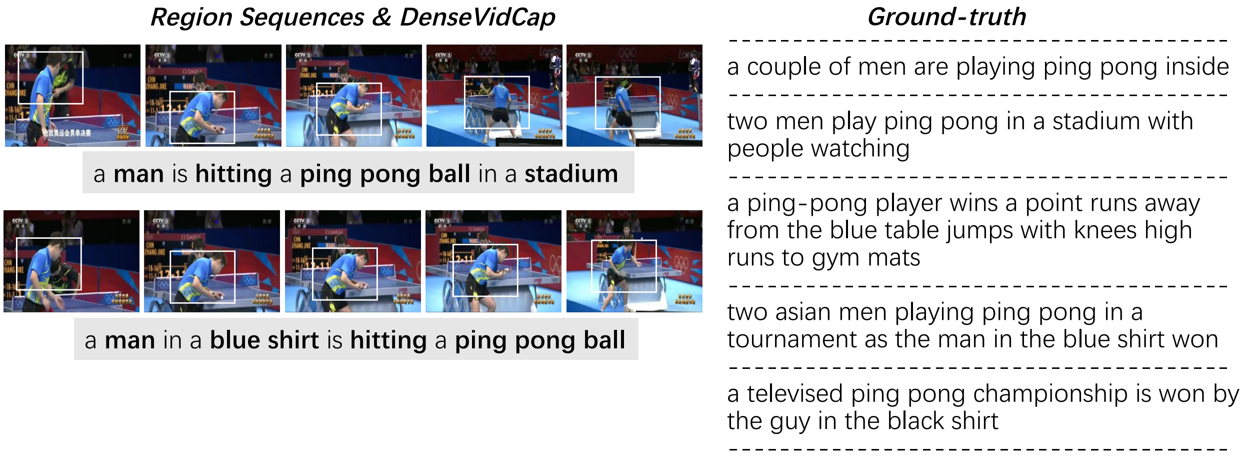
this is a cooking video

a woman is giving a cooking instructional video

creating a thick chocolate pudding

girl cooking chocolate in the kitchen

Figure 5. Left: Examples of dense sentences produced by our ***DenseVidCap*** method and corresponding ***region sequences***; Right: Ground-truth (video6911).



Region Sequences & DenseVidCap

a **man** is **hitting** a **ping pong ball** in a **stadium**

a **man** in a **blue shirt** is **hitting** a **ping pong ball**

Ground-truth

a couple of men are playing ping pong inside

two men play ping pong in a stadium with people watching

a ping-pong player wins a point runs away from the blue table jumps with knees high runs to gym mats

two asian men playing ping pong in a tournament as the man in the blue shirt won

a televised ping pong championship is won by the guy in the black shirt

Figure 6. Left: Examples of dense sentences produced by our ***DenseVidCap*** method and corresponding ***region sequences***; Right: Ground-truth (video6973).